

Communications of **HUAWEI RESEARCH**

November 2024
Issue 7



Machine Learning in Communications:
A Road to **Intelligent Transmission** and Processing p02

AI in the 5G-A Era: Scenarios, Key Technologies, and Evolution Trends p21

Ten Issues of **NetGPT** p52

Semantic Digital Twins: Enhancing Performance in
Wireless Communication and LLM Inference p128

Digital Twin **Online Channel Model**: Vision, Progress, and Challenges p135



From Connected People and Things to **Connected Intelligence**



Editorial Note

In 2018, Huawei introduced the concept of "Connected Intelligence" for 6G. Today, as the AI revolution gains momentum across various sectors, our mission is to develop a next-generation mobile platform that offers intelligent services to every person, family, and organization, at any time and from anywhere.

In November 2023, the International Telecommunication Union-Radiocommunication Sector (ITU-R) reached a global consensus on 6G framework recommendations, recognizing AI and communication integration as one of the key pillars for 6G. This milestone signifies that AI for Network and Network for AI are the iconic generational technologies of 6G, driving unprecedented advancements in devices and supercomputing centers toward full AI. Just as Johannes Gutenberg's printing press ignited the information revolution and the industrial revolution enabled machines to surpass human physical capabilities, modern AI technologies will empower machines to outperform humans in intelligence. Looking ahead, 6G mobile networks will serve as the foundation for AI and communication integration, delivering super-intelligent services to every person and everything.

Technological innovation will, as always, drive the evolution of the mobile industry, and the 6G era is no exception. With a market window spanning from 2035 to 2045, 6G networks and devices must adapt to meet the burgeoning demands of future consumers and vertical industries. Two decades ago, the Internet was the enabler of new technologies, and its adoption enabled mobile communication to unprecedented commercial success. Now, AI has become the enabler of new technologies, fueling exponential growth and propelling humanity toward a super-intelligent digital world. The convergence of 6G mobile communication and AI will redefine the paradigm for mobile networks, creating a seamless interface between the digital and physical worlds and unlocking the potential for artificial general intelligence (AGI) and embodied AI. 6G networks should integrate AI throughout the entire process of sensing, inference, decision-making, and physical operations in devices, wireless networks, and core networks, rather than being confined to generative AI. By harnessing the power of AI and communication integration, 6G networks will deliver integrated sensing and communication (ISAC), making native network intelligence a tangible reality.

In terms of 6G device evolution, AI is crucial for sparking revolutionary breakthroughs and driving the entire ecosystem forward. As we enter the post-mobile-broadband era, breakthroughs in device technologies will be pivotal in shaping the future of the mobile industry, and 6G mobile devices will undergo a transformative shift toward full AI. This elevation of 6G networks through enhanced device capabilities will ultimately lead the wireless industry to success.

In terms of architectural design, 6G must transcend the conventional service-based architecture (SBA) and leverage Agentic AI for technological reconstruction. This strategic move will pave the way for application-driven generative networks (ADGNs), setting 6G apart from its predecessor, 5G. Rather than simply building upon existing technologies and architectures, 6G seeks to catalyze far-reaching innovations across the mobile industry and beyond.

Centering on the applications of AI in wireless communication, this special issue brings together a diverse range of research findings from various stakeholders. In addition to insights from Huawei experts, it showcases the visions and achievements of the academic community and our partners. By sharing these findings, we aim to help realize the 6G vision of "Connected Intelligence."



Dr. Wen Tong
Huawei Fellow



Dr. Peiyong Zhu
Huawei Fellow

Editor-in-Chief:

Heng Liao

Executive Editors:

Wen Tong, Peiyong Zhu

Editorial Board:

Heng Liao, Wen Tong, Xinhua Xiao,
Banghong Hu, Huihui Zhou, Feng Bao,
Jeff Xu, Haibo Chen, Pinyan Lu,
Jianbing Wang, Ruihua Li, Bo Bai

E-mail: HWResearch@huawei.com

CONTENTS

Outlook



Machine Learning in Communications: A Road to Intelligent Transmission and Processing 02

Shixiong Wang, Geoffrey Ye Li

AI in the 5G-A Era: Scenarios, Key Technologies, and Evolution Trends 21

Yingpei Lin, Yan Chen, Yi Qin, Yan Sun, Rui Xu, Yuwen Yang,
Zhengming Zhang, Jiakuan Chen, Yang Tian, Youlong Cao,
Xiaomeng Chai, Hongzhi Chen, Hong Qi, Xu Pang

Anticipating 6G: Communication + AI 37

Xiongyan Tang, Youxiang Wang, Tengfei Sui

Computing as a Service: Unlocking the Boundless Potential of User Equipment 42

Yannan Yuan, Qi Wu, Yanchao Kang, Jiankang Liu, Xiaowen Sun,
Dajie Jiang, Fei Qin

Ten Issues of NetGPT 52

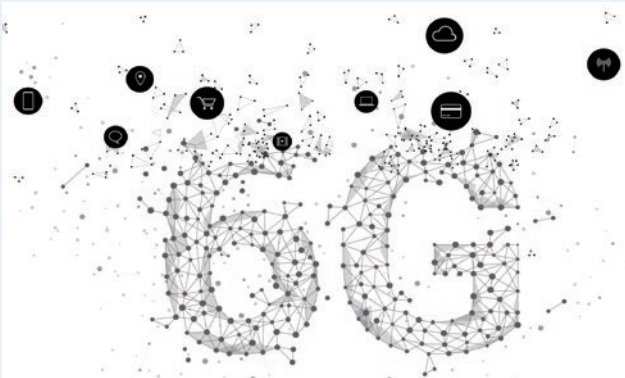
Wen Tong, Chenghui Peng, Tingting Yang, Fei Wang, Juan Deng,
Rongpeng Li, Lu Yang, Honggang Zhang, Dong Wang, Ming Ai,
Li Yang, Guangyi Liu, Yang Yang, Yao Xiao, Liexiang Yue,
Wanfei Sun, Zexu Li, Wenwen Sun

Key Issues and Technological Explorations of Large Models in 6G Network Communications 58

Tingting Yang, Ping Zhang, Mengfan Zheng, Nan Li, Shuai Ma



Technologies



Performance Requirements and Evaluation Methodology for AI and Communication in 6G 64

Gongzheng Zhang, Jian Wang, Rong Li, Yan Chen, Jiafeng Shao, Hui Lin, Jun Wang, Jianglei Ma, Peiying Zhu

Data Plane Design for AI-Native 6G Networks 73

Xueqiang Yan, Xinran Zhang, Junfan Wang, Yi Zhang

In-Network Learning for Distributed RAN AI 83

Distributed LLMs via Latent Structure Distillation

Abdellatif Zaidi, Romain Chor, Piotr Krasnowski, Milad Sefidgaran, Rong Li, Fei Wang, Chenghui Peng, Shaoyun Wu, Jean-Claude Belfiore

A Novel Federated Learning Method for Distributed Generation of 6G Air Interface Data 92

Mingfeng Xu, Yang Li, Wei Zhou, Hui Liu, Jiamo Jiang

Service Quality Assurance for 6G Networks with Native AI 98

Guangyi Liu, Kaiyue Wang, Juan Deng, Jiajun Wu, Huanran Hu, Guanchen Lin

Joint Orchestration and Management of Multidimensional Resources in 6G Intelligent Endogenous Networks 109

Dong Wang, Jianzhang Guo

An Exploration on 6G-oriented Transmission and Reception Solutions for Non-Orthogonal Superimposed Pilots 120

Han Xiao, Wenqiang Tian, Xufei Zheng, Wendong Liu, Jia Shen

Practices



Semantic Digital Twins: Enhancing Performance in Wireless Communication and LLM Inference 128

Peiyao Chen, Yiqun Ge, Qifan Zhang, Wuxian Shi, Zheyuan Wei

Digital Twin Online Channel Model: Vision, Progress, and Challenges 135

Junling Li, Weitian Zhang, Chengxiang Wang, Chen Huang

AI-Driven Innovations in RF and Antenna Design 144

Guangjian Wang, Lingjun Yang, Jimmy Jian, Chandan Roy, Li Pan, Guolong Huang, Hua Cai, Wen Tong

Physics-inspired Intelligent Communication: Opportunities, Advances, and Trends 158

Ziqing Xing, Ridong Li, Zirui Chen, Zhaohui Yang, Zhaoyang Zhang

Robots Empowered by AI Foundation Models and the Opportunities for 6G 168

Massimiliano Maule, Anh Vu Vu, Hanwen Cao, Tingzhong Fu, Mohamed Gharba, Daniel Gordon, Joseph Eichinger, Shenfei Zhang, Yiqun Wu, Xueli An, Lei Lu

LLM Application in Wireless Communication Knowledge Management 179

Hongwei Hou, Chixiang Ma, Lihong Du, Junhui Li



Machine Learning in Communications: A Road to Intelligent Transmission and Processing

Shixiong Wang, Geoffrey Ye Li

Abstract

Prior to the era of artificial intelligence (AI) and big data, research into wireless communications primarily followed a conventional route involving problem analysis, model building and calibration, algorithm design and tuning, and holistic and empirical verification. However, this methodology often faced limitations when dealing with large-scale and complex problems and managing dynamic and massive data, resulting in inefficiencies and limited performance of traditional communication systems and methods. As such, wireless communications have embraced the revolutionary impact of AI and machine learning (ML), leading to the development of more adaptive, efficient, and intelligent systems and algorithms. This technological shift paves the way to intelligent information transmission and processing. This paper discusses the typical roles of ML in intelligent wireless communications, as well as its features, challenges, and practical considerations.

Keywords

machine learning, intelligent transmission, intelligent processing

1 Introduction

Since the 19th century, radio communications have started a new era of information transmission for human society. The early stages of radio transmission technology, such as Morse codes and telegraph machines, relied heavily on manual operations, which limited the efficiency and reliability of information exchange. Aiming to automate the task of information transmission and processing at a sophisticated level, the first-generation concept of "intelligent transmission and processing" emerged. Subsequently, the 20th century witnessed significant developments in module-based communication systems. These systems encompass essential modules such as source coding, channel coding, modulation, transmit beamforming, wireless channel transmission, receive beamforming, demodulation, signal detection, channel decoding, and source decoding [1–3]. Methodologically, the development of module-based wireless communications and signal processing follows a systematic research trajectory that includes problem analysis, model development and calibration, algorithm design and optimization, and empirical validation, feedback, and improvement. Notably, every step in this methodological loop demands large volumes of human intellectual endeavors.

In the 21st century, wireless communication systems are expected to deliver extremely vast amounts of data in various formats, such as audio, video, and text, while ensuring low latency, high data rates, and reliability. Furthermore, the incorporation of new network topologies (e.g., internet-of-things networks, unmanned-aerial-vehicle relay networks) and cutting-edge functions (e.g., integrated sensing and communications [ISAC], integrated computing and communications [ICAC]) has added complexity to the design of modern communication systems. This complexity is particularly evident in the following three aspects:

- Addressing different types of modeling uncertainties in not only holistic systems but also individual modules
- Leveraging various forms of big data generated by user equipment and base stations
- Solving challenging algorithmic problems in realizing the networks

Traditional design methods rely heavily on the intensive human intellectual efforts and are proven inadequate for managing large-scale and complex issues and handling dynamic extensive data. This inadequacy results in inefficiencies and limited performance of information

transmission and processing. In response, wireless communications and signal processing have embraced the transformative potential of artificial intelligence (AI) and machine learning (ML) [4]; for comprehensive surveys of ML on communications, see [5–11]. This technological and methodological shift has enabled the development of more adaptive, efficient, robust, and intelligent systems and algorithms. Consequently, the second-generation concept of "intelligent transmission and processing" is emerging, aiming to significantly reduce the need for human intellectual efforts and improve the integrated performances of communication systems.

Figure 1 illustrates the philosophical connotations of intelligent transmission and processing. A technical visualization of ML-empowered intelligent transmission and processing is shown in Figure 2.

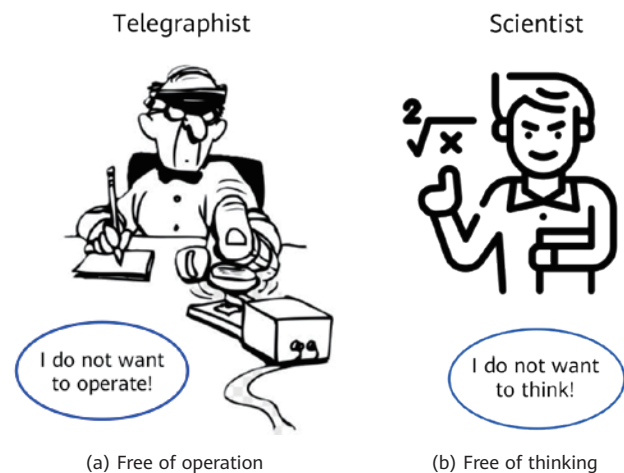


Figure 1 Connotations of intelligent transmission and processing (Icon credit: CLEANPNG.com and FLATICON.com)

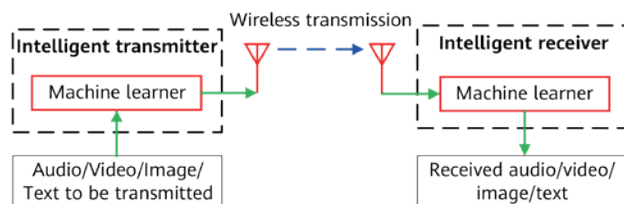


Figure 2 An end-to-end structure of intelligent transmission and processing systems. The intelligent transmitter and receiver act as end-to-end information processors f , where each f is learned by machines from data; the transmitter and receiver can automatically adapt to the real-time characteristics of wireless channels. The intelligence is reflected in the sense that human intellectual efforts are no longer explicitly required to study large-scale, dynamic, and uncertain information transmission mechanisms and processing solutions.

To showcase the power of ML techniques in enabling intelligent transmission and processing, this paper reviews trending ML applications in communication systems and

methods, including physical-layer communications [6, 12], semantic communications [13, 14], resource allocations in communications [15], ICAC [16], and ISAC [17]. However, the ambition of this paper is not to offer an exhaustive list of all existing works in the area. Rather, we aim to illuminate the path towards intelligent transmission and processing.

Although ML has the potential to reform the theory and practice of wireless communications, the challenges and disadvantages of utilizing ML-based approaches accompany its opportunities and advantages [4], for example, the reliability issue due to the lack of interpretability of black-box learning methods (e.g., deep learning), the generalization issue due to the limited training data and the non-stationarities of the underlying data-generating laws, and the resource deficits in training and storing large ML models (e.g., deep learning); see Figure 3 for a motivational understanding. In addition to the three primary challenges exemplified, other instances may also arise, e.g., the scalability issue caused during the reconfiguration of system topology or hardware (e.g., removing or adding antennas; which can be seen as a kind of generalization problem) and the security and privacy issue in networked learning [16]. The main message is that in advancing communication theories and developing communication systems, the role of ML should not be overstated: ML (especially data-driven deep learning) can be a valuable factor to consider rather than an absolute rule to follow; problem analyses and mechanism modeling are always important; see [18–20] for technical investigations and justifications; see also the example below for a motivational understanding.

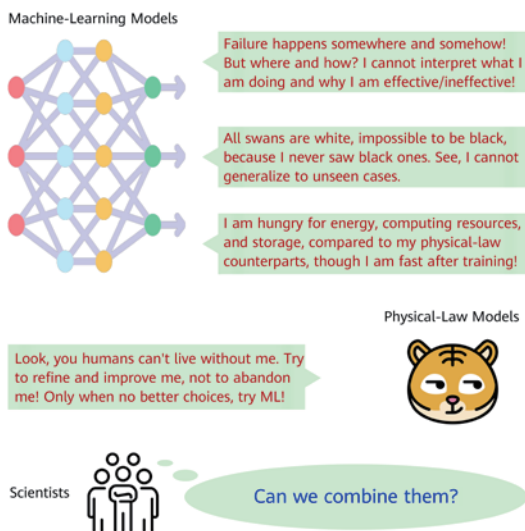


Figure 3 To choose an ML model or a physical-law model, this is a question! Nothing is free, although some are cheap! Choose the one that fits your situation and expectations well! Whenever possible, combine them to improve the overall system performance. (Icon credit: FLATICON.com)

Example (Ice Cream Sales and Shark Attacks): Regression analysis using historical data shows a positive relationship between ice cream sales and shark attacks, which is illogical. However, the primary factor driving this correlation is temperature: higher temperatures lead to increased ice cream sales and beach attendance; more beach visitors result in more shark attacks [21]. Hence, mechanism modeling is vital.

Before digging into ML applications in communications, we quickly review the essentials of ML concepts and methods in Section 2, especially those of trustworthy ML. The aim is to highlight the primary considerations, including philosophical and technical facets, of using ML in wireless communications.

2 Machine Learning Concepts and Methods

ML is concerned with discovering hidden information and patterns from data. The primary advantage is its ability to explain data automatically, thus freeing humans from studying the underlying data-generating mechanisms. This feature inherently enables machine intelligence in the practice of communications, specifically, in the transmission and processing of information [5, 7, 8, 11, 22].

Depending on the characteristics of tasks, ML can be categorized into four genres: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Mathematically, the key to all ML tasks is to find a function f , called a hypothesis, that maps the observed data to a desired decision; see Figure 4 for a conceptual illustration. Specific examples of ML are as follows.

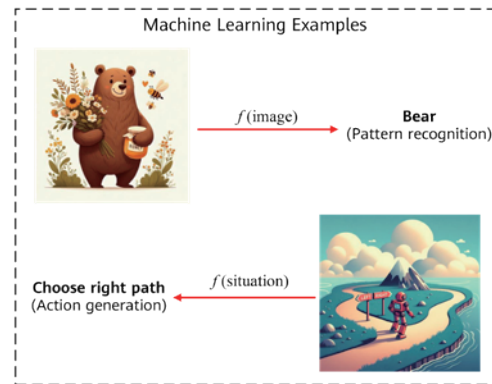


Figure 4 Conceptual illustration of ML. ML is to find a mapping f from the input data to a decision. The upper example is a supervised ML problem where an image classifier f recognizes the image as a bear. The lower example is a reinforcement learning problem where an action generator f recommends the robot to choose the right path in the current situation. (The two images are generated by Microsoft Copilot.)

- **Supervised Learning:** Supervised learning summarizes the hidden information of labeled data and includes regression and classification as two main types of tasks. Given the deterministic or random variable pair (x, s) where x is the feature vector and s is the continuous-valued expected response, regression aims to find a functional relation f from x to s such that predicted label $f(x)$ can be as close as possible to the target label s . The closeness is measured by a loss function L through $L[s, f(x)]$, for example, the mean-squared error $[s - f(x)]^2$. In handling the deterministic variable pair (x, s) , there potentially exists a function f such that $f(x)$ is exactly equal to s for every realization of (x, s) , i.e., $L[s, f(x)] \equiv 0$. As for the random variable pair (x, s) , however, the exact equality cannot be generally guaranteed. Instead, the loss is calculated under the joint distribution $\mathbb{P}_{x, s}$ of (x, s) , for example, the expectation of the loss $\mathbb{E}_{(x, s) \sim \mathbb{P}_{x, s}} L[s, f(x)]$. When s is one-dimensional and takes discrete values, we have the classification problems where $L[s, f(x)]$ is defined by, e.g., the indicator function $\mathbb{I}\{s \neq f(x)\}$. In this case, s indexes different target classes; (e.g., for binary classification, $s \in \{-1, 1\}$).
- **Unsupervised Learning:** In unsupervised learning, there are no labels s for the collected data, and only feature data x is present. Therefore, we focus on discovering the hidden information from the realizations of the datum variable x . Clustering is a typical task of unsupervised learning. Different from classification which predicts the categorical labels s of new data points x based on a training dataset with known labels, clustering aims at grouping similar data points x together based on their features without predefined categorical labels. In summary, clustering is to find a function f that maps data x into a suitable group. Feature transformation is another example of unsupervised learning, which transforms original feature data x into another feature space using a learned mapping f ; Specifically, $y = f(x)$ —compare this with a time-domain signal x and its Fourier transform y . Autoencoders, a type of artificial neural network, provide an excellent example of feature transformation through their encoding and decoding operations. Yet another important example of unsupervised learning is distribution estimation, i.e., to estimate the data-generating distribution that best fits (or describes) the collected data. Distribution estimation is particularly vital in generative tasks such as producing a new sample based on collected samples; for example, given a group of cat images, determining how to produce a new cat image by drawing from the fitted distribution?
- **Semi-supervised Learning:** Semi-supervised learning can be considered a variant of supervised learning as it extends the principles of supervised learning by incorporating a mixture of labeled data (x, s) and unlabeled data x' . While supervised learning relies entirely on labeled data to train the model, semi-supervised learning aims to improve model performance and generalization capabilities by leveraging the additional unlabeled data. The semi-supervised approach is particularly useful when acquiring large amounts of labeled data is expensive or time-consuming, while unlabeled data is abundant and easy to obtain. By leveraging the information from the unlabeled data along with that from the labeled data, semi-supervised learning can find a model f that has better prediction performance (of the label s associated with the data x) compared to purely supervised learning methods that rely solely on labeled data.
- **Reinforcement Learning:** Reinforcement learning is concerned with decision-making problems in a dynamic and uncertain environment. Unlike supervised learning, which uses labeled data, and unsupervised learning, which finds patterns in unlabeled data, reinforcement learning involves the agent interacting with the environment, receiving feedback in the form of rewards or penalties, and using this feedback to learn optimal behaviors or strategies over time. To be specific, the agent autonomously learns to make decisions in an environment by performing actions a , in response to current states s , in order to maximize the cumulative reward. Therefore, mathematically, an action-generating function f from state s to action a needs to be learned.

For specific applications of the four ML genres in wireless communications, refer to [8, 23].

Data-Driven and Model-Driven Learning: Considering the degree of human intelligence and domain knowledge involved, ML can be classified into data-driven and model-driven approaches. Data-driven ML relies entirely on historical data and does not involve any analysis of underlying data-generating mechanisms. In contrast, model-driven ML incorporates, to varying extents, studying the underlying physical mechanisms and data-generating models. Intelligent information transmission and processing can benefit, in terms of improving overall performance, from the collaboration between communication-systems modeling and big data discovery [12, 23]. For example, in

signal detection, suppose that we have T pilot data pairs $\{(s_1, x_1), (s_2, x_2), \dots, (s_T, x_T)\}$ where x_i are the received signals and s_i are the transmitted symbols, for $i = 1, 2, \dots, T$. Data-driven ML directly utilizes all the data to train a detector f from x to s . In contrast, model-driven ML first considers the signal-transmission model $x = Hs + v$ where H denotes the channel matrix and v the channel noise, and then finds a detector f based on the above underlying data-generating mechanism. For detailed technical treatments and discussions, see [19, 20, 24, 25].

Hypothesis Space and Deep Learning: To locate a best decision function f , a candidate function space \mathcal{H} (called hypothesis space) from which f is drawn, needs to be specified. To clarify further, for instance, supervised statistical ML can be formulated as:

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, \mathbf{s}) \sim \mathbb{P}_{\mathbf{x}, \mathbf{s}}} L(\mathbf{s}, f(\mathbf{x})),$$

where the joint distribution $\mathbb{P}_{\mathbf{x}, \mathbf{s}}$, which is unknown in practice, can be estimated using collected historical data (e.g., using empirical distribution). As an example, signal detection problems can be characterized as described earlier, where f is a detector, x is the antenna-received signal, and s is the transmitted symbol (e.g., constellation points) [20]; the loss function L can be mean-squared error or symbol-error rate. Canonical examples for hypothesis space \mathcal{H} are as follows.

- **Linear Function Space:** \mathcal{H} only includes the linear transforms of input x . In the signal detection case, \mathcal{H} contains only linear detectors.
- **Reproducing Kernel Hilbert Space:** \mathcal{H} includes all linear transforms of the nonlinearly-lifted-feature $\varphi(x)$ of the original feature x , using a feature mapping function φ . In essence, \mathcal{H} contains some specific types of nonlinear functions of input x .
- **Neural Network Function Space:** \mathcal{H} is represented (or structured, characterized) by neural networks, for instance, multi-layer perceptron, recurrent neural networks, convolutional neural networks (CNNs), radial basis neural networks, autoencoders, or transformers. Each given neural network defines a special type of function space \mathcal{H} . When the employed neural network has deep structures with many hidden layers being included, \mathcal{H} denotes a deep-neural-network function space. Upon operating with deep neural networks, ML is referred to as deep learning.

On the other hand, with the involvement of domain knowledge and expert designs, a hypothesis space \mathcal{H} can be accordingly adapted or tailored to a domain-specific

problem [12, 18, 23]. Therefore, model-driven ML is to devise an ad-hoc and structured candidate space \mathcal{H} , by leveraging known problem characteristics and data-generating mechanisms.

Explainability, Reliability, and Sustainability: Modern ML research addresses several advanced concerns, including explainability, reliability, and sustainability of learning models [26, 27]. Explainable ML seeks to make learning models transparent, interpretable, and accountable through techniques such as feature engineering and physical modeling [28]; model-driven ML, which leverages underlying physical data-generating mechanisms, can be seen as such a scheme [12, 18]. Reliable ML focuses on creating robust and accurate learning models that generalize well to new data (that are not used in the training stage), tackling issues such as overfitting, generalization, knowledge migration, and limited-sample learning [20, 29–31]. Sustainable ML aims to develop learning models with minimal negative impact on the environment and society, addressing energy efficiency, privacy and security, and fairness and bias [16, 32]. In the context of intelligent information transmission and processing, the three considerations (i.e., explainability, reliability, and sustainability) are of natural importance and significance. Therefore, they are the primary considerations in developing ML-based solutions for wireless communications.

Centralized and Distributed Learning: ML models can be trained using various approaches depending on the structure of the data distribution and the architecture of the computation. Two primary paradigms in this context are centralized learning and distributed learning [33, 34]. Centralized learning involves collecting and storing all training data in a single central location, such as a data center or cloud server, and the ML model is trained on this aggregated dataset. Distributed learning, on the other hand, involves training ML models in a distributed manner across multiple devices (or nodes), each of which holds a portion of the data. A prime example of distributed learning is federated learning, where multiple clients (e.g., smartphones, internet-of-things devices, or different organizations) collaboratively train a model without sharing their local data. Instead, each client trains the model on its local data and only shares the model updates (gradients or weights) with a central server, which aggregates these updates to form a global model. Both centralized and distributed learning methods are beneficial for advancing future-generation communication systems because they can adapt to diverse modern communication network typologies.

3 Physical Layer Communications

Physical layer communications aim to reliably transmit raw data streams, e.g., binary bits, through physical mediums. Figure 5 presents a traditional architectural diagram of wireless communications, featuring various functional modules (or blocks) that are meticulously designed by humans in accordance with fundamental mathematical and physical principles. This block-based diagram is structurally different from the ML-empowered architectural diagram shown in Figure 2, where interconnected functional modules are taken over by end-to-end operating parts.

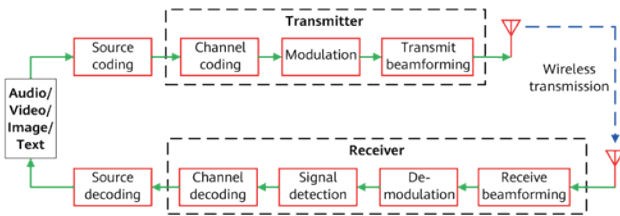


Figure 5 The module-based structure of traditional transmission and processing systems. Every block acts as an information processor f , where f is elaborately designed by scientists based on underlying physical mechanisms and mathematical laws.

In addition to the highly integrated (i.e., highly intelligent) structure in Figure 2, ML-based transmission and processing systems can also be partially intelligentized. For example, in one scenario, only the channel coding or decoding block is managed by ML, meaning that the channel coding scheme is designed by machines rather than information scientists. In another scenario, ML is used solely for the transmit beamformer or the receive beamformer. A conceptual illustration is shown in Figure 6.

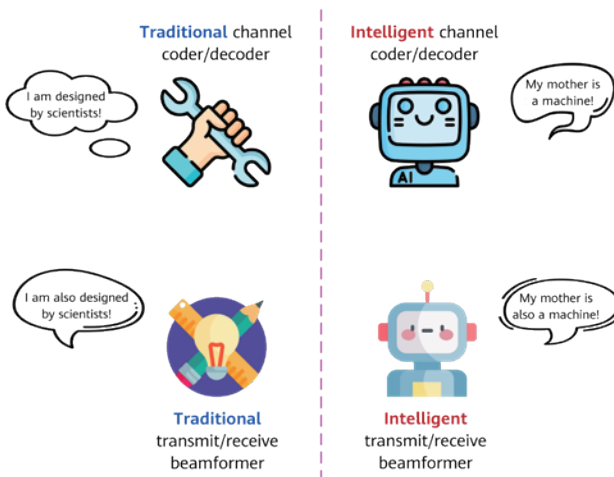


Figure 6 In contrast to the highly integrated structure in Figure 2, each individual module in Figure 5 can be augmented by ML. (Icon credit: FLATICON.com.)

In technical details, applications of ML in physical layer communications include overall end-to-end system design [35, 36] (Figure 2) and individual module design (Figure 6). The latter, to be specific, encompasses:

- coding/decoding techniques, for example, source coding [37], channel coding [38, 39], and joint source-channel coding (JSCC) [40, 41]
- signal modulation and detection [25, 42]
- transmit and receive beamforming [20, 43–46], for example, beam alignment and beam tracking [47–50]
- channel estimation and feedback [51, 52]

among many others. For comprehensive and recent surveys, see [6, 53, 54].

Coding and decoding techniques are vital in digital communications, ensuring efficient and reliable data transmission. Recently, ML has been increasingly applied to enhance these techniques, encompassing source coding, channel coding, and JSCC. Traditionally, source coding (i.e., data compression) reduces data redundancy for efficient transmission and storage. ML techniques, such as neural-network-based autoencoders, have revolutionized this field. Autoencoders learn efficient representations of data by encoding it into a lower-dimensional space and then reconstructing it, achieving high compression rates with minimal loss of information [55]. Channel coding adds redundancy to data in order to detect and correct transmission errors caused by noisy channels. ML models, particularly deep learning techniques, have been applied to develop novel error correction codes. For example, neural decoders have been designed to decode complex schemes like low-density parity-check (LDPC) [56] and Turbo codes [57], offering improved performance over traditional algorithms, especially in highly noisy environments. JSCC integrates source and channel coding to optimize overall system performance. ML models, such as variational autoencoders (VAEs) [58], CNNs [59], and generative adversarial networks (GANs) [60], are used to jointly learn the representation and error correction codes. These models can adapt to the characteristics of both the source and the channel, achieving better compression and error resilience than traditional methods. Overall, ML-based coding and decoding techniques represent a significant advancement in digital communications. By leveraging the predictive and adaptive capabilities of ML, these techniques enhance data compression, error correction, and overall transmission efficiency. This lays the foundation for creating communications systems that are more robust and efficient.

Signal modulation and detection, which enable the transmission and interpretation of data over various channels, are fundamental processes in digital communications. Recently, ML techniques have been applied to enhance these processes, improving efficiency and reliability. Modulation involves altering a carrier signal's properties, such as amplitude, frequency, or phase, to encode information. Traditional modulation schemes include amplitude modulation (AM), frequency modulation (FM), and phase shift keying (PSK). ML techniques, particularly deep learning models, are now used to design adaptive modulation schemes. These models can dynamically adjust modulation parameters based on the channel conditions, optimizing performance in real time. For instance, neural networks can learn complex modulation patterns that maximize data throughput and minimize error rates [61]. Detection involves demodulating the received signal to recover the transmitted information. Traditional methods rely on predefined algorithms to estimate the transmitted data but often assume specific channel characteristics. ML approaches, such as fully connected deep neural networks [42] and transfer learning [62], have been employed to enhance signal detection. These models can learn from data to accurately detect signals under varying and complex channel conditions, improving robustness against noise and interference. Overall, the integration of ML techniques in signal modulation and detection represents a significant leap forward in communications technology — it enhances data transmission efficiency, resilience to noise, and overall system performance.

In wireless communication systems, transmit and receive beamforming techniques are essential for enhancing signal quality and increasing data throughput. Beamforming directs the transmission or reception of signals in specific directions using antenna arrays, improving signal strength and reducing interference. Recently, ML techniques have significantly advanced beamforming performance and adaptability. For transmit beamforming, traditional methods, such as phased array systems, use predefined algorithms to adjust the phase and amplitude of signals from multiple antennas. In contrast, ML techniques, particularly deep learning models, optimize this process by learning from environmental data. As an example, reinforcement learning can dynamically adjust beamforming patterns in real time based on feedback from the communication environment, enhancing performance in complex and changing scenarios [63, 64]. For receive beamforming, conventional methods, such as minimum variance distortionless response (MVDR) and maximal ratio combining (MRC), rely on statistical

models of the signal environment. ML approaches, such as CNNs, improve upon these by learning optimal beamforming weights directly from data, allowing for more accurate and robust signal reception in diverse and dynamic environments [65, 66]. Beam alignment and tracking are crucial subcategories of beamforming, particularly important in high-frequency bands like millimeter-wave (mmWave) and terahertz communications. These techniques ensure that the transmitter and receiver maintain optimal beam alignment to maximize signal strength and data throughput [49, 50, 67]. Traditional alignment methods rely on exhaustive search or iterative algorithms, which are time-consuming and computationally intensive. ML approaches, such as supervised learning, multi-armed bandits, and reinforcement learning, provide more efficient solutions by predicting optimal beam directions from historical data, significantly reducing the search space. Beam tracking maintains alignment as the transmitter or receiver moves or as the environment changes. ML techniques, particularly deep learning models, enhance tracking by predicting beam direction changes in real time. Recurrent neural networks and long short-term memory networks, which capture temporal dependencies, are particularly effective for this purpose. For technical details on ML-based beam alignment and tracking, see [47–50, 67]. In summary, the integration of ML into beamforming, including beam alignment and tracking, is critical for next-generation networks such as 5G and beyond, because these ML-driven techniques can leverage predictive and adaptive capabilities to enhance signal quality, reduce interference, and optimize system performance.

Channel estimation and feedback techniques are of high importance in wireless communication systems for accurately characterizing the communication channel and ensuring efficient data transmission. These processes involve measuring the channel's properties and providing necessary feedback to transmitters. Recently, ML techniques have been applied to enhance these processes, offering significant improvements in accuracy and efficiency [68, 69]. Channel estimation involves predicting the state of the communication channel to optimize signal transmission and reception. Traditional methods, such as minimum mean-squared error (MMSE), rely on statistical models and require significant computational resources. ML approaches, especially deep learning models, have introduced new ways to perform channel estimation with higher accuracy, and potentially, lower computational complexity. For instance, CNNs can learn to estimate channel states directly from received signal data, providing more robust and adaptive

solutions in complex environments [70]. Long short-term memory networks are particularly effective for capturing temporal dependencies in channel conditions, improving estimation accuracy [71]. Feedback mechanisms forward channel state information (CSI) from the receiver back to the transmitter, allowing for real-time adaptation of transmission setups. Traditional feedback methods often involve quantizing and encoding the CSI, which may incur delays and inaccuracies. ML techniques, such as autoencoders and CNNs, improve feedback efficiency by compressing and reconstructing the CSI with minimal loss of information [52]. This allows for more precise and timely adjustments to transmission setups. In addition, ML models can simultaneously handle channel estimation and feedback, optimizing both processes in an integrated manner [69]. This holistic approach leverages the strengths of ML to enhance overall system performance.

4 Semantic Communications

Semantic communications, unlike conventional physical-layer communications, focus on transmitting semantic information conveyed in original data (e.g., image, text, audio) rather than bit-wise raw information. The primary benefit of semantic communications is that the transmission overloads of wireless channels can be significantly reduced compared to bit-wise transmission. Consequently, the information transmission speed and efficiency can be considerably improved. For comprehensive and recent surveys in semantic communications, see [72–75].

The key to semantic communications is to extract the semantic information from raw data. Therefore, semantic communications can be realized by elegantly designing the source coding and decoding strategies. It can also be actualized through JSCC and decoding. The difficulty, however, is that the semantic information of given raw data is specific to a task (see Figure 7), due to which, a generally well-accepted mathematical analysis, modeling, and computing framework for semantic communications is still lacking; for exploring works in this direction, see [76]. Therefore, for a specified communication task, the semantic coding and decoding schemes need to be elaborated. In the context of intelligent transmission and processing, semantic communications can be implicitly realized in highly integrated end-to-end transceivers, shown in Figure 2.

ML approaches play a pivotal role in semantic communications, enabling systems to understand, process, and convey meaning more accurately. Techniques such as

deep learning models, including transformers, CNNs, and recurrent neural networks, have been widely utilized to analyze and predict the semantic relevance of data, thus optimizing bandwidth usage and improving communication efficiency. To be specific, natural language processing (NLP) algorithms allow for the extraction and interpretation of semantic content from text and audio, facilitating more meaningful data compression and transmission [77, 78]; computer vision methods, on the other hand, enable the extraction and interpretation of semantic meaning from image and video [79].

Recent research has demonstrated the potential of ML-driven semantic communications in various applications. For example, in [13], transceiver neural networks have been designed to directly transmit text semantic meaning, which significantly reduces the demand on communication resources and improves the overall transmission performance. Another example is [80], in which an efficient system for video conferencing is developed to improve transmission efficiency. Most studies in semantic communications focus on JSCC to save communication resources. However, this approach requires changing the existing communication infrastructures and therefore hinders practical implementation. As such, a pragmatic approach to wireless semantic transmission through revising some modules in existing infrastructures is reported [14]. To guarantee semantic transmission reliability and communication efficiency, the spectral efficiency in the semantic domain and the semantic-aware resource allocation issues have been investigated in [81]. In addition to the above representative applications, the synergy between semantic communications and emerging technologies, such as the internet of things (IoT) [82] and edge computing [83], is fostering new opportunities for intelligent and context-aware communication systems. By leveraging distributed ML models, semantic communication systems can dynamically adapt to changing environmental conditions and user requirements, ensuring robust and efficient information exchange [84].

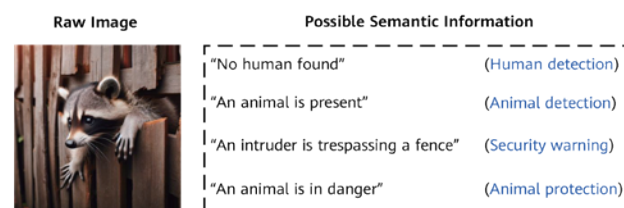


Figure 7 The semantic information of raw data is task-specific. Losslessly transmitting a high-definite image is time- and resource-consuming. However, accurately transmitting a semantic message can be relatively simpler and cheaper. (The image is generated by Microsoft Copilot.)

In summary, semantic communications, underpinned by advanced ML techniques, define a brilliant future direction for communication systems. This innovative approach promises to reform how information is transmitted and understood, offering profound implications for the efficiency and effectiveness of future communication networks.

5 Resource Allocation in Communications

In wireless communications, resource allocation is concerned with how to efficiently manage and utilize spectra, power, computing, space, and time resources, thus improving the overall communication network performance, e.g., higher throughput, lower latency, larger coverage, higher reliability, to name a few [15, 85, 86]. Typical applications encompass link scheduling, message routing, power allocation, channel selection, beamforming, spectra access and management, and division protocol design (i.e., time division, frequency division, etc.). From the mathematical programming perspective, resource allocation is often formulated as optimization problems. From the operations research perspective, assignment and scheduling are two pivotal techniques; the former handles static resource allocation problems, while the latter addresses dynamic ones; the static and dynamic features are with respect to time. From the computational and algorithmic perspective, standard and trending solution frameworks include the following:

- Continuous optimization, discrete (e.g., combinatorial, integer) optimization, and mixed optimization
- Single-objective optimization and multi-objective optimization
- Linear programming and nonlinear programming
- Convex optimization and non-convex optimization
- Smooth optimization and non-smooth optimization
- Min-Max optimization (e.g., game theory, worst-case robust analyses)
- Deterministic programming and stochastic programming (i.e., whether random variables are involved; if involved, associated distributions are considered)
- Single-stage optimization (i.e., static programming) and multi-stage optimization (i.e., dynamic programming)
- Heuristic optimization (e.g., genetic algorithm, particle swarm optimization, simulated annealing)
- Surrogate optimization which is also known as black-box optimization (e.g., Bayesian optimization)

- ML-based optimization (e.g., solution methods based on reinforcement learning and deep learning)

The canonical applications and solution frameworks of resource allocation in wireless communications are shown in Figure 8. For introductory and motivational reading on this topic, refer to [85, 87, 88].

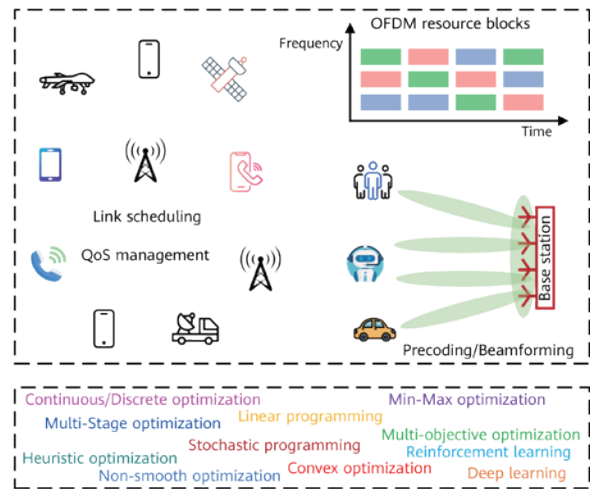


Figure 8 Typical applications and solution frameworks of resource allocation in wireless communications. OFDM: orthogonal frequency division multiplexing; QoS: quality of service. (Icon credit: FLATICON. com.)

Nowadays, the advent of ISAC [89] has slightly changed the connotation of traditional resource allocation. This is because the best resource allocation scheme for communication is not necessarily the same as (or in accordance with) that for sensing; see [90]. For example, the optimal waveforms for communication and sensing are usually dissimilar [91, 92] because the two radio functions have different or even contradicting design preferences. Therefore, diverse resources, including radio, computing, power, time, beam, etc., should be delicately allocated to satisfy the individual performance requirements of communication and sensing. The same dilemma holds for ICAC, e.g., edge computing [93] and networked control* [94], because limited resources need to be elegantly distributed to computing and communication. For further details, see [95, 96].

ML techniques have emerged as powerful tools to address the challenges brought by resource allocation in wireless communications. Recent advancements have demonstrated the potential of ML in various resource allocation tasks. For instance, deep reinforcement learning has been applied to optimize spectrum allocation, power control, and user

*Controllers are, by their nature, information processors and are therefore ad-hoc computing modules.

association in heterogeneous networks, showing significant improvements over conventional methods [22, 97–102]. Similarly, supervised learning algorithms have been used to efficiently solve complex optimization problems in resource allocation such as mixed integer nonlinear programming (MINLP) [103]. In addition, unsupervised learning techniques can also be employed to solve resource allocation problems and refine the solutions, e.g., the graph embedding trick in link scheduling [104].

Traditional resource allocation methods rely heavily on human intellect to build exact models and develop ad-hoc solution methods, which can be suboptimal and even inflexible in dynamic, complex, and large-scale scenarios. ML, particularly deep learning and reinforcement learning, offers the ability to model complex interactions with environments, predict communication-network states, and optimize decisions in a real-time manner, thereby enhancing the overall performance and adaptability of wireless networks [15]. To be specific, communication channels in practice are often time-varying, however, mathematically considering such model uncertainties is not straightforward. This is because we do not exactly know how the channels evolve over time. Even worse, the resultant mathematical programming models are computationally complex, and therefore, hard to be efficiently and optimally solved. The role of ML, in this sense, is to leverage accessible real-world data, discover the hidden knowledge and patterns that the data convey, and automatically find satisfactory resource allocation decisions. In technical details, on the one hand, ML can assist in solving computationally difficult optimization problems because resource allocation optimization can be seen as a mapping from parameters to decisions. This data-to-decision mapping benefits from the powerful function-fitting ability of supervised learning based on deep neural networks, where labeled data-decision pairs are generated by well-behaved artifact solution methods. On the other hand, ML can treat the utility function of a resource allocation problem as the loss function in the training stage. This strategy allows ML to generate high-quality resource allocation decisions without relying on legacy human-made algorithms. In addition to the above two ML schemes in resource allocation, another archetype, called algorithm unrolling [18], employs neural networks to unroll existing efficient iterative algorithms. Specifically, each neural network layer acts as an iteration step of an iterative algorithm. By cascading several layers, an iteration process of the algorithm can be mimicked. This algorithm-instructed archetype is also referred to as model-driven deep learning [12], where the architecture of a deep

neural network is tailored considering domain knowledge, thus improving the generalizability of the network and reducing the required size of the training data set. The fourth promising ML paradigm in resource allocation is to use reinforcement learning to explore unknown and hard-to-model environments (e.g., complex and dynamic physical transmission channels). By interacting with environments, intelligent resource allocation solutions can be learned. The four typical roles of ML in resource allocation are summarized in Figure 9. The first benefit of using ML methods is their fast computing speed in the running stage, although the training stage might be computationally heavy (when compared with the first three schemes). The second benefit of using ML methods is the ability to respond to dynamic and uncertain (even unknown) environments without explicit physical modeling (when compared with the fourth scheme, i.e., the reinforcement learning scheme).

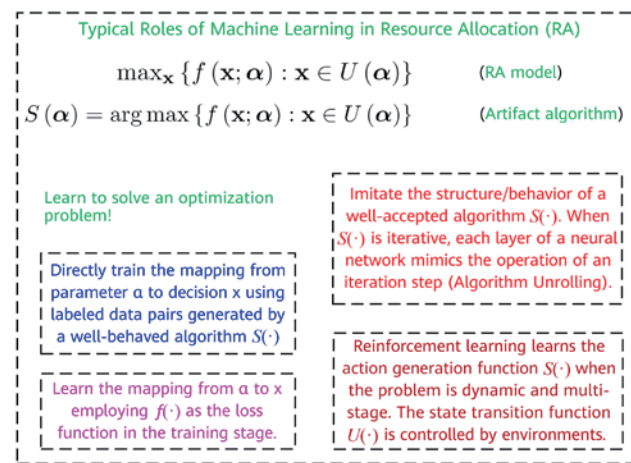


Figure 9 Four typical roles of ML in resource allocation —learning to solve optimizations.

6 Beyond Data Transmission: Sensing and Computing

Wireless communication systems are undergoing transformative changes driven by the increasing demand for low-latency and high-speed connectivity, the growing need for sensing abilities (e.g., to localize and track users) assisting high-performance communications, and the proliferation of connected devices enabling collaborative computing. This evolution has led to the development of innovative system paradigms such as ISAC [89, 105] and ICAC [106, 107], which aim to unify traditionally disparate functionalities to optimize resource usage, reduce hardware costs, and enhance overall system capabilities. For example, environmental and users' sensory data can be

utilized to enhance communication performance through beam management and resource allocation, while sharing sensing data across network nodes enables real-time network monitoring and situational awareness for better sensing accuracy and larger coverage. Another example is local data processing at the edge, which can reduce latency for real-time communications, while high-speed communications enable efficient distributed computing for large-scale data analytics. As discussed in previous sections, ML (especially deep learning) techniques are indispensable in modern communication systems. These techniques offer sophisticated algorithms that can learn from vast amounts of data, and can therefore, optimize various aspects of communication networks, including resource allocation, signal processing, and fault detection. These benefits are also applicable to emerging ISAC and ICAC systems. In ISAC, deep learning models, such as CNNs and transformers, can improve sensing accuracy and robustness [108], while realizing semantic information transmission [109]. In ICAC, ML algorithms, such as federated learning, can protect users' data privacy and optimize computational tasks, facilitating efficient data processing and communication [110, 111]. In short, the synergy between ML/deep learning and the developing integrated paradigms enables more intelligent, adaptive, and efficient communication systems.

6.1 Integrated Sensing and Communication

ISAC is a paradigm that merges sensing and communication functionalities into a single system, leveraging shared infrastructure and spectral resources. This integration is essential in applications where both capabilities are crucial, such as autonomous vehicles, smart cities, and advanced surveillance systems. ISAC enhances the efficiency and performance of these systems by enabling simultaneous data acquisition and communication, thus reducing hardware costs and spectral congestion. However, this integration complicates the design of communication waveforms, the allocation of system and hardware resources, interference management, and overall network operations [112, 17]. These challenges drive the need for novel approaches to unlock the potential of ISAC systems in real-world applications. ML and deep learning techniques are, therefore, pivotal in ISAC, providing advanced data processing and decision-making capabilities. For comprehensive and motivational surveys on ML for ISAC, refer to [112, 17].

6.2 Integrated Computing and Communications

ICAC represents the convergence of computing and communication functionalities, aiming to meet the increasing computational demands of modern applications while maintaining high and robust communication performance. This integration is driven by the necessity to handle massive data processing tasks close to the source — handling tasks closer to the source helps reduce latency and improve efficiency of communications in edge computing environments, and enables intelligence of all connected devices. ICAC, which facilitates real-time data processing and analytics, is essential for applications like industrial automation, virtual reality, and the internet of things. Typical examples of ICAC include edge computing, federated learning, pervasive computing, fog computing, internet of things/vehicles, and autonomous systems. ML and deep learning are integral to ICAC, enabling dynamic resource allocation, adaptive system configurations, and real-time information analytics. These techniques ensure that computing and communication resources are utilized optimally, providing enhanced performance and responsiveness. For comprehensive and motivational surveys on ML for ICAC, refer to [16, 113, 114]. Note that, swarm intelligence and network control [94] are closely related to ICAC because controllers are, in nature, information processors (mapping the system's state signals to the system's control input signals). They are therefore ad-hoc computing modules.

7 Discussions and Conclusions

This paper discusses several pivotal aspects where ML can reform wireless communications, including but not limited to physical-layer communications, semantic communications, resource allocation, ISAC, and ICAC (e.g., federated learning, edge computing). These applications demonstrate ML's potential to upgrade various facets of communication systems, ranging from signal processing algorithms to overall network management. Nevertheless, the adoption of ML in communications is not without challenges and its role should not be overstated. Issues, such as the interpretability and troubleshooting of ML models, the need for large and rich training datasets, and the high computational resources (e.g., power, processing speed) required for training and deployment, must be addressed. In addition, particular focus should be given to the reliability and security of ML-

based systems, especially in scenarios where data privacy (e.g., federated learning [115]), data freshness (e.g., few-shot learning [116, 117]), and real-time decision-making (e.g., autonomous driving) are critical. To address these challenges, the hybrid methodology, which combines the strengths of traditional physical-law models with emerging data-driven ML models, is advocated. Such a synergistic strategy can leverage the reliability and interpretability of physical mechanisms while harnessing the adaptability and learning capabilities of ML, thus enhancing overall communication system performance; see Figure 10 for features, challenges, and future considerations of intelligent transmission and processing. Among all the challenges that we can imagine, the following three items are crucial in real-world operations because they are the minimum requirements for implementing ML-based communications systems:

- How do we interpret the performance gains and failures of machine-learned models, and how do we troubleshoot and repair failures when systems are down, thus improving the overall reliability of systems? In this sense, the paradigm in Figure 6 is more reliable and manageable than that in Figure 2.
- How do we use practically limited data for better generalization and how do we integrate newly available data to improve the generalization capability [20, 31]? This consideration also includes determining how to quickly adapt the learned model to new data, for example, when the environment's data-generating laws change over time [62, 117]. In ML terminologies, data freshness, sample efficiency, and data-distributional robustness are closely related to this issue.
- How do we build domain-knowledge-informed ML models (beyond general-purpose deep neural networks such as multi-layer perception) and design computationally efficient training algorithms (beyond popular stochastic gradient descent) to diminish response times and power consumption [12, 108, 118, 119]? In addition, how do we reduce the model sizes (especially those of deep neural networks) to save storage space [120]? The three considerations above are particularly vital for embedded and edge devices.

In summary, the convergence of ML and communication systems marks a significant technological advancement, which offers the possibility for more intelligent, efficient, and reliable communication networks.

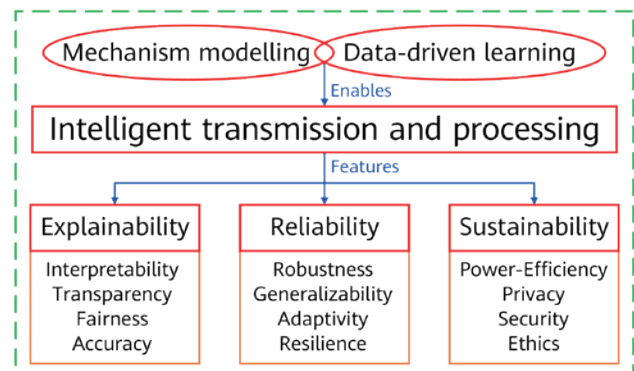


Figure 10 Features, challenges, and considerations of intelligent transmission and processing (in Figure 3) Although data-driven ML is powerful, mechanism modeling (including discovering physical/mathematical laws) is always important to improve explainability, reliability, and sustainability.

References

- [1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.
- [3] G. L. Stüber and G. L. Steuber, *Principles of Mobile Communication*, 4th ed. Springer, 2017.
- [4] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.
- [5] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2016.
- [6] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 93–99, 2019.
- [7] D. Gündüz, P. De Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. Van der Schaar, "Machine learning in the air," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2184–2199, 2019.
- [8] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472–1514, 2020.
- [9] W. Yu, F. Sotiriou, and T. Jiang, "Role of deep learning in wireless communications," *IEEE BITS the Information Theory Magazine*, vol. 2, no. 2, pp. 56–72, 2022.
- [10] A. Alhammedi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, and S. A. Saad, "Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 8845070, 2024.
- [11] A. Celik and A. M. Eltawil, "At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence," *IEEE Open Journal of the Communications Society*, 2024.
- [12] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.
- [13] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.
- [14] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic transmission via revising modules in conventional communications," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 28–34, 2023.
- [15] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep learning-based wireless resource allocation with application to vehicular networks," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 341–356, 2019.
- [16] M. A. Ferrag, O. Friha, B. Kantarci, N. Tihanyi, L. Cordeiro, M. Debbah, D. Hamouda, M. Al-Hawawreh, and K.-K. R. Choo, "Edge learning for 6G-enabled internet of things: A comprehensive survey of vulnerabilities, datasets, and defenses," *IEEE Communications Surveys & Tutorials*, 2023.
- [17] S. Lu, F. Liu, Y. Li, K. Zhang, H. Huang, J. Zou, X. Li, Y. Dong, F. Dong, J. Zhu et al., "Integrated sensing and communications: Recent advances and ten open challenges," *IEEE Internet of Things Journal*, 2024.
- [18] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.
- [19] N. Shlezinger and T. Routtenberg, "Discriminative and generative learning for the linear estimation of random signals [lecture notes]," *IEEE Signal Processing Magazine*, vol. 40, no. 6, pp. 75–82, 2023.

- [20] S. Wang, W. Dai, and G. Y. Li, "Distributionally robust receive beamforming," *arXiv preprint arXiv:2401.12345*, 2024.
- [21] G. James, D. Witten, T. Hastie, R. Tibshirani et al., *An Introduction to Statistical Learning*, 2nd ed. Springer, 2021.
- [22] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3072–3108, 2019.
- [23] Y. C. Eldar, A. Goldsmith, D. Gündüz, and H. V. Poor, *Machine Learning and Wireless Communications*. Cambridge University Press, 2022.
- [24] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.
- [25] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1702–1715, 2020.
- [26] B. Thuraisingham, "Trustworthy machine learning," *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 21–24, 2022.
- [27] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.
- [28] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Leanpub, 2020.
- [29] K. Kawaguchi, Z. Deng, K. Luh, and J. Huang, "Robustness implies generalization via data-dependent generalization bounds," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 866–10 894.
- [30] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and S. Y. Philip, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.
- [31] S. Wang and H. Wang, "Distributional robustness bounds generalization errors," *arXiv preprint arXiv:2212.09962*, 2024.
- [32] A. Van Wynsberghe, "Sustainable AI: AI for sustainability and the sustainability of AI," *AI and Ethics*, vol. 1, no. 3, pp. 213–218, 2021.
- [33] S. AbdulRahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5476–5497, 2020.
- [34] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3579–3605, 2021.
- [35] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3133–3143, 2020.
- [36] H. Ye, G. Y. Li, and B.-H. Juang, "Deep learning based end-to-end wireless communication systems without pilots." *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 3, pp. 702–714, 2021.
- [37] S. Manouchehri, J. Haghghat, M. Eslami, and W. Hamouda, "A delay-efficient deep learning approach for lossless turbo source coding," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6704–6709, 2022.
- [38] H. Ye, L. Liang, and G. Y. Li, "Circular convolutional autoencoder for channel coding," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.
- [39] Y. Zhang, H. Wu, and M. Coates, "On the design of channel coding autoencoders with arbitrary rates for ISI channels," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 426–430, 2021.

- [40] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Deep joint source-channel coding for wireless image retrieval," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5070–5074.
- [41] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source-channel coding for wireless image transmission with OFDM," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.
- [42] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, 2017.
- [43] S. Mohammadzadeh, V. H. Nascimento, R. C. de Lamare, and N. Hajarolasvadi, "Robust beamforming based on complex-valued convolutional neural networks for sensor arrays," *IEEE Signal Processing Letters*, vol. 29, pp. 2108–2112, 2022.
- [44] A. M. Elbir, K. V. Mishra, M. R. B. Shankar, and B. Ottersten, "A family of deep learning architectures for channel estimation and hybrid beamforming in multi-carrier mm-wave massive MIMO," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 642–656, 2022.
- [45] D. d. S. Brilhante, J. C. Manjarres, R. Moreira, L. de Oliveira Veiga, J. F. de Rezende, F. Müller, A. Klautau, L. Leonel Mendes, and F. A. P. de Figueiredo, "A literature survey on AI-aided beamforming and beam management for 5G and 6G systems," *Sensors*, vol. 23, no. 9, p. 4359, 2023.
- [46] N. Shlezinger, M. Ma, O. Lavi, N. T. Nguyen, Y. C. Eldar, and M. Juntti, "Artificial intelligence-empowered hybrid multiple-input/multiple-output beamforming: Learning to optimize for high-throughput scalable MIMO," *IEEE Vehicular Technology Magazine*, 2024.
- [47] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, "Deep learning-based beam tracking for millimeter-wave communications under mobility," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7458–7469, 2021.
- [48] F. Sahrabi, T. Jiang, W. Cui, and W. Yu, "Active sensing for communications by learning," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1780–1794, 2022.
- [49] Y. Wei, Z. Zhong, and V. Y. Tan, "Fast beam alignment via pure exploration in multi-armed bandits," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3264–3279, 2022.
- [50] W. Yi, W. Zhiqing, and F. Zhiyong, "Beam training and tracking in mmWave communication: A survey," *China Communications*, 2024.
- [51] Q. Hu, F. Gao, H. Zhang, S. Jin, and G. Y. Li, "Deep learning for channel estimation: Interpretation, performance, and comparison," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2398–2412, 2020.
- [52] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2827–2840, 2020.
- [53] B. Ozpoyraz, A. T. Dogukan, Y. Gevez, U. Altun, and E. Basar, "Deep learning-aided 6G wireless networks: A comprehensive survey of revolutionary PHY architectures," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1749–1809, 2022.
- [54] N. Ye, S. Miao, J. Pan, Q. Ouyang, X. Li, and X. Hou, "Artificial intelligence for wireless physical-layer technologies (AI4PHY): A comprehensive survey," *IEEE Transactions on Cognitive Communications and Networking*, 2024.
- [55] Y. Yang, S. Mandt, L. Theis et al., "An introduction to neural data compression," *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 2, pp. 113–200, 2023.
- [56] S. Han, J. Oh, K. Oh, and J. Ha, "Deep-learning for breaking the trapping sets in low-density parity-check codes," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 2909–2923, 2022.
- [57] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Turbo autoencoder: Deep learning based channel codes for point-to-point communication channels," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

- [58] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1182–1192.
- [59] E. Boursoulatzé, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.
- [60] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.
- [61] E. Bobrov, D. Kropotov, H. Lu, and D. ZaeV, "Massive MIMO adaptive modulation and coding using online deep learning algorithm," *IEEE Communications Letters*, vol. 26, no. 4, pp. 818–822, 2021.
- [62] N. Van Huynh and G. Y. Li, "Transfer learning for signal detection in wireless networks," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2325–2329, 2022.
- [63] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1581–1592, 2019.
- [64] M. Chu, A. Liu, V. K. Lau, C. Jiang, and T. Yang, "Deep reinforcement learning based end-to-end multiuser channel prediction and beamforming," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 271–10 285, 2022.
- [65] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1065–1069, 2019.
- [66] P. Ramezanpour, M. J. Rezaei, and M. R. Mosavi, "Deep learning-based beamforming for rejecting interferences," *IET Signal Processing*, vol. 14, no. 7, pp. 467–473, 2020.
- [67] K. Chen, C. Qi, C.-X. Wang, and G. Y. Li, "Beam training and tracking for extremely large-scale MIMO communications," *IEEE Transactions on Wireless Communications*, 2023.
- [68] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8017–8045, 2022.
- [69] J. Guo, T. Chen, S. Jin, G. Y. Li, X. Wang, and X. Hou, "Deep learning for joint channel estimation and feedback in massive MIMO systems," *Digital Communications and Networks*, vol. 10, no. 1, pp. 83–93, 2024.
- [70] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Dual CNN-based channel estimation for MIMO-OFDM systems," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5859–5872, 2021.
- [71] R. Shankar, "Bi-directional LSTM based channel estimation in 5G massive MIMO OFDM systems over TDL-C model with rayleigh fading distribution," *International Journal of Communication Systems*, vol. 36, no. 16, p. e5585, 2023.
- [72] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.
- [73] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.
- [74] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, Z. Zhao, and H. Zhang, "Semantics-empowered communications: A tutorialcum- survey," *IEEE Communications Surveys & Tutorials*, 2023.
- [75] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," *IEEE Communications Surveys & Tutorials*, 2024.

- [76] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.
- [77] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of Artificial Intelligence*, pp. 603–649, 2020.
- [78] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2023.
- [79] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Nature, 2022.
- [80] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2022.
- [81] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1394–1398, 2022.
- [82] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.
- [83] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, and K. B. Letaief, "Semantic communication meets edge intelligence," *IEEE Wireless Communications*, vol. 29, no. 5, pp. 28–35, 2022.
- [84] H. Tong, Z. Yang, S. Wang, Y. Hu, O. Semiari, W. Saad, and C. Yin, "Federated learning for audio semantic communication," *Frontiers in Communications and Networks*, vol. 2, p. 734402, 2021.
- [85] Z. Han and K. R. Liu, *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*. Cambridge University Press, 2008.
- [86] Y. Teng, M. Liu, F. R. Yu, V. C. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2134–2168, 2018.
- [87] R. Zheng and C. Hua, *Sequential Learning and Decision-Making in Wireless Resource Management*. Springer, 2016.
- [88] E. Hossain, M. Rasti, and L. B. Le, *Radio Resource Management in Wireless Networks: An Engineering Approach*. Cambridge University Press, 2017.
- [89] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.
- [90] F. Dong, F. Liu, Y. Cui, W. Wang, K. Han, and Z. Wang, "Sensing as a service in 6G perceptive networks: A unified framework for ISAC resource allocation," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3522–3536, 2022.
- [91] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu et al., "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.
- [92] S. Wang, W. Dai, H. Wang, and G. Y. Li, "Robust waveform design for integrated sensing and communication," *IEEE Transactions on Signal Processing*, 2024.
- [93] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [94] X. Ge, F. Yang, and Q.-L. Han, "Distributed networked control systems: A brief overview," *Information Sciences*, vol. 380, pp. 117–131, 2017.

- [95] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2131–2165, 2021.
- [96] H. Djigal, J. Xu, L. Liu, and Y. Zhang, "Machine and deep learning for resource allocation in multi-access edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2449–2494, 2022.
- [97] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.
- [98] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.
- [99] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3163–3173, 2019.
- [100] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.
- [101] X. Xiong, K. Zheng, L. Lei, and L. Hou, "Resource allocation based on deep reinforcement learning in IoT edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 6, pp. 1133–1146, 2020.
- [102] K. Xu, N. Van Huynh, and G. Y. Li, "Distributed-training-and-execution multi-agent reinforcement learning for power control in HetNet," *IEEE Transactions on Communications*, 2023.
- [103] M. Lee, G. Yu, and G. Y. Li, "Learning to branch: Accelerating resource allocation in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 958–970, 2019.
- [104] —, "Graph embedding-based wireless link scheduling with few training samples," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2282–2294, 2020.
- [105] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3834–3862, 2020.
- [106] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for 5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 9–39, 2023.
- [107] D. Wen, X. Li, Y. Zhou, Y. Shi, S. Wu, and C. Jiang, "Integrated sensing-communication-computation for edge artificial intelligence," *IEEE Internet of Things Magazine*, vol. 7, no. 4, pp. 14–20, 2024.
- [108] B. Zhang and G. Y. Li, "White-box 3D-OMP-transformer for ISAC," *arXiv preprint arXiv:2407.02251*, 2024.
- [109] B. Zhang, Z. Qin, and G. Y. Li, "Compression ratio learning and semantic communications for video imaging," *IEEE Journal of Selected Topics in Signal Processing*, 2024.
- [110] S. Zhou and G. Y. Li, "FedGiA: An efficient hybrid algorithm for federated learning," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1493–1508, 2023.
- [111] —, "Federated learning via inexact ADMM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9699–9708, 2023.
- [112] U. Demirhan and A. Alkhateeb, "Integrated sensing and communication for 6G: Ten key machine learning roles," *IEEE Communications Magazine*, vol. 61, no. 5, pp. 113–119, 2023.
- [113] S. H. Alsamhi, A. V. Shvetsov, S. Kumar, J. Hassan, M. A. Alhartomi, S. V. Shvetsova, R. Sahal, and A. Hawbani, "Computing in the sky: A survey on intelligent ubiquitous computing for UAV-assisted 6G networks and industry 4.0/5.0," *Drones*, vol. 6, no. 7, p. 177, 2022.

- [114] V. A. Nugroho and B. M. Lee, "A survey of federated learning for mmWave massive MIMO," *IEEE Internet of Things Journal*, 2024.
- [115] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 487–500, 2022.
- [116] O. Wang, S. Zhou, and G. Y. Li, "Few-shot learning for new environment adaptation," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 351– 356.
- [117] O. Wang, J. Gao, and G. Y. Li, "Learn to adapt to new environments from past experience and few pilot blocks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 2, pp. 373–385, 2022.
- [118] Y. Liu, Z. Qin, and G. Y. Li, "Energy-efficient distributed spiking neural network for wireless edge intelligence," *IEEE Transactions on Wireless Communications*, 2024.
- [119] O. Wang, S. Zhou, and G. Y. Li, "BADMM: Batch ADMM for deep learning," *arXiv preprint arXiv:2407.01640*, 2024.
- [120] H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce memory, not parameters for efficient on-device learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 285– 11 297, 2020.



AI in the 5G-A Era: Scenarios, Key Technologies, and Evolution Trends

Yingpei Lin, Yan Chen, Yi Qin, Yan Sun, Rui Xu, Yuwen Yang, Zhengming Zhang, Jiaxuan Chen, Yang Tian, Youlong Cao, Xiaomeng Chai, Hongzhi Chen, Hong Qi, Xu Pang

Research Dept, WN

Abstract

As the core driving force behind today's technological revolution, artificial intelligence (AI) is revolutionizing the communications field when combined with 5G-A networks. In this paper, we explore the evolution trend of AI, analyze its key values in 5G-A networks, and discuss emerging application scenarios like artificial intelligence generated content (AIGC) and embodied intelligence. We also identify the key requirements and challenges of these application scenarios for 5G-A networks and describe how to address these challenges by leveraging AI to improve 5G-A network performance and enabling 5G-A networks to provide high-quality AI services. Finally, we conclude this paper by considering the deep integration of AI and 5G-A networks in the future and envisioning a new era of intelligent and personalized communications.

Keywords

AI, 5G-A, network performance, AI service, AIGC, embodied intelligence

1 Introduction

The development of communications technologies has always played an important role in driving social transformation. 5G-A networks — the latest generation of communications network — provide a broad platform for various emerging technologies thanks to their high throughput, low latency, and high reliability. With the integration of artificial intelligence (AI) and 5G-A networks, a revolutionary leap in the communications field is now possible.

Featuring powerful data processing capabilities and intelligent decision-making, AI has become the core driving force behind today's technological revolution. Its influence is continuously expanding from basic theoretical research to extensive industry applications. In the mobile communications field, AI applications are already reshaping the network architecture and service mode. This paper delves into the key roles and development trends of AI technologies in the 5G-A era, analyzes how AI helps improve 5G-A network performance, and explores the potential application of AI technologies in the future communications field.

In this paper, we review the development of AI models, computing power, data, and application paradigms, summarize the AI evolution trend (see Section 2), and analyze the key values of AI in 5G-A networks (see Section 3). We also discuss emerging application scenarios like artificial intelligence generated content (AIGC) and embodied intelligence in the AI era. Then, we identify the key requirements and challenges of these application scenarios for 5G-A networks (see Section 4), and describe how to address these challenges by leveraging AI to improve 5G-A network performance and enabling 5G-A networks to provide high-quality AI services (see Section 5). Finally, we conclude this paper by considering the deep integration of AI and 5G-A networks in the future and envisioning a new era of intelligent and personalized communications (see Section 6).

2 AI Evolution Trends

In addition to being the core driving force behind today's technological revolution, AI also has a significant and far-reaching impact on various industries by extracting high-value information from data. AI is widely used in various fields and, as a multi-disciplinary subject, it involves research of basic theories and cutting-edge technologies, creating

significant influence and business value. Breakthroughs in AI technologies, especially in industries such as the mobile Internet, have not only redefined what is possible, but also brought about profound changes in society and the economy. Incredible breakthroughs, skepticism, and constant evolution form the history of AI. As we prepare for the further development of AI, we are fully aware of its infinite potential. And with the advent of the 5G-A era, the explosive development of AI requires us to actively embrace this technique in order to achieve continuous breakthroughs and progress.

2.1 AI Models

AI models represented by neural network models were first proposed by Warren Sturgis McCulloch and Walter Pitts in 1943 [1]. Then in 1993, following years of evaluation and development, perceptron-based AI models [2] were proposed by Frank Rosenblatt, setting off a new wave of development. Since then, machine hardware has undergone rapid iterations and updates, while both computing capability and storage space have been significantly improved. At the same time, AI models have gradually been adopted in various fields and attracted attention from all sectors of enterprises. Multi-layer perceptrons also emerged, which can work on hardware entities and be used for image recognition, after researchers created them by adding more layers to perceptrons.

Researchers found that multi-layer perceptrons had strong potential in solving complex problems such as image recognition and began work on inventing new AI models to model data like text. In 1997, the long short-term memory (LSTM) [3] recurrent neural network was proposed and had a profound impact on subsequent AI research. Thanks to the iterative upgrade of computing and storage devices, the LSTM architecture involving memory units controlled by three gates (input gate, forget gate, and output gate) can determine the memory increase/decrease and output of AI models through logic gates. As a representative of recurrent neural networks, LSTM shows great ability in dealing with long sequence problems. It has become a classical neural network architecture for sequence tasks, such as text classification, sentiment analysis, speech recognition, image title generation, and machine translation. However, this type of AI model involves large-scale parameters and high computing costs in order to achieve an optimal effect, and it cannot meet requirements in the case of limited computing power.

In 2006, Geoffrey Hinton invented the restricted Boltzmann machine (RBM) model and deep belief network (DBN) to train multi-layer neural networks, and officially named multi-layer neural networks as deep learning (DL) [4]. It was at this moment that AI models entered the deep neural network (DNN) era. AI models with more layers and smarter structures based on DNNs emerged and achieved exciting results. For example, AlexNet [5], which consists of five convolutional layers, one max pooling layer, three fully connected layers, and one softmax layer, won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The industry quickly realized that deep convolutional neural networks (DCNNs) are adept at handling visual recognition tasks. Thus ensued an era of large-scale development and application of convolutional neural networks (CNNs). The most representative visual geometry group (VGG) [6] and ResNet [7] were proposed successively, and improved the performance of computer vision tasks to an unprecedented level. AI models had evolved from simple two-layer neural network models to VGG and ResNet models with more than 10 and 50 convolutional layers, respectively. Additionally, the ability of AI models had changed from recognizing simple images such as optical character recognition (OCR) to implementing advanced tasks like semantic segmentation and instance segmentation. However, researchers noticed that these CNNs still had certain limitations in natural language processing (NLP) tasks, and AI models underperformed in understanding natural languages.

With the emergence of Transformer [8], AI models have made significant breakthroughs in natural language understanding. Transformer is a neural network model based on the attention mechanism and does not use recurrent networks or convolution. This type of model consists of multi-head attention, residual connection, layer normalization, fully connected layer, and positional coding. It mines the correlation of sequence information while retaining the sequence order. Transformer has revolutionized NLP and quickly became a core AI model that dominates other fields (e.g., computer vision). In NLP, this type of AI model is often used for machine translation, text summarization, speech recognition, text completion, and document search. It was a precursor to large language models (LLMs) such as ChatGPT. Released by OpenAI in November 2022, ChatGPT has become a phenomenal application of generative AI. This AI model is based on the GPT-3.5 architecture and trained through reinforcement learning. It enables the transformation of AI models from image recognition to image understanding and then to generative AI [9]. The number of parameters in such AI models has also changed from thousands to millions or even billions.

2.2 AI Computing Power

Computing power is a key driving force for AI development and progress. Over the past decade, the amount of computing power used to train AI models has increased by 350 million times. Many of the advances in AI stem from the significant increase in the computing power used to train and run AI models. Developers have harnessed immense computing power to train LLMs [9], AlphaGo [10], protein folding models [11], and autonomous driving models [12] on huge datasets, adopting a different approach from the previous training and deployment of small-scale AI models. As a result, AI models are now capable of solving problems. In many AI fields, researchers have found a scaling rule, namely, the performance of a training objective (such as predicting the next word) increases with the amount of computing power used to train the model.

Thanks to hardware improvements, the computing power of end devices, base stations, and clouds has been improved to an unprecedented level. With the development of advanced computing units such as CPUs, GPUs, and TPUs, computing is not limited to traditional data centers. Instead, it is gradually moving to the edge, especially in full-stack scenarios. Thanks to the deployment of intelligent devices with computing capabilities, clouds, edges, and devices can jointly provide computing power for AI, effectively improving the computing efficiency and performance. Clouds, which store the collected data in data centers, perform computing and processing at central points, while base stations are edge computing devices that offer powerful data computing and processing capabilities. With end devices becoming more intelligent and AI models being iteratively upgraded, the performance of intelligent device processors has been significantly improved, and the computing power of end devices has been enhanced.

2.3 AI Data

Training AI models requires high-quality and large-scale data, which is one of the decisive factors for AI success. In the early stage of AI development, researchers manually built extremely limited datasets for AI model training and evaluation. However, as AI models grew in size and capabilities, researchers found that training data had become a bottleneck for achieving higher-performance AI models. The desire for high-quality training data has become a major challenge for AI development. For example, existing large-scale AI language models are built using text

obtained from the Internet. Such text includes scientific research, news reports, and Wikipedia entries, and is further decomposed into lexical elements. According to researchers, the training data used by the GPT-4 model contains up to 12 trillion lexical elements. If AI models continue to grow at the same pace, it is possible that training will require 60 to 100 trillion lexical elements. To address issues with data exhaustion, the AI field demands stronger data acquisition capabilities.

2.4 AI Application Paradigms

When AI first emerged, AI applications mainly focused on function fitting. That is, neural network models were used to simulate specific functions or tasks, such as classification, prediction, and optimization. These applications were usually based on rules or statistical models, with the goal of improving efficiency and accuracy. Nowadays though, with the development of big data technologies, AI applications are shifting to data-driven models, which rely on large amounts of data to train algorithms and implement more accurate function fitting.

AI foundation models have developed rapidly in recent years, enabling AI to process data more efficiently and create new data and content. For instance, generative AI [13] can creatively synthesize and augment data by digging into the underlying distribution of the input data, understanding the joint probability distribution of data and labels, and generating content similar to training data. AIGC expands the scope of AI applications and transforms content generation from a manual-based to an AI-based process, making AIGC a promising development direction in the future.

AIGC can generate text, voice, and video content. The great potential of AI in creative text generation is demonstrated by advances in natural language generation technologies, such as ChatGPT launched by OpenAI. In addition to generating answers based on the patterns and statistical rules learned in the pre-training phase, ChatGPT can also complete a variety of tasks such as writing papers, emails, and scripts, performing copywriting, and undertaking translation and encoding. In the image and video fields, DL technologies, such as generative adversarial networks (GANs) and extended models, are also used to generate realistic image and video content, including animations, simulated scenarios, and special effects. OpenAI's Sora demonstrates AI's ability to generate a complex 60-second video with multiple characters, specific types of motions,

precise themes, and background details based on text instructions [14].

In summary, the AI application paradigms have evolved from simple function simulations to complex intelligent systems capable of understanding, learning, and creating. As a new chapter in the development of AI, AIGC brings new possibilities to the AI field, and provides innovative tools and solutions for various industries, signifying the arrival of a new era full of imagination and creativity.

3 Key Values of AI in 5G-A Networks

3.1 Improving Network Performance

AI's technological breakthroughs in language, audio, and image processing demonstrate the performance advantages of data-driven methods in multimodal information feature extraction and problem solving. With the rapid development of AI technologies, the convergence of wireless networks and AI is deepening. Improving wireless network performance with AI technologies has become a popular research topic in the communications field. Research in both academic and industry circles shows that AI technologies can be used to enhance wireless networks from multiple dimensions, such as air interface and core network intelligence. In particular, the design of wireless intelligent air interface technology is a key direction in improving network performance and promoting standardization.

- **Network performance improvement:** Traditional communication modules are designed using the model-driven method. That is, the communications system is simplified and modeled based on factors such as Gaussian hypothesis and linear hypothesis, consisting of modules such as modulation and demodulation, encoding and decoding, and channel measurement. AI technologies, adopting a data-driven design method, learn input and output mapping relationships of different communication modules based on training data other than the non-real hypotheses in system modeling. This makes it possible to enhance the performance of different communication modules with reference to data-driven AI technologies, for example, improved channel measurement precision and enhanced signal demodulation performance. Ultimately, the overall communication network performance is enhanced, including improved user-perceived rate and communication coverage.

- **Deterministic service assurance:** The time-varying nature of a communications environment leads to unstable performance of communication links. Ensuring deterministic transmission of services in a changing communications environment has always been an important research direction of wireless networks. Data-driven AI technologies learn the change characteristics of the communications environment and adaptively adjust communication policies, making it become an important technical direction to achieve deterministic service assurance.

The evolution of AI-based intelligent air interface design has also promoted the discussion of radio access network (RAN) AI standards. Research into the design of AI-based air interfaces was initiated in 3GPP Release 18. This research project explored the impact of AI-based design on the overall wireless network framework, the performance of some typical use cases, and their impact on standardization [15]. The project defined basic AI concepts, simulation and verification methodologies, and base station and end device cooperation modes. It also studied each phase in the lifecycle management process, including model/function registration, data transmission, model transmission, model/function selection, and model/function activation and deactivation.

3.2 Providing High-Quality AI Services

The development of AI-based wireless networks can significantly improve network performance. And as a system with powerful communication connections, distributed computing power deployment, distributed data processing, and AI algorithms, wireless networks can provide broad possibilities for the extensive construction of high-quality AI services such as AI-based image enhancement, gaming, extended reality (XR), immersive communication, and other experience-oriented services. However, such services pose extremely high requirements on network performance.

With the design and application of foundation models in the AI field becoming a development trend, the exponential growth of the foundation model scale also brings important challenges to the implementation of AI services. The technical architecture of distributed training and inference has emerged to address such challenges and is considered as a fundamental feature of the next-generation AI architecture. The edge AI computing power deployed on base stations, combined with device-

network convergence, makes it possible to build distributed training and inference capabilities in wireless networks. An advantage of this architecture is that, by offloading AI computing requirements from the central cloud server to network devices closer to users, it can effectively optimize the latency and energy consumption of AI services, thereby improving user experience.

As services expand, the data processing capabilities of wireless networks become richer, providing efficient support for end-to-end data collection, transmission, storage, and sharing. These capabilities are deeply integrated with distributed training and inference capabilities in wireless networks, enabling larger-scale, more intelligent model optimization, training, and inference. However, providing data to internal or external network functions in a secure and efficient manner is still a subject that needs to be further studied.

Looking ahead, the ubiquitous edge computing power of networks will provide powerful support for AI services. With the platform advantages of network-integrated communication, sensing, and computing, we can open up a new market space for network participation in AI services, providing a strong driving force achieving prosperity in the AI era. Specifically, a RAN has natural advantages in transmission, collaboration, and sensing, and can meet requirements of AI services for low latency, high intelligence, wide coverage, and low power consumption based on the technology roadmap of "communication-computing convergence and sensing-computing convergence", accommodating explosive growth of AI services in the future.

4 5G-A AI Use Cases

4.1 Use Case Analysis

4.1.1 AIGC Applications Based on Artificial General Intelligence Devices

AIGC's multimodal processing capability can significantly improve production efficiency and reduce labor costs [16]. As one of the most popular artificial general intelligence (AGI) applications, AIGC is gradually changing the way we create and consume content. In this section, we describe the capabilities and functions of AIGC in three scenarios — e-commerce livestreaming, cloud gaming, and video calling — from the perspective of 5G-A networks.

- **E-commerce livestreaming:** AIGC can generate attractive livestreaming content, including automatically generating product descriptions, answering frequently asked questions (FAQs), and even creating virtual streamers for 24/7 livestreaming. By analyzing audience interaction and feedback, AIGC can also adjust livestreaming content in real time, improving user engagement and purchase conversion rate. Automated content generation saves human resources, and provides a more personalized and diversified shopping experience [17].
- **Cloud gaming:** AIGC provides personalized game recommendations and dynamically generated game content in cloud gaming services. It can generate customized game levels or tasks according to players' gaming history and preferences (e.g., it can customize diverse NPC designs) to bring a unique gaming experience to players and increase game diversity and interest. AIGC further enables AI game battles and provides enjoyable gaming for players who pursue fiercer competition [18].
- **Video calling:** AIGC can improve call quality and provide functions such as real-time background replacement, voice quality enhancement, and sentiment analysis. By analyzing call content, AIGC automatically generates minutes of meeting (MOM), keyword tags, or emotional feedback, helping users better understand and review call content. This makes video calling more intelligent and efficient, especially in remote work and online collaboration scenarios.

The core advantage of AIGC is its high degree of automation and intelligence. With DL models, AIGC can analyze large amounts of data, learn human creative habits and styles, and independently generate high-quality content. From writing news reports and stories, to designing visual arts, or even producing music and videos, AIGC can deliver a unique perspective and creativity.

Furthermore, AIGC's customizability provides users with great flexibility to set different parameters and conditions according to their requirements. This allows AIGC to generate content with a specific theme, style, or emotion. Such customized services are especially popular in the advertising, marketing, and entertainment industries. Innovation is another highlight of AIGC. It is not restricted by traditional thinking modes or creative boundaries. Instead, it can explore unknown fields, create new forms of content, bring new horizons for artistic creation, scientific research, and education, and stimulate infinite imagination and creativity.

As technologies develop, we can expect AIGC to deliver more exciting applications and achievements in the future. For example, Apple Intelligence [19] demonstrated at the WWDC 2024 integrates GPT-4o to enable iOS, completely transforming Siri into the "ultimate virtual assistant." Additionally, Apple is preparing to develop it as "the most powerful killer AI application." Personalized interaction and intelligence implemented through collaborative on-device processing and private cloud computing will become the standard configuration of the iPhone 16 and subsequent models. The on-device inference and edge training of AGI devices also extend the subject of traffic consumption from humans to machines, posing urgent requirements for faster and more reliable communication pipes. AIGC may be the next super application of operators' ToC services.

4.1.2 Embodied Intelligence Applications Based on Robots

With the rapid development of AI technologies, robots have moved from science fictions to becoming a reality. They have shown great potential and value in multiple fields including industry, healthcare, service, and home business [20]. Embodied intelligence, as a new trend in the AI field, embeds intelligent systems into physical machines, so that they can directly interact with the environment. Achieving such intelligence relies on the integration of multiple disciplines, including mechanical engineering, electronics, computer science, and cognitive science. Based on this integration, intelligent systems can learn how to adapt to the environment, optimize their own behavior, and make decisions in complex scenarios.

Advanced AI models like GPT-4o have achieved multimodal interaction of text, audio, and images, and possess certain sentiment analysis capabilities. However, the key to embodied intelligence lies in the ability of intelligent systems to interact with the physical world. This represents the transition of AI from relying on manual prompts to running in a more autonomous and intelligent form. The intelligent agents can be robots, drones, unmanned vehicles, or other forms of automation equipment. They sense the outside world through an integrated sensor network and translate the sensing data into an understanding and response to the environment.

In practical applications, embodied intelligence has shown a wide range of potential. It plays an increasingly important role in improving production efficiency in industrial

automation, providing more personalized customer experience in the service industry, or performing high-risk tasks in exploring unknown fields. As the AGI industry and foundation models develop, and as technologies such as computer vision, computer graphics, NLP, and cognitive science become mature, embodied intelligence is rapidly evolving from theory to practice and from labs to daily life.

Oriented to the 5G-A+ era, the ToC embodied intelligence robots serving as personal/home assistants will be used in a series of application scenarios, such as parcel pickup, household shopping, and care assistance.

- **Companion robots** are a typical application of embodied intelligence in households. These robots can talk with humans, recognize their emotional state through facial expressions, voice intonation, body language, and the like, and understand and respond to human emotions and needs. For the elderly, companion robots monitor their health status, remind them to take medicine, or offer help in case of an emergency. And for children, the robots provide educational content (to enable online learning) and entertainment content, such as games, music, and storytelling, to serve as a friend and teacher.
- **Transportation robots** are playing an increasingly important role in the logistics and express delivery industries. These robots use sensor information and map data for navigation, obstacle avoidance, and optimal path planning. This improves delivery efficiency, reduces logistics costs, and provides convenient and fast service experience for users.

The high speed, low latency, and high reliability features of 5G-A networks will enable more accurate environmental perception for embodied intelligence devices and provide highly reliable end-to-end deterministic latency services for indoor and outdoor robot applications. This will have a far-reaching impact on society and usher in a more intelligent and personalized era.

4.2 Key Requirements and Challenges

With the continuous progress of AI technologies, the application fields of AIGC and intelligent robots are expanding. This poses new requirements and challenges to wireless networks in order to support richer and more complex content generation and interaction experience.

4.2.1 Requirements and Challenges of AIGC Applications

AIGC technology has strict requirements on low latency for 5G-A network transmission, especially in application scenarios that require real-time interaction, such as online gaming, virtual reality (VR), and remote control. Low latency ensures real-time content generation and interaction, significantly improving user experience. For example, in virtual livestream shopping, the total latency from when a user sends a request to when the content is displayed must be within 70 ms to 100 ms. Additionally, the one-way latency of the air interface must be within 5 ms to 10 ms.

Because AIGC applications may involve the transmission of large amounts of data, including HD images, video streams, and complex model parameters, a high-bandwidth network is critical to support fast transmission of the data. This is necessary to meet AIGC's data processing requirements. For example, the typical upload bit rate of 1080p videos ranges from 5 Mbit/s to 8 Mbit/s, while the downlink rate of AIGC-generated content may reach 100 Mbit/s [20, 21]. Additionally, wireless networks need to support efficient transmission of multimodal data, including text, images, audio, and videos, and provide differentiated QoS assurance for AIGC applications, ensuring that necessary network resources are allocated to mission-critical applications.

4.2.2 Requirements and Challenges of Embodied Intelligence

Thanks to the rapid iteration of AI technologies, especially LLMs, intelligent robots are becoming increasingly smart and able to quickly and reliably execute tasks in unstructured environments. However, the uncertainty of such environments and the diversity of tasks pose a series of challenges.

First, robots need relatively high computing power and will consume a lot of power if they perform all calculations. However, lightweight design is an important requirement for intelligent robots to work in real-world environments, making it impossible to equip such robots with numerous CPUs/GPUs or large-capacity batteries. Second, robots have relatively limited sensing data. For targets outside the field of view, relying only on this data may result in a low task success rate or a long completion time.

A key solution to addressing these challenges is to offload computing to networks. The multimodal data collected by robot sensors is transmitted to the network together with task instructions. The network performs inference to generate the final output, such as target detection and path planning results, and then sends the output back to the robot for execution. And thanks to its superior sensing capability, the network can provide comprehensive environmental information to assist the robot in task execution, such as path planning.

In these scenarios, the end-to-end inference latency must be within 200 ms, and the inference accuracy must be at least 90% [22]. Furthermore, the network must be capable of accommodating sufficient intelligent robots that meet the latency and accuracy requirements of the AI services. Specifically, each cell must support the stable running of at least 30 intelligent robots.

5 Key 5G-AI Technologies

5.1 Key Technologies for Improving Network Performance

Based on Shannon information theory, wireless air interface transmission technology has undergone long-term development since the 1950s and 1960s. It has been split into various sub-fields, such as modulation and demodulation, pilot and channel estimation, channel measurement, and waveform, which have been widely used in commercial communications systems of 2G to 5G cellular networks.

AI/machine learning (ML) has also undergone long-term development and accumulation. For example, Turing proposed the famous Turing test in 1954, and the concept of "artificial intelligence (AI)" was first proposed at the Dartmouth Conference in 1956. Then, AI underwent two rounds of technical development. After 2006, with DL algorithms and large datasets emerging as a new breakthrough point, the third wave of AI development quickly swept through various fields.

Combining AI with 5G-AI networks to improve network performance is a cross-field technology that connects communication theories and AI methods. By effectively combining and extending the mathematical model, system architecture, and algorithm design in the two fields, we can build an intelligent kernel for wireless communications

networks, providing better transmission performance, higher O&M efficiency, and tailored user experience. In this section, we describe the key technologies used by AI to improve network performance in terms of constellation design, flexible pilots, TDD-oriented high-precision channel measurement, and coverage enhancement waveforms.

5.1.1 AI-based Constellation Design

Constellation modulation is a digital modulation technology that carries information bits on carrier signals. Modulated signals can be vividly represented on a 2D plane by using a constellation diagram. In existing wireless communications systems, quadrature amplitude modulation (QAM) is usually used, that is, amplitude modulation is performed on two quadrature carriers I and Q. QAM can be further classified into N -QAM based on the number of constellation points in the constellation diagram, where N is the QAM order. Each modulation symbol can carry $\log_2 N$ bits. Under normal circumstances, the candidate sets of amplitude modulation on quadrature carriers I and Q are the same. Therefore, N is usually an even power of 2, for example, 16QAM, 64QAM, 256QAM, or 1024QAM, as shown in Figure 1.

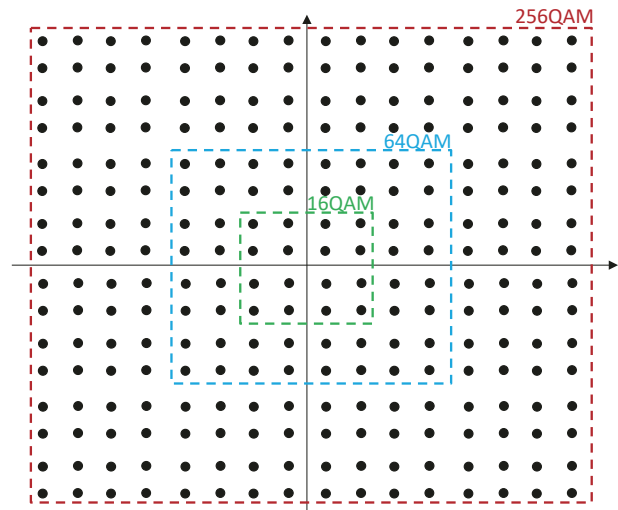


Figure 1 Constellations of 16QAM, 64QAM, and 256QAM

QAM is a regular modulation technology, with which the relationship between each constellation point and its corresponding information bit can be represented using the same formula. For example, the correspondence between the 16QAM constellation point and 4 bits can be represented as

$$s = \frac{1}{\sqrt{10}}((1 - 2b_1)(2 - (1 - 2b_3)) + 1j * (1 - 2b_2)(2 - (1 - 2b_4))),$$

which makes QAM modulation and demodulation easier to implement. However, this regularity also restricts the performance of QAM. In an Additive White Gaussian Noise (AWGN) channel, theoretically, the closer the constellation diagram is to Gaussian distribution, the closer the performance is to the Shannon channel capacity. The QAM constellation diagram is not represented as Gaussian distribution, accounting for a gap between the performance and Shannon channel capacity. The gap becomes larger as the number of orders increases, as shown in Figure 2.

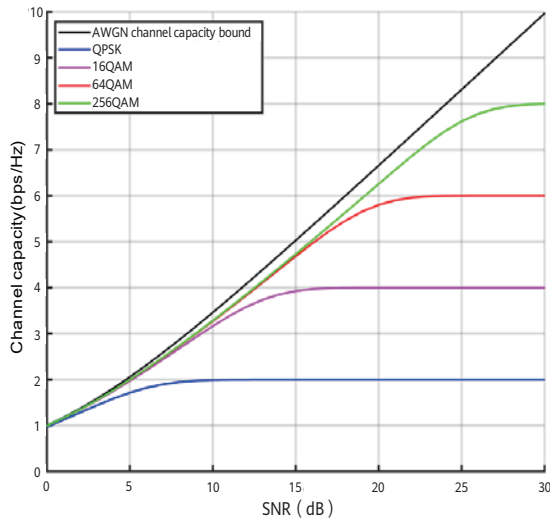


Figure 2 Channel capacity of QAM

Conventional irregular constellation modulation is classified into geometric shaping and probabilistic shaping. Geometric shaping changes the position coordinates of constellation points in the constellation diagram, so that these points tend to follow Gaussian distribution. And although probabilistic shaping does not change the geometric shape of the constellation diagram (e.g., the QAM constellation diagram), it does change the appearance probability of constellation points of transmitted signals, making them closer to Gaussian distribution.

Theoretically, both methods can make the distribution of the constellation diagram closer to Gaussian distribution. In reality though, given a non-ideal receiver, the optimal constellation distribution may not be Gaussian distribution, because it varies under different channel conditions (such as SNR). Consequently, it is theoretically difficult to provide an optimal constellation design.

An optimal constellation design that best adapts to the target channel conditions can be obtained by AI with end-to-end training. Compared with QAM, geometric shaping, and probabilistic shaping, AI-based constellation design is more flexible. As shown in Figure 3, in an AWGN channel, the geometric shaping, bit mapping, and corresponding demodulator of a constellation diagram can be jointly trained in an end-to-end manner, making the performance of the constellation diagram approximate the Shannon channel capacity. At the transmit end, the AI model outputs the constellation diagram, including positions (values of carriers I and Q) of all constellation points (modulation symbols), and bit mapping is represented by the sequence of constellation points. For example, for a 4-order constellation diagram, bit00 corresponds to the first constellation point, bit01 corresponds to the second constellation point, and so on. At the receive end, the AI model is fed with modulation symbols with noise, and outputs a log-likelihood ratio (LLR) for each bit. The loss function can be a binary cross entropy (BCE) function, so that the LLRs output by the AI demodulator approximates the sent information bits. After bits pass through the channel equalization module, most of the impacts from channel and multi-user interference have been eliminated. For the demodulator, it can be approximately considered that bits pass through the AWGN channel. Therefore, in fading channel and multi-user scenarios, constellation diagrams can be obtained through training in AWGN channels.

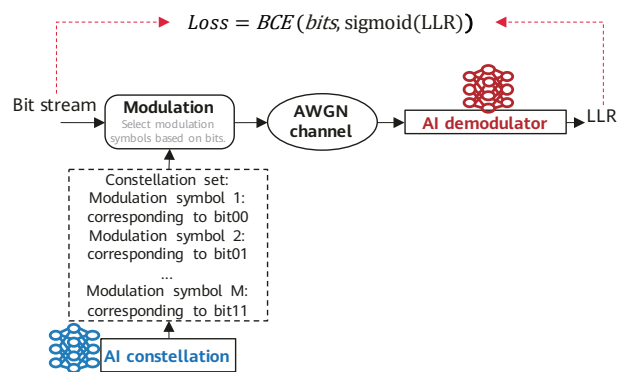


Figure 3 Training process of AI-based constellation design

Figure 4 shows two examples of irregular constellation diagrams designed by AI. Although AI-designed constellation diagrams approximate Gaussian distribution much more than QAM constellation diagrams do, they are still more flexible and can easily adapt to various channel conditions when compared with conventional geometric shaping.

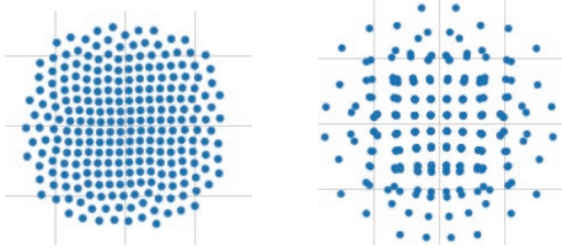


Figure 4 AI-based constellation design

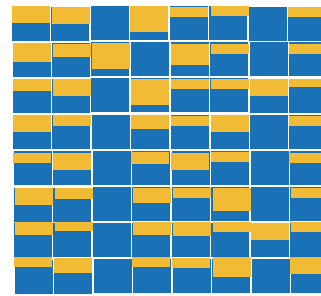


Figure 5 Transmission pattern of data and pilot symbols in RBs in superimposed pilot mode

5.1.2 AI-based Flexible Pilots

In 5G systems, there are various reference signals with different functions. Demodulation reference signals (DMRSs) are one such type of reference signal and are used to estimate the channel response of time-frequency resources occupied by data. However, there is a trade-off between channel estimation accuracy and DMRS density/overheads. If the channel has a relatively larger frequency selectivity (i.e., the channel changes significantly in the frequency domain), the DMRS density in the frequency domain should be increased. Similarly, if the channel changes significantly in the time domain, more resources need to be occupied in the time domain to deploy DMRSs. After determining the DMRS density in both the time and frequency domains, we need to further consider the positions of DMRSs in time-frequency resource blocks (RBs). For example, when the channel is stable, DMRSs can be evenly allocated in the frequency and time domains in order to reduce interpolation errors and implementation complexity. Because DMRSs do not transmit any data signals that are useful to users, DMRSs need to be allocated at an appropriate density to maximize throughput.

In existing protocols and solutions, DMRSs and data signals are still orthogonally allocated with time-frequency resources. Although this guarantees channel estimation performance, more resources need to be reserved for DMRSs (especially in a mobility scenario), leading to a contradiction between channel estimation performance and available data resources. A more flexible DMRS design can be generated using AI-based flexible pilot, which achieves the following functions by superimposing DMRSs and data signals for transmission:

- Providing a larger optimization space, releasing DMRS resources, and maximizing resource efficiency.
- Breaking the dependency on orthogonal DMRSs and the limitation on the number of DMRS ports.

- Resolving complexity issues by introducing AI to allocate power between data signals and DMRSs in superposition transmission mode and jointly design receivers (using AI to integrate multiple receive modules, including at least channel estimation and equalization modules).
- Implementing DMRS coverage on all available time-frequency resources, and forming an irregular DMRS pattern through power allocation to improve the system's adaptability to the communication environment.

Specifically, when each resource element (RE) carries both modulation symbols and reference signals, power normalization is still required. Figure 5 shows the pattern of pilot and data symbols on each RE of a single RB in watermark pilot transmission mode. In the figure, the blue lines indicate the power proportion of data symbols, and the yellow lines indicate the power proportion of pilot symbols. The receive end can perform channel estimation and data demodulation on superimposed pilots using a corresponding AI receiver.

AI is deployed on base stations in uplink scenarios, and deployed on UEs in downlink scenarios. The power allocation factor is obtained through AI training, and base stations allocate power based on the number of paired layers. The carried pilot sequence can reuse the sequence generation manner in existing standards.

5.1.3 TDD-oriented High-Precision Measurement

To improve coverage and performance of wireless networks, it is necessary to expand the antenna scale. However, this will lead to a significant increase in the amount of occupied air interfaces needed for high-precision channel measurement. Implementing high-precision channel measurement in a massive MIMO system with limited air interface occupation becomes a key challenge to the 5G-A time division duplex (TDD) system.

To address this challenge, it is critical to effectively use limited channel information obtained based on different reference signals to restore full-dimensional channel information. For example, UEs obtain downlink channel information from downlink reference signals, and base stations obtain uplink channel information from uplink reference signals. However, because the channel information from different reference signals originates from the same communication environment, the measurement can be considered as being performed for the same communication environment from different perspectives or in different modes.

By utilizing AI technologies that fuse multimodal information, we can effectively integrate channel information obtained from different reference signals, and implement high-precision restoration for the communication environment and channels.

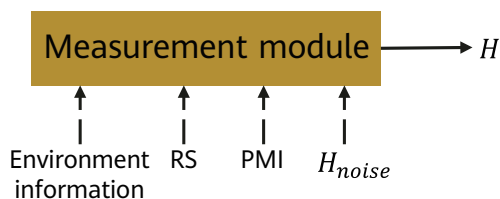


Figure 6 High-precision restoration for the communication environment and channels using AI technologies that fuse multimodal information

5.1.4 AI-based Coverage Enhancement Waveforms

A New Radio (NR) system currently has two types of waveforms in the uplink: multi-carrier orthogonal frequency division multiplexing (OFDM) waveform and single-carrier OFDM waveform. The latter is formed by adding an additional discrete Fourier transform (DFT) before a conventional OFDM processing process is executed and is therefore also called DFT-spread OFDM (DFT-s-OFDM) waveform. In contrast, the multi-carrier OFDM waveform involves superposing signals of multiple subcarriers. If subcarrier signals are superposed in the same direction, very high peak values are generated at times. Consequently, the single-carrier OFDM waveform has a low peak to average power ratio (PAPR), meaning a better coverage performance. However, for UEs requiring ultimate deep coverage (e.g., cell edge UEs or UEs with multi-time building penetration loss), the PAPR of the DFT-s-OFDM waveform is still high. It is necessary to design a new waveform with better coverage performance than the DFT-s-OFDM waveform for these UEs.

Conventional technologies used to reduce the PAPR of OFDM waveforms mainly include filter and clipping. Filter technology reduces the PAPR by designing a proper frequency domain filter to change the time domain waveform, whereas clipping technology reduces the PAPR by scaling down the signals that exceed the threshold. Although both technologies can optimize the PAPR of OFDM waveforms to a certain extent, they do so at the cost of losing some data transmission throughputs.

AI-based coverage waveform design optimizes the waveform coverage performance through AI training. This approach, when compared with the conventional waveform PAPR reduction technology, achieves a trade-off between coverage and throughput through multi-objective joint optimization, improving coverage while reducing throughput loss. A prime example of this is tone reservation (TR) technology based on AI.

In TR technology, several subcarriers in addition to the transmission bandwidth are reserved. As shown in Figure 7, signals on the subcarriers are optimized through AI training to change the waveform shape and reduce the PAPR. The input of the AI model is the time domain signals of the original data symbols, and the output is the signals on the reserved subcarriers. To avoid wasted power on the reserved subcarriers, the subcarrier power can be constrained during training.

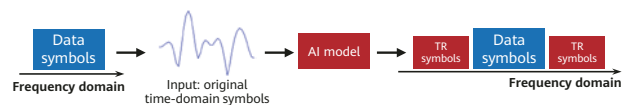


Figure 7 AI-based TR design

5.2 Key Technologies for Providing High-Quality AI Services

5.2.1 Distributed Inference

The increasing scale of parameters used in large models requires more powerful hardware for training and inference, pushing AI models to be deployed closer to the edge. Base stations, being closest to UEs, can detect channel changes in real time and fully utilize the deep coupling of communication and computing resources to optimize AI service inference performance.

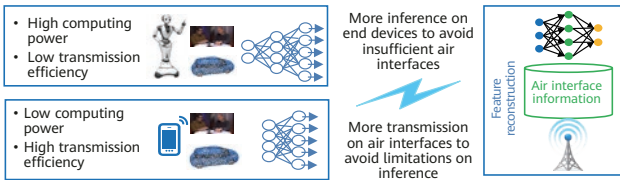


Figure 8 Dynamic scheduling of air interface resources based on joint communication and computing

Currently, there are four types of techniques for optimizing edge AI distributed inference:

- Sparse quantization and structure freezing:** It optimizes inference by reducing computing workload and memory usage. For example, quantization compresses model parameters from high-precision floating-point numbers (such as 32-bit) to low-precision integers (such as 8-bit) or fixed-point types, reducing the model size and computing workload. To achieve the same objectives, we can freeze parts of the model that do not need to be updated (such as pre-trained layers), preventing them from participating in training and inference. Sparsification is a process of removing unnecessary connections or parameters from a model to reduce the model complexity, thereby reducing the computing workload and storage space. Common sparsification methods include weight pruning and structured sparsification.
- Pipeline serialism:** Different layers of a model are allocated to different edge devices for serial inference. For example, the image preprocessing layer is allocated to a device with low power consumption, the convolutional layer is allocated to a device with higher performance, and finally the classification layer is allocated to a base station or cloud. This method fully utilizes the advantages of different devices to improve inference efficiency.
- Tensor parallelism:** Model computing tasks are allocated to multiple edge devices for parallel inference. For example, a large-scale matrix multiplication operation is decomposed into a plurality of sub-operations that are allocated to different devices for execution, and the results are finally integrated. This method fully uses the parallel computing capability of multi-core processors to accelerate model inference.
- Batch processing:** Multiple inference requests are combined for one-time processing to improve inference efficiency. For example, image recognition requests of multiple users are combined into a batch for inference, and the results are separately returned to the users. This method effectively reduces the inference delay and improves resource utilization.

Implementing dynamic sparsification, pipeline serialism, tensor parallelism, and batch processing based on dynamic changes of communications and computing resources is challenging from the perspective of air interfaces. By utilizing the unique advantages of communication-computing integration of base stations, we can design a dynamic device-network computing power collaboration solution, opening up a new space for future networks to participate in the AI computing field.

5.2.2 Retrieval Augmentation

In the world of rapidly developing wireless communications, the processing power, storage, and computing capabilities of end devices and access network devices often limit their performance. These limitations pose a challenge for large-scale model training. However, the emergence of distributed training effectively addresses this challenge. It enables computing tasks to be shared across multiple devices, thereby reducing the load of a single device while enabling each device to use its own data.

Distributed training for wireless communications systems involves several key technologies that constitute the technical pillar of this field:

- Model parallelism:** It is an effective method for enabling large-scale model training. When dealing with large numbers of model parameters that cannot be carried by a single device, model parallelism splits the model into multiple parts and distributes them to different computing devices. This reduces the memory requirements for each device and significantly improves computing efficiency. Model segmentation is a core step in this process and can be performed based on the model hierarchy or with reference to the functions and features of end devices, access network devices, and cloud servers.
- Distributed knowledge base:** Building a distributed knowledge base is crucial given the varying information that can be obtained by different end devices and access network devices. This process includes data collection, preprocessing, and knowledge representation design. For example, a base station with global sensing capabilities must collect data from various environments and preprocess it in a unified manner to ensure the consistency of its format and quality. We then need to determine how to represent this knowledge in the knowledge base, potentially involving the representation of features, model parameters, or intermediate results.

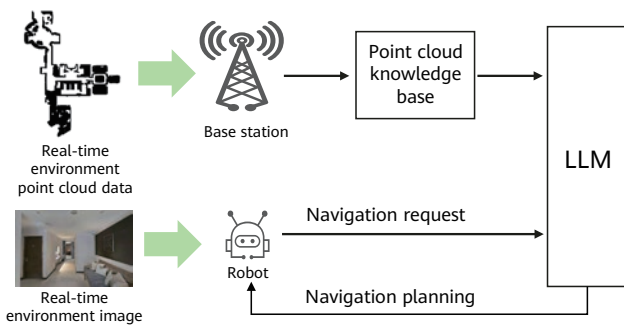


Figure 9 Building a distributed knowledge base

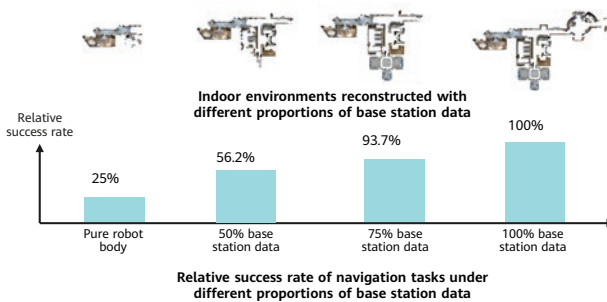


Figure 10 Success rate of robot navigation planning with the help of the base station's point cloud data

Figure 9 depicts how a distributed knowledge base is built. In the figure, the base station collects point cloud data of the environment to build a point cloud knowledge base, which is then shared to the LLM for further analysis and use. When executing a navigation task, the robot collects environmental information (video) in real time and sends the information to the LLM together with a navigation request. The LLM comprehensively analyzes the point cloud data and the video information to formulate an accurate navigation path. Finally, the LLM sends the path back to the robot, guiding the robot to complete the navigation task.

In indoor robot navigation planning tasks, the success rate is heavily dependent on the percentage of global environmental point cloud data provided by the base station. As shown in Figure 10, as the percentage of point cloud data increases, the robot's success rate in performing navigation planning tasks indoors also increases. This shows that detailed environmental data is critical to improving the accuracy and reliability of robot navigation. Specifically, the richness of point cloud data directly affects a large model's cognition of the environment, helps the large model optimize its path planning algorithm, and enables the robot to complete the navigation task more efficiently.

- High-precision small model design:** It is an important technique for improving the overall system performance. Multiple small models are designed based on the features of base stations and cloud servers, and are trained using high-quality labeled data. During training, advanced techniques such as transfer learning can be used to improve model learning efficiency. Additionally, small model prediction and large model verification techniques can also be utilized. The prediction result of a small model needs to be converted into a format suitable for the input of a large model, for example, a vector or encoding format. The large model then further verifies and adjusts the result of the small model to ensure the accuracy of the final output.

Through comprehensive application of these techniques, distributed training in wireless communications systems can overcome the performance limitations of devices while maximizing each device's advantages to implement more efficient and accurate model training.

5.2.3 Feature-based Communication

Feature-based communication is a new communication paradigm that focuses on the bit streams of data and further analyzes and understands the meaning that the data represents. It deeply understands the semantic content of the transmitted data. This helps the receive end identify the most valuable information for obtaining the intent of the transmit end, thereby improving communication efficiency and accuracy.

Semantic-aware air interface communication technology integrates semantic information understanding and processing into the communication process, and optimizes the transmission policy based on semantic features. This technology is reflected in two aspects: (1) differentiated transmission of feature flows over air interfaces, and (2) air interface communication with the fault tolerance of feature flows. The following describes these two aspects.

- Differentiated transmission of feature flows over air interfaces**
 In conventional communications systems, data packets are transmitted and received based on technical counters such as signal strength and error rate. However, feature-based communication can process data more intelligently by introducing semantic understanding. For example, when a data packet including important information is transmitted over the air interface, its

transmission priority can be increased, ensuring that this packet arrives at the destination quickly and reliably.

Additionally, feature-based communication can dynamically adjust the encoding and modulation policy based on the contribution of different feature flows to semantic recovery at the receive end.

Figure 11 shows an example of differentiated transmission of feature flows over air interfaces based on semantic guidance. Data on the signal source side is divided into multiple feature flows through semantic conversion. Before transmission, these feature flows are prioritized according to their semantic importance. Different channel encoding policies are then used based on these priorities. Take feature flows that can contribute more to semantic recovery at the receive end as an example. In such a case, more transmission resources are allocated, or a more reliable modulation and encoding policy is used. This ensures that the whole system for

differentiated transmission of feature flows over air interfaces based on semantic guidance can select the most appropriate transmission manner according to semantic importance. Furthermore, this optimizes the use of spectrum resources while also guaranteeing the communication quality.

- Air interface communication with the fault tolerance of feature flows

This technique can resist wireless channel instability and possible errors by recovering the semantic content of original information when some feature information is damaged or lost.

As shown in Figure 12 [23], conventional communication techniques focus on the correct transmission of bit streams. If part of the bit transmission fails, the receive end cannot recover the intent or information of the transmit end during image recovery. In this case, errored bits appear as mosaics, and artifacts appear on the entire image.

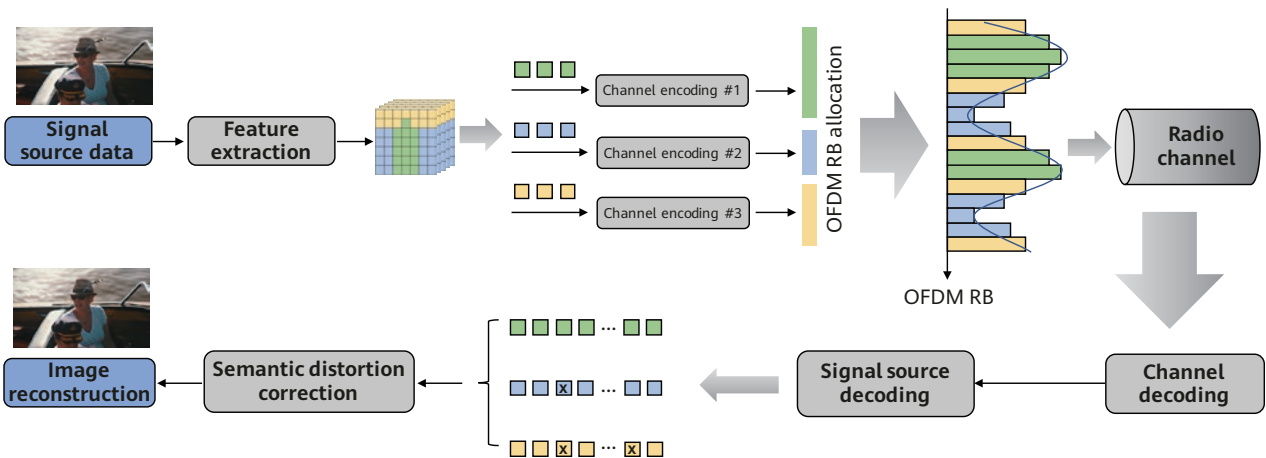


Figure 11 Differentiated transmission of feature flows over air interfaces based on semantic guidance

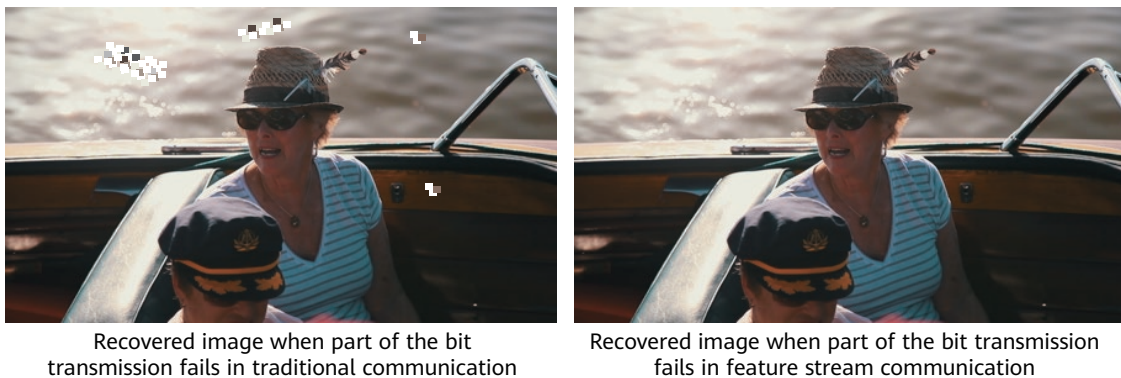


Figure 12 Air interface communication with the fault tolerance of feature flows

In contrast, air interface communication with the fault tolerance of feature flows ensures that the receive end can recover or infer the intent or information of the transmit end, even if part of the bit transmission fails, meaning that the image can be recovered [24]. To sum up, by leveraging fault tolerance of feature flows, we can significantly improve the tolerance and recovery capabilities of air interface communication.

6 Conclusion and Outlook

This paper discussed the evolution trend, key technologies, and application prospects of AI in 5G-A networks. It covers the evolution of AI models, the rapid progress of computing power and data, and the key values and use cases of AI in 5G-A networks, comprehensively demonstrating the opportunities and challenges faced by 5G-A AI technologies. In particular, AI's key role in improving network performance, providing high-quality AI services, and enabling embodied intelligence and AIGC applications underscores its core position in future communication and network development.

The in-depth integration of 5G-A and AI will herald the arrival of a new intelligent era. As technology advances, the potential of AI will become more apparent in more fields, especially in NLP, image and video generation, and multimodal interaction. This will enable more personalized and intelligent services to improve user experience and provide innovative solutions for various industries. Furthermore, AI's huge demand for computing power and data will continuously drive the upgrade of hardware and network infrastructures to empower more efficient distributed training and inference. Key technologies in improving wireless network performance, such as AI-based constellation design, flexible pilot, high-precision channel measurement, and coverage enhancement waveform, will further optimize the use of network resources and improve communication efficiency.

Although AI has a promising future, it also faces many challenges:

- **New breakthroughs in network capabilities:** The continuous evolution of 5G-A technologies will further improve the network speed, connection density, and latency.

- **New network O&M mode:** Intelligent and automated tools will change the network O&M mode and improve efficiency and accuracy.
- **Popularization of intelligent services:** AI technologies will be widely used in various industries to significantly improve production efficiency and user experience.
- **Technological innovation and integration:** New technologies such as edge computing and network slicing will be deeply integrated with AI to provide customized services for specific application scenarios.
- **Data security and privacy protection:** As the amount of data processed by AI increases, technologies and regulations that ensure data security and user privacy need to be strengthened.
- **AI explainability:** It is necessary to improve the transparency and explainability of AI decision-making to ensure the fairness and security of the system.

In summary, AI in the 5G-A era marks a new beginning, and its development will have a profound influence on the future social structure and human life.

References

- [1] McCulloch W S and Pitts W, "A logical calculus of the ideas immanent in nervous activity[J]," *The bulletin of mathematical biophysics*, 1943, 5: 115–133.
- [2] Rosenblatt F, "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms[R]," Cornell Aeronautical Lab Inc Buffalo NY, 1961.
- [3] Schmidhuber J and Hochreiter S, "Long short-term memory[J]," *Neural Comput*, 1997, 9(8): 1735–1780.
- [4] LeCun Y, Bengio Y, and Hinton G, "Deep learning[J]," *nature*, 2015, 521(7553): 436–444.
- [5] Krizhevsky A, Sutskever I, and Hinton G E, "ImageNet classification with deep convolutional neural networks[J]," *Advances in neural information processing systems*, 2012, 25.
- [6] Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition[J]," *arXiv preprint arXiv:1409.1556*, 2014.
- [7] He K, Zhang X, Ren S, *et al.*, "Deep residual learning for image recognition[C]," *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 770–778.
- [8] Vaswani A, Shazeer N, Parmar N, *et al.*, "Attention is all you need[J]," *Advances in neural information processing systems*, 2017, 30.
- [9] Wu T, He S, Liu J, *et al.*, "A brief overview of ChatGPT: The history, status quo and potential future development[J]," *IEEE/CAA Journal of Automatica Sinica*, 2023, 10(5): 1122–1136.
- [10] Silver D, Huang A, Maddison C J, *et al.*, "Mastering the game of Go with deep neural networks and tree search[J]," *nature*, 2016, 529(7587): 484–489.
- [11] Jumper J, Evans R, Pritzel A, *et al.*, "Highly accurate protein structure prediction with AlphaFold[J]," *nature*, 2021, 596(7873): 583–589.
- [12] Liu H X and Feng S, "Curse of rarity for autonomous vehicles[J]," *nature communications*, 2024, 15(1): 4808.
- [13] Ho J, Jain A and Abbeel P, "Denoising diffusion probabilistic models[J]," *Advances in neural information processing systems*, 2020, 33: 6840–6851.
- [14] <https://openai.com/index/sora/>
- [15] TR 38.843 v18.0.0, "Study on artificial intelligence (AI)/ machine learning (ML) for NR air interface," (Release 18), December 2023.
- [16] ALGC Panorama Report of Application in China, QbitAI Insights, March 2024.
- [17] Analysys, "Insights into China's live streaming e-commerce development in 2023 [R]," 2023. <https://www.analysys.cn/article/detail/20020949>.
- [18] China Academy of Information and Communications Technology, "Research report on in-depth observation and trend analysis of the global cloud gaming industry[R]," Beijing: China Academy of Information and Communications Technology, 2023.
- [19] WWDC 2024 — June 10 Apple, <https://www.youtube.com/watch?v=RXeOiIDNnek>.
- [20] CES 2023 Latest Robot [Technical Overview], "This year's Consumer Electronics Show has these differences," https://www.youtube.com/watch?v=_UflhU8pUU.
- [21] Sim, D. You, Y. Lee, *et al.*, "User-selectable stereoscopic video streaming using video enhancement information: System design and implementation."
- [22] Tanya Stivers, "Universals and cultural variation in turn-taking in conversation."
- [23] <http://toflow.csail.mit.edu/>
- [24] Dai J, Zhang P, Niu K, *et al.*, "Communication beyond transmitting bits: Semantics-guided source and channel coding[J]," *IEEE Wireless Communications*, 2022.



Anticipating 6G: Communication + AI

Xiongyan Tang, Youxiang Wang, Tengfei Sui
China Unicom Research Institute

Abstract

From the development history of mobile internet, we can see that the 3G/4G era mainly focused on connectivity. In March 2016, the emergence of AlphaGo marked the beginning of a new era where Artificial Intelligence (AI) enables networks, known as "AI for Net". The introduction of network data analytics functions (NWDAFs) in the 5G core network indicates the network is now moving toward automation and intelligence. At the same time, 5G has introduced technologies such as SDN/NFV and edge computing, which promotes the process of network-enabled AI, or "Net for AI", accelerating the development and practical application of the AI industry.

In the era of 6G, AI and networks will be deeply integrated. AI-enabled network ("AI for Net") will be transformed from add-on AI to native AI, and network-enabled AI ("Net for AI") will be efficiently supported through a unified endogenous AI architecture. 6G will go beyond mere connectivity and evolve toward a comprehensive digital information infrastructure that integrates connectivity, computing, and intelligence, achieving ubiquitous and inclusive intelligence.

Keywords

6G, native AI, foundation model

1 Multiplier Effects Brought by 5G-AI Integration

5G is characterized by high bandwidth, low latency, and massive connectivity, providing a solid foundation for the development of intelligent applications. This has sparked extensive research and exploration into the combination of 5G and artificial intelligence (AI), paving the way for rapid development of intelligent 5G networks.

3GPP introduced a new core network element (NE) called *network data analytics function (NWDAF)* in Release 15. This NE is primarily used to collect and analyze network data, and to provide information about network slice loads. With continued advancement of standardization, the functions of NWDAF continue to expand and improve, with the ultimate aim of achieving comprehensive support for centralized intelligent network architecture, service-oriented interfaces, and hierarchical and trustworthy intelligent network architecture. Release 18 covers two important aspects of AI applications in the network. First, federated learning is introduced, which allows NWDAFs to analyze and process data during the model training and inference stages. Second, supporting AI/machine learning (ML)-based services in 5G networks is introduced, including collaboration across radio access networks and core networks, model identification/management consistency, AI/ML application training, and related quality of service (QoS) enhancements. In Release 19, 5G intelligence-related standardization works will continue to advance, supporting continuous network optimization and further resource utilization, network efficiency, and customer experience improvements.

3GPP started research on wireless AI/ML in Release 17, mainly focused on functions of network energy saving, load balancing, mobility optimization, etc. Release 17 also defined a basic operational structure, including data collection, model training, model inference, and execution. Release 18 conducted research on enhanced channel state information (CSI) feedback, millimeter wave (mmWave) beam management, and positioning accuracy improvement, ushering in a new stage of system-terminal collaboration. Release 19 will study new AI/ML use cases, including support for distributed learning.

AI can enhance network intelligence and improve network performance and user experience, leading to a comprehensive increase in network production. AI has already been widely deployed in network planning,

construction, maintenance, optimization, and operation. Take network planning and construction as an example. AI can achieve network topology optimization, resource/capacity planning, intelligent acceptance, and new site selection to improve network service quality and user experience. As for network operation and maintenance (O&M), optimization, and energy saving, AI can achieve self-optimization of the network, including fault identification, prediction, intelligent diagnosis, and self-healing, thus improving network stability and reliability. In terms of network operation and customer service, AI can provide personalized service recommendations based on user behavior and preferences, thus improving customer satisfaction.

As an important digital infrastructure, 5G provides strong support and empowerment for AI applications by virtue of its characteristics such as high speed and low latency, as well as its powerful arithmetic and data aggregation capabilities, promoting the deep integration of AI technology with various industries, and facilitating the comprehensive digital and intelligent transformation of industries. 5G achieves cloud-edge-device computing power collaboration by distributed edge computing, and supports AI algorithms to run data more efficiently among the clouds, edges, and devices. In this way, AI inference tasks can be offloaded from the cloud to edges and devices. Innovative AI applications and services cover various fields, from smart homes to smart cities, and to smart healthcare and intelligent transportation systems, promoting the rapid derivation and development of the intelligent economy.

2 Symbiotic Harmony and Mutual Enablement Between 6G and AI

In June 2023, the International Telecommunication Union - Radiocommunication Sector (ITU-R) Working Party 5D (WP 5D) completed a Recommendation on the "Framework and overall objectives of the future development of IMT for 2030 and beyond" (hereinafter referred to as the "*Recommendation*"). The *Recommendation* proposed six typical usage scenarios for 6G, namely, "Immersive Communication", "Massive Communication", "Hyper Reliable and Low-Latency Communication", "AI and Communication", "Integrated Sensing and Communication", and "Ubiquitous Connectivity." Among them, the first three scenarios are the evolution and enhancement of the three major scenarios of 5G. "AI and Communication" is

a new scenario. The introduction of AI makes networks more intelligent and provides stronger support for models, algorithms, and data analysis.

The *Recommendation* also proposed the overarching aspects of 6G system design, including sustainability, security/resilience, connecting the unconnected, and ubiquitous intelligence that improves overall system performance. Built upon traditional connectivity capabilities, the 6G network is expected to integrate connectivity, computing, and intelligence, becoming the next-generation digital-intelligent service enabling platform.

2.1 AI-Native 6G Network Architecture

China Unicom has developed a hierarchical 6G network architecture based on the design philosophy of "beyond connectivity", "diversified users", and "platform as a service." This architecture inherits the connotative concepts of CUBENet3.0 and aims to become a next-generation digital-intelligent service enabling platform. In the architectural design, the concept of native AI is interwoven throughout all elements from top to bottom. The management and control layer is responsible for AI management and task orchestration. Control plane functions perform collaborative control, with user plane functions, data functions, and computing functions combined to jointly deploy, schedule, and execute AI tasks, forming an AI-native network architecture.

Figure 1 shows the design of China Unicom's AI-native 6G network architecture. The architecture is divided into four

layers (from bottom to top): resource layer, function layer, management and control layer, and service layer. Security management is implemented in a unified manner across all four layers [1].

- The resource layer comprises fundamental infrastructure such as computing, storage, networking, and spectrum. It supports the deployment, operation, and service of functions and serves as the foundation for the entire network operation.
- The function layer executes instructions from the management and control layer and leverages control plane functions to perform collaborative control. Specifically, user plane functions, data functions, and computing functions are combined to jointly deploy, schedule, and execute AI tasks. Among these functions, the user plane ones are responsible for access anchoring, service data forwarding, and network data forwarding. Data functions include data collection, preprocessing, storage, and sharing. Computing functions are divided into two modules: AI task execution control and AI task execution collaboration. These modules collaborate to schedule elements (e.g., compute nodes) for executing AI tasks, including service deployment and adjustment.
- The management and control layer is responsible for AI management and task orchestration, supporting the service layer in providing AI use cases and policies, QoS requirement analysis, algorithm/model management, AI task decomposition, network resource scheduling, etc.
- The service layer directly faces users and provides services such as connectivity, computing, data, and AI services.

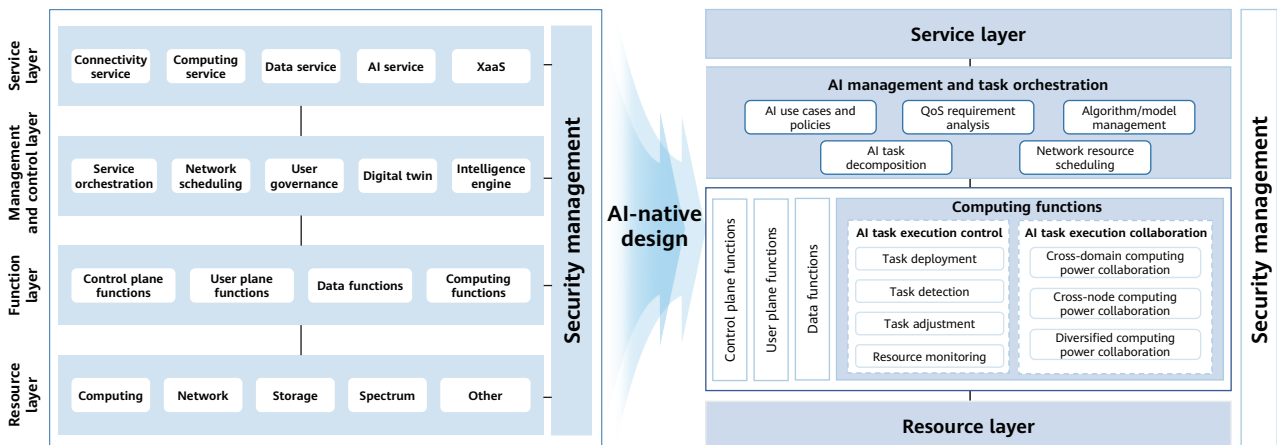


Figure 1 Design of China Unicom's AI-native 6G network architecture

AI-native designs must meet four fundamental criteria [2]: (1) adopting standard data formats to promote data interoperability and simplify AI model training and integration; (2) defining a quality of artificial intelligence service (QoAIS) system that extends beyond traditional network QoS to include new indicators such as AI service response time, reliability, and accuracy; (3) integrating the computing capabilities of AI with the transmission capabilities of communication networks to achieve optimal data processing; and (4) offering trustworthy AI systems that conduct transparent and fair decision-making, protect user privacy, and comply with ethical and regulatory requirements.

In the coming years, AI capabilities will be available in the form of AI as a service (AlaaS), facilitating internal network function calls and third-party capability openness. This will enable the development of distributed, efficient, energy-saving, and secure AI services and an open ecosystem by utilizing the resources and functions available in the 6G network, including connections, computing power, data, and models across the core network, radio access network, and UEs [3].

2.2 New Opportunities Presented by Foundation Models for Communication Networks

From Google's release of the Transformer model in 2017 to the unveiling of Sora (text-to-video model) in 2024, foundation models have gained significant momentum. They have evolved from merely perceiving and understanding the world to generating and creating it, and have transitioned from single-modal to multi-modal approaches. Instead of being purpose-built, foundation models are now more general, expanding their reach into diverse fields and application scenarios. As AI gradually enters the era of general intelligence, the emergence of foundation models presents new opportunities for the integration of 6G and AI.

The AI-native design of 6G enables each NE to natively integrate communication, computing, and storage capabilities, forming a multi-level computing power collaboration architecture that spans cloud, edges, and devices. Robust network connectivity enables control and dynamic, on-demand scheduling of a wider range of computing resources. Regarding the deployment and application of foundation models in the future, a cloud-edge-device collaboration approach is adopted to facilitate the entire process from model training to inference. Offering

ample storage and computing resources, the central cloud meets the demands of model training. Data and model transmissions are handled by the network. And the edges and devices utilize the trained model to complete the inference process, achieving faster response times, higher reliability, and better resource utilization.

The integration of AI foundation models and 6G will enable networks to reach unprecedented heights in data processing, analytics, optimization, and so on. This will result in higher network efficiency and achieve intelligent network operations. With the support of 6G, AI foundation models can be trained and inferred within the network, providing adaptable AI capabilities for various application scenarios and delivering high-performance AI services to users. Such models are more general than traditional small models, as they can generalize model capabilities and diversify tasks through content generation training. However, foundation models are weaker in specialized capabilities. This is where AI Agent comes in. It combines foundation models with small models, representing an important trend for future AI products. Figure 2 illustrates the "foundation model + small model" collaboration pattern in an example network optimization scenario. This pattern can better meet the differentiated needs of various application scenarios and improve the performance and efficiency of AI products.

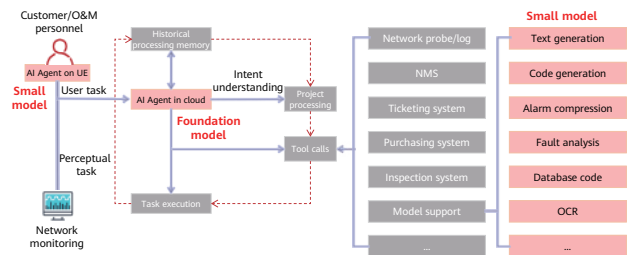


Figure 2 "Foundation model + small model" collaboration organized by AI Agent for network optimization

3 Joint Promotion of AI-Communication Integration

AI has injected new vitality into the evolution of 5G and the development of 6G, establishing itself as a key driving force for communication technology iteration and system/device innovation. As the launch of 6G standards approaches, China Unicom will continue to conduct in-depth research on the key technologies of 6G-AI integration. Through strengthened cooperation with the industry, China Unicom will strive to promote the intelligence of 6G networks and contribute to the development of the digital economy.

References

- [1] China Unicom, "China Unicom 6G network architecture white paper[R]," June 2023.
- [2] 6GANA, "6G network native AI technical requirements white paper[R]," Jan. 2021.
- [3] IMT-2030 (6G) Promotion Group, "6G network architecture vision and key technology outlook white paper[R]," Sept. 2021.



Computing as a Service: Unlocking the Boundless Potential of User Equipment

Yannan Yuan, Qi Wu, Yanchao Kang, Jiankang Liu, Xiaowen Sun, Dajie Jiang, Fei Qin
vivo Mobile Communication Co., Ltd.

Abstract

The usage scenario "AI and Communication" has introduced computing as a new service in 6G. Computing as a service in 6G will help drive the development of new types of user equipment (UE), enhance user experiences, and provide more possibilities for the evolution of UE. We recommend a method to establish performance indicators for computing services based on typical use cases (such as virtual humans), including computing power density, connection density for computing, peak computing power per UE, and computing latency per UE. To support the "AI and Communication" scenario, this paper suggests that potential technical directions for convergence of mobile networks and computing in 6G include the evolution-based path and revolution-based path. The evolution-based path includes the enhanced edge computing (EC) and enhanced IP multimedia subsystem (IMS). The computing control function, compute node, and computing data transmission channel (i.e., computing bearer/session) are introduced in the revolution-based path. This paper also describes a prototype for convergence of mobile networks and computing developed by vivo. The test results of the prototype show that the AI and communication services provided by the 6G network offer higher accuracy and lower latency compared to local computing on the UE, when computing power and transmission bandwidth are higher than the thresholds.

Keywords

6G, AI and communication, enhanced EC, enhanced IMS, computing control, prototype for convergence of mobile networks and computing

1 Introduction

The International Telecommunication Union - Radiocommunication Sector (ITU-R) proposed six usage scenarios (Immersive Communication, Massive Communication, Hyper Reliable and Low-Latency Communication, AI and Communication, Integrated Sensing and Communication, Ubiquitous Connectivity) in the "Framework and overall objectives of the future development of IMT for 2030 and beyond" [1]. In addition to the high user experienced data rate and low latency required in traditional communication scenarios, AI and communication applications require the convergence of AI and computing-related functions into the 6G system. This convergence will enable the 6G system to support model sharing and inference between different nodes, data collection from multiple data sources, and distributed AI model training. In this way, AI and communication applications are expanding the role of computing as a new service in 6G.

Currently, cloud computing is the main form of computing services including AI and rendering and the cloud computing market continues to expand [2]. Looking ahead, in which areas will 6G AI and communication give full play to their strengths and provide differentiated services? We think that in the 6G era, which aims for a freely connected physical and digital integrated world, user equipment (UE) will serve as a bridge between these two worlds. Therefore, the AI and communication usage scenario in 6G should focus on applications related to UE, expanding computing capabilities based on the existing mobile communication infrastructure. When AI and communication applications related to UE exhibit the following characteristics, 6G may offer differentiated services that surpass current cloud computing in terms of performance, efficiency, or security.

- AI and communication applications require low-latency computing that far exceeds the computing capabilities of most UE — for example, interactive high-fidelity 3D virtual humans.
- Partial data of AI and communication applications is provided by the 6G network — for example, AI model training, where the training dataset consists of network data.
- AI and communication applications require real-time scheduling of communication and computing resources considering channel status and communication performance fluctuation due to UE mobility. Based

on real-time information about channel status and computing resource status, the 6G system can better implement enhanced collaborative scheduling between computing and communication.

- Operators offer native AI and communication services based on an IP multimedia subsystem (IMS), like the voice service or short message service (SMS). UE can access AI and communication services without having applications installed.
- AI and communication applications require UE and network nodes collaborate to complete complex tasks — for example, distributed AI agents.

For the 6G AI and communication usage scenario, potential UE service flows include on-demand computing offloading and on-demand in-network computing. As shown in Figure 1, on-demand computing offloading, represented by a green dashed line, means that the UE determines whether to offload computing to the network based on the local computing power, computing task requirements, power consumption and so on. In this case, the same node transmits and receives computing data streams. If the UE needs to offload computing, it sends data to the 6G network and retrieves the computing result from the network. Regarding on-demand in-network computing indicated by blue dotted lines, computing services are provided on demand during network transmission. Applications can be hosted by the 6G IMS application server or over-the-top (OTT) application server to provide services. Based on the transmit and receive nodes of computing data streams, in-network computing can be classified as follows:

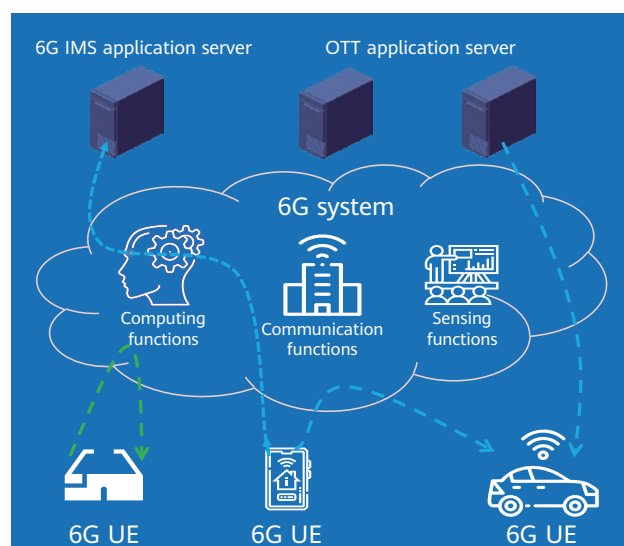


Figure 1 Computing service flows related to UE

- The UE transmits data. After transmission and computing through the 6G network, the application function receives the data.
- The application function transmits data. After transmission and computing through the 6G network, the UE receives the data.
- The UE transmits data. After transmission and computing through the 6G network, the other peer UE receives the data.

2 Computing Service Performance Indicators

AI and rendering are some of the computing services provided by a 6G system. Performance indicators of AI services include reachable performance (e.g., normalized mean square error and cosine similarity), AI model complexity, convergence speed (or training time), generalization capability, data dependency, inference time, computing overhead for training, and overheads for model transfer and storage [3]. The performance indicators of AI services will depend on the development level of computer technologies, such as AI algorithms and big data, in 2030 and beyond.

This paper focuses on potential performance indicators [4] determined by the combined computing and communication resources deployed in a 6G system and their performance. Use interactive 3D virtual humans as an example. Assume that virtual humans must respond to people's instructions and actions at a total latency of no more than 200 ms to ensure a real-time interaction experience. Considering other

overheads, such as transmission delay, it is estimated that the computing latency in a 6G system should range from 10 ms to 100 ms. Considering the precision and computing complexity of virtual humans, it is estimated that the computing power required for driving and rendering high-fidelity, intelligent interactive virtual humans with more than 500,000 faces (or geometries) should be not less than 10 Tera floating-point operations per second (FLOPS). Considering factors such as UE power consumption and capability limitations, UE can use on-demand computing offloading so that 6G can help support interactive 3D virtual humans. In addition, assume that the inter-site distance is 100 m in the three-sector coverage solution, the density of active users is one UE per 5 m², the average time for each UE to use virtual humans per day is 30 minutes, and the concentration rate (i.e., the ratio of the volume of computing in the busiest hour to the volume of computing throughout the day) is 10%. Table 1 lists the performance indicators [5] of using interactive 3D virtual humans in this typical instance.

3 Potential Technical Directions for Convergence of Mobile Networks and Computing

This section will delve into the potential technical directions of 6G in supporting AI and communication services based on existing standards. AI and communication services should be intrinsically built upon the convergence of mobile networks and computing [6], and the key to such a convergence is breaking the boundaries between layers.

Table 1 Comparison between FT- and CS-based imaging

Computing Service Performance Indicator		Definition	Requirement in the Typical Application of Virtual Humans
System-level performance indicators	Computing power density	Computing power provided by a mobile communication network per unit coverage area	~100,000 Tera FLOPS/km ²
	Connection density for computing	Number of computing connections provided by a mobile communication network per unit coverage area	~10,000/km ²
UE-level performance indicators	Peak computing power	Peak computing power per UE	~10 Tera FLOPS
	Computing latency	Total latency from the time when a UE initiates a computing service request to the time when the UE receives the computing response	10–100 ms

Computing resources at the bottom layer should be aligned with computing and communication requirements from upper-layer applications. This is also an important task of the computing control function in different technical directions for the convergence of mobile networks and computing. Typically, for AI services, computing control may further integrate functions such as data and AI model control. In this way, functions related to computing, data, and algorithms can be customized and optimized on demand for AI services to meet the requirements of AI model inference and training for application functions.

As shown in Figure 2, the potential technical directions of the convergence of mobile networks and computing in 6G are classified into two categories: evolution-based path and revolution-based path. The evolution-based path includes enhanced EC and enhanced IMS. In the revolution-based path, new computing-related network functions, such as computing control function, compute node, and computing data transmission channel (i.e., computing bearer/session), can be introduced to mobile networks. For EC enhancement and IMS enhancement, computing resources are located on the edge application server and IMS application server, respectively, instead of network nodes such as UPF or base station. In the evolution-based path, computing services and communications protocols are upper-layer applications and lower-layer transmission protocols, respectively. As for the revolution-based path, the collaboration between computing services and communication can be implemented in the control plane protocol of 3GPP standards. Therefore, the difference between the preceding two categories of technical directions lies in the relationship between computing services and communication protocols.

The technical direction based on EC enhancement is to extend the UE-side edge DNS client (EDC) [7], network-side edge application server discovery function (EASDF), and edge application server (EAS). The computing control node at the edge receives the computing requirements of applications from the UE through the user plane. The computing control node may also obtain real-time status information of the EAS compute node. Based on network and computing status information, the 6G network selects the EAS that best matches the computing requirements of the UE. The 6G network also supports application server redirection and dynamic route planning based on the computing status. In this way, UE requirements are dispatched in real time to the EAS with the optimal computing and shortest access latency to improve the user experience from the computing dimension.

The technical direction based on IMS enhancement is to extend the existing IMS or introduce IMS standards that support new services. In this technical direction, both computing control information and computing data between the UE and the network are transmitted through the user plane. Take an extension based on the existing IMS as an example. The existing IMS supports audio calls, video calls, and data channels [8]. For application scenarios such as AR or virtual humans, the data channel application server (DCAS) or media function (MF) provides image and video computing capabilities (for instance, image rendering). However, currently, media service processing supports only the preceding data channel service, and the IMS does not have the control capability of a compute node that the data channel service requires. Therefore, it would be a potential evolution direction for 6G to support on-demand

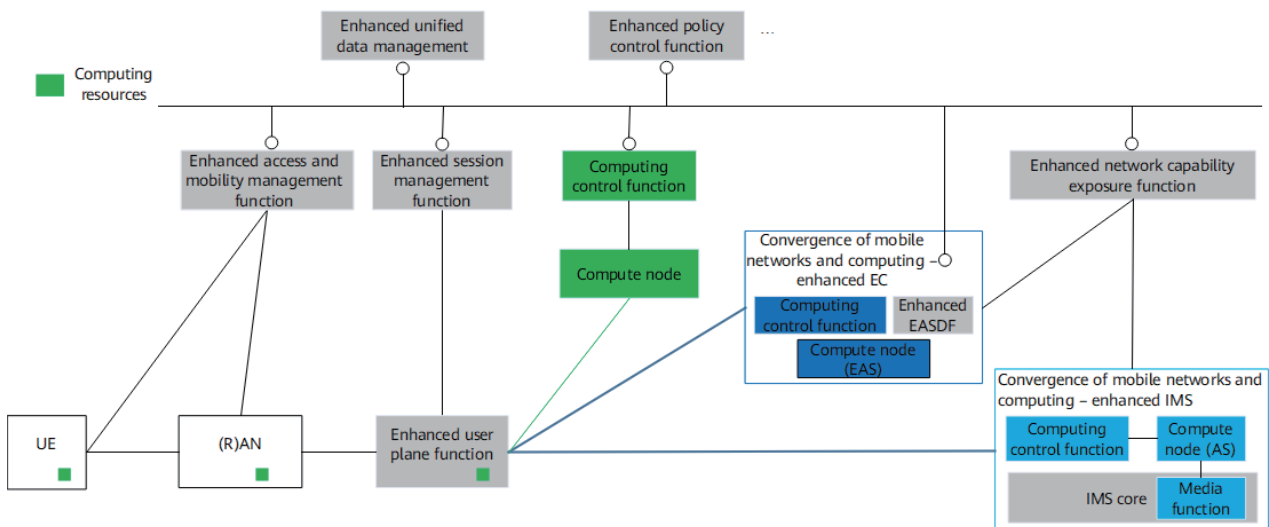


Figure 2 Potential directions for convergence of mobile networks and computing in 6G

computing based on data channels. This requires that media application management be enhanced in the process of establishing and using a data channel in order to manage the computing resource information accessible through the current data channel. The computing power and computing load status of application servers, and even the computing capabilities of UE are essential for using and establishing a data channel. Furthermore, a negotiation mechanism between the UE and network needs to be introduced so that different types of UE can maximally access data channel media services with high computing power requirements.

As for the revolution-based path, a computing control function needs to be introduced to the control plane of the core network. The computing control function may further obtain dynamic computing information of application servers. The function can effectively measure and efficiently manage heterogeneous computing resources in the network. A service-based interface is used for the computing control function, and computing control information between the UE and network is transmitted through the control plane. From the perspective of the communication protocol layer, the computing control function supports the end-to-end (E2E) computing control protocol layer equivalent to that of the UE. It focuses on computing control information exchange, including computing requirement exchange and computing resource allocation and update. The computing control function collaborates with the communication control function based on a policy, selects an appropriate compute node among the nodes with available computing resources, as shown in Figure 2, and manages the computing bearers or sessions required for computing data transmission. In this way, communication and computing resources can be dynamically scheduled in a coordinated manner to meet diverse service experience requirements.

4 Prototype Verification for Convergence of Mobile Networks and Computing

As for the revolution-based path, vivo has developed a prototype for the convergence of mobile networks and computing. Section 4.1 describes the system architecture and technical solution of the prototype. Section 4.2 describes the computing resource evaluation indicators and scheduling policies. Section 4.3 describes a typical demonstration case and compares the test results (i.e., latency) under different computing resources or communication resources.

4.1 Prototype System Overview

As shown in Figure 3, the prototype for the convergence of mobile networks and computing consists of four parts: a vivo UE verification platform, a base station, a core network that integrates communication and computing, and a management and orchestration system. The UE verification platform requests and uses network computing services. The base station provides wireless data transmission for the UE. Currently, the prototype does not involve the optimization or testing of base station. The core network responds to computing service requests from the UE and provides computing services for the UE. The management and orchestration system centrally manages the server cluster where the core network and computing service application functions are deployed through Kubernetes (K8s).

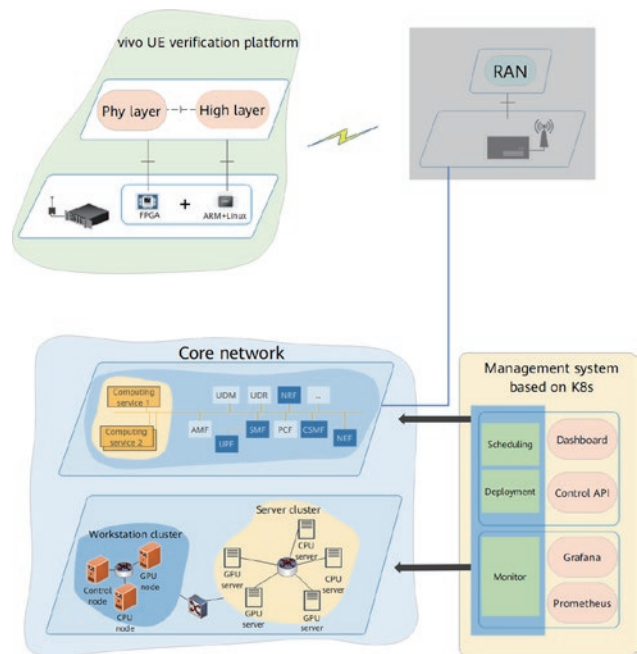


Figure 3 Architecture of vivo's prototype for convergence of mobile networks and computing

The UE verification platform uses the Arm+FPGA architecture. As shown in Figure 4, the hardware of the platform includes the FPGA accelerator card, universal subscriber identity module (USIM) card reader, software-defined radio (SDR), and host computer. The software includes the RF processing unit, physical-layer baseband processing unit, upper-layer protocol stack, and host computer program. The upper-layer protocol stack and the physical-layer baseband processing module run on the Arm and FPGA, respectively. Currently, the preceding test UE supports the main features of New Radio (NR) Rel-15 and certain features of NR Rel-17 and can be used to

verify the principles of new features such as 6G-oriented large bandwidth and low latency. To meet the verification requirements of the convergence of mobile networks and computing, the non-access stratum (NAS) module of the UE verification platform has been developed to process computing requirements from the application function and send computing requests to the network. Specifically, a computing request from the host computer (that is, the UE application processor) is obtained by running an ATtention (AT) command, and the computing request and the response message are separately carried in NAS messages. This achieves interactive computing and communication between the UE and the core network.

The 5G core network is enhanced with the computing service management function (CSMF) and computing server (CS). The CSMF schedules, manages, and maintains computing resources. The CS monitors the status of the current compute node and responds to the management and scheduling information of the CSMF. The network also enhances the session management function (SMF) and user plane function (UPF). The SMF collects statistics on the network status between each compute node and the UPF and responds to the CSMF's subscription to various network status information.

To implement load balancing and visualized monitoring between compute nodes, the management and orchestration system leverages the synergy between the core network and the existing cloud monitoring technology. The local status information of compute nodes is monitored using a solution built upon the monitoring database Prometheus, node monitoring unit Node-exporter, and visualization tool Grafana. The CSMF periodically obtains monitoring information from the Prometheus database and dynamically schedules computing and communication resources based on the information.



Figure 4 vivo UE verification platform

4.2 Computing Resource Evaluation Indicators and Scheduling Policies

This section describes the computing resource evaluation indicators and scheduling policies used by the prototype for the convergence of mobile networks and computing. Computing resource evaluation indicators are important to the integration of computing and mobile communication. A unified computing power model is fundamental to ensuring effective network access, management, and operation of compute nodes. The evaluation indicators of computing resources for the prototype are classified into two categories: computing performance and network transmission performance. The computing performance is represented by the computing amount, computing speed, and usage of a node, involving the specifications and usage of main hardware such as CPU, memory, and GPU. The network transmission performance is represented by parameters such as network bandwidth, latency, and jitter between the compute node and UE, as well as the current network load status. The network transmission performance parameters are related to the physical distance between the compute node and UE, the current network route, and the degree of congestion.

In the preceding two categories of evaluation indicators, load statuses in the network transmission performance and computing performance change with time. Therefore, a set of dynamic evaluation indicators for real-time monitoring needs to be established so that proper compute nodes can be allocated to computing services. In addition, different applications have varying requirements for computing resources. Compute-intensive applications, such as AI model training and inference, tend to use nodes with higher computing performance. For example, such applications prefer compute nodes with a higher CPU frequency and more CPU and CUDA cores. For bandwidth-intensive applications such as real-time video streaming and cloud gaming, network transmission performance indicators such as higher bandwidth, lower latency, and less jitter are more important. Therefore, different evaluation indicators and scheduling policies need to be designed for different computing services.

Computing service requirements may be rigid or elastic in actual application scenarios. For instance, the requirement for GPU memory is usually rigid for computing tasks such as AI model training or inference. If the requirement cannot be met, the computing service will be unavailable. For GPUs that meet the memory requirement, a GPU with higher

performance usually delivers higher computing service quality. However, a GPU with lower performance can also support the computing service. Therefore, both rigid and elastic requirements should be considered in the design of resource evaluation indicators and scheduling policies to ensure that services can run properly.

To sum up, a set of diverse service requirements should be designed, and whether the requirements are rigid or elastic should be easily distinguishable. The following formula indicates the dynamic evaluation indicators and scheduling policies based on real-time monitoring:

$$S = \prod_{i=0}^n r_i \cdot (\omega^T \cdot f)$$

where f and w are n -dimensional vectors, with n representing the number of quantifiable computing resource evaluation parameters. f is the normalized parameter vector corresponding to elastic requirements, such as CPU frequency or memory. w is the weight vector of elastic requirement parameters, which varies with service types. r_i indicates whether each rigid requirement is met, and its value is **0** or **1**. It should be noted that the evaluation parameters corresponding to rigid and elastic requirements dynamically change with factors such as load and network status. In other words, the preceding formula is time-varying.

After computing resources are evaluated based on the preceding principles, they are sorted in descending order to obtain a list of optimal compute nodes for the computing service. The prototype for the convergence of mobile networks and computing uses such a method. In addition to computing service quality for UEs, the power consumption of network should also be considered in the design of scheduling policies in actual scenarios. Assume that in an extreme case, there are n nodes with the same computing performance and same network transmission performance in the network. When n same computing service requests are being processed, if only optimal service quality for UEs is used as the scheduling criterion, each UE tends to occupy one compute node. As a result, none of the compute nodes in the network enter the sleep state, and thus, power consumption increases significantly. This problem is prominent in a large-scale network. Therefore, scheduling policy design should achieve a trade-off between overheads such as UE service quality and network power consumption to strike a balance between service quality and costs.

4.3 Demonstration Case and Test Results

This section uses AI-based real-time object detection as a use case to demonstrate the prototype described in section 4.1. By comparing video processing done remotely on a communication and computing integrated network with local processing on the UE, we have demonstrated the advantages of convergence of mobile networks and computing in this application scenario. The prototype passed the mobile computing network certification test by the IMT-2030 (6G) Promotion Group in 2023.

In the demonstration case, the UE implements real-time object detection for the video collected by a camera and labels the identified objects and persons in the video. AI-based real-time object detection is usually used in scenarios such as real-time facial recognition, obstacle detection, and security surveillance. Due to limitations such as UE hardware capabilities and power consumption, local processing directly on the UE may not achieve the expected result. For example, the latency may not meet real-time requirements, and AI models with higher accuracy may not run due to hardware limitations. Moreover, local processing on the UE may increase the power consumption of the UE and affect the computing resources of other applications. If the UE invokes computing resources on the network, the preceding problem can be resolved to better complete AI real-time object detection. In this case, the UE can request to offload computing to the network and utilize the high-performance GPU on the server to complete the processing. The test case described in this section demonstrates the implementation of this process as follows:

- After being started, the compute node automatically registers with the CSMF network element of the core network and periodically maintains its own information.
- When the vivo UE verification platform receives a real-time object detection request, the UE uses local computing capability.
- If the local computing capability cannot meet the requirements or the local capability does not support this type of service, the host computer instructs the UE to initiate a special PDU session establishment request to the network through an AT command, calling for providing the real-time object detection service and carrying computing workloads.
- The CSMF network element in the core network selects a proper compute node based on the computing request

of the UE and starts computing on the selected compute node. In this prototype, computing workloads are real-time image detection and video stream pushing and pulling.

- The compute node establishes a computing connection to the UE. By using the video stream pushing method, the UE obtains the real-time object detection result from the network to implement fast real-time object detection with higher accuracy.

In the test, the computing capability of the compute node in the 6G network is higher than that of the local computing hardware of the UE. Figure 5 shows the performance comparison between the network-side computing and local computing on the UE. The E2E latency is the sum of the bidirectional transmission latency and frame processing

latency. Due to the limitations of the CPU and GPU performance, the frame processing latency on the UE is higher than that on the network. However, because network transmission is not involved, the transmission latency is approximately 0. Thus, the E2E latency of computing on the network is lower than that on the UE. The computing latency gain of frame processing on the network is greater than the latency overhead of bidirectional transmission. Therefore, the computing latency on the network is lower than that on the UE. In addition, since local computing is subject to the UE hardware, network-side computing can support an AI model with a larger quantity of parameters, implementing object detection with higher accuracy.

In this instance, network-side computing can achieve a lower average latency and less jitter than local computing.

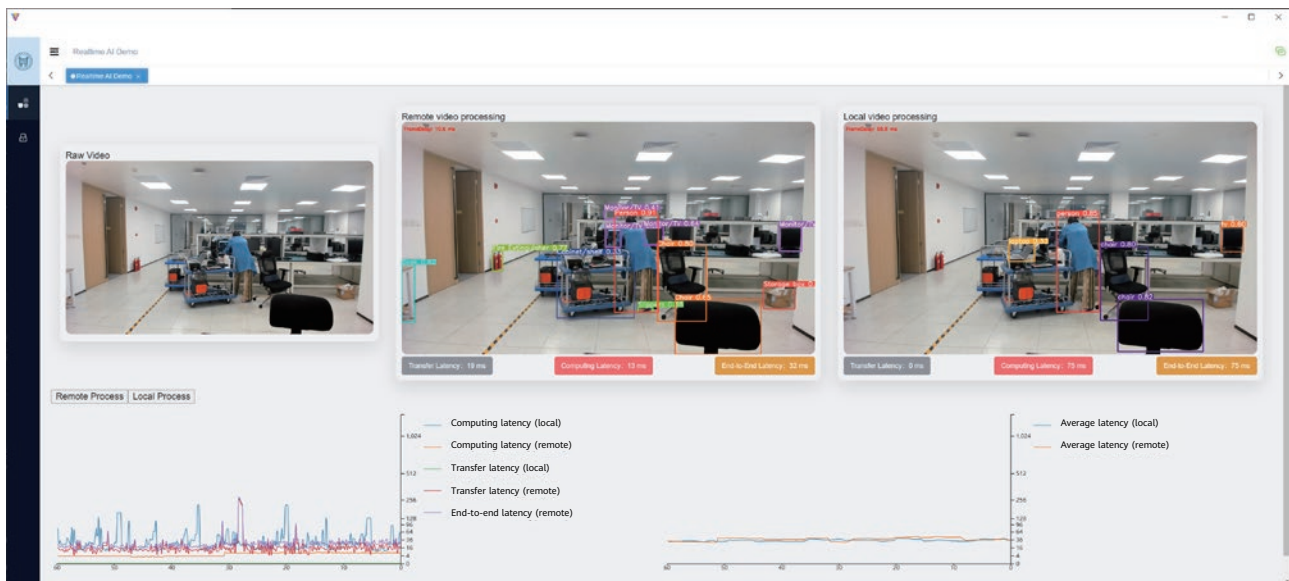


Figure 5 Demonstration of real-time object detection

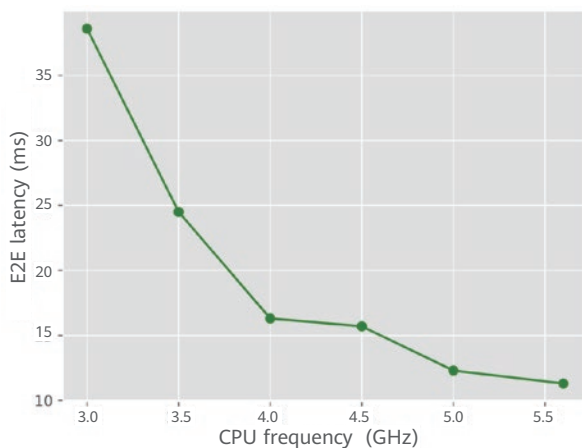


Figure 6 Relationship between the E2E latency and CPU frequency

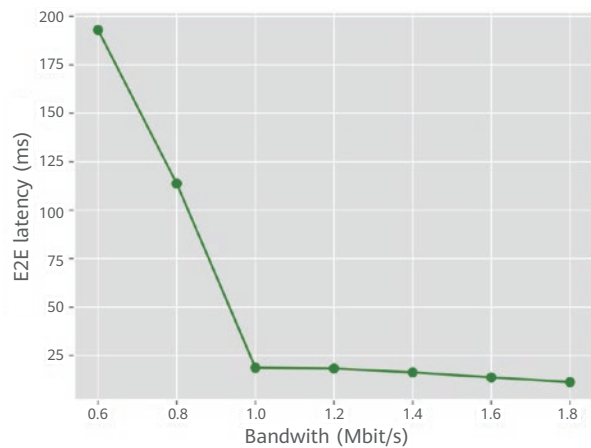


Figure 7 Relationship between the E2E latency and bandwidth

To further illustrate the impact of node computing performance and network transmission performance on the E2E latency or service result, we fix the video stream resolution and object detection model and vary the CPU frequency and network bandwidth of the compute node. Figure 6 and Figure 7 show the latency curves varying with them.

It can be learned from Figures 6 and 7 that the processing latency gradually decreases along with the CPU frequency increase. However, when the network bandwidth is lower than the specific threshold, performance is severely limited, unable to complete the service normally. When the network bandwidth is higher than the threshold, there is no obvious gain in performance along with bandwidth increase. In addition, the latency of local computing is not always higher than that of computing over the 6G network. When the network bandwidth is low (e.g., 0.6 Mbps), or the performance of the compute node in the network is slightly different from the local computing performance of the UE, the transmission latency introduced by computing over the 6G network is higher than the decrease in computing latency. Therefore, local computing on the UE is more suitable in this instance. It should be noted that a common UE is usually based on the Arm architecture, and the computing power of the vivo UE verification platform in the prototype is much higher than that of the common UE in real-time object detection. Therefore, in actual practice, the UE computing latency will be higher than that of the prototype test result described in this paper. In other words, offloading computing to the network could result in a greater latency gain.

5 Summary and Prospects

The AI and communication usage scenario introduces new services beyond communication in 6G. We recommend a method to establish performance indicators for computing services based on typical use cases (such as virtual humans), including computing power density, connection density for computing, peak computing power per UE, and computing latency per UE. To support AI and communication applications, this paper suggests that potential technical directions for convergence of mobile networks and computing in 6G include the evolution-based path and revolution-based path. The technical direction of the evolution-based path includes enhanced EC and the enhanced IMS. The computing control function,

compute node, and computing data transmission channel (i.e., computing bearer/session) are introduced in the technical direction of the revolution-based path. Finally, the technical direction of the revolution-based path is verified by a prototype for the convergence of mobile networks and computing. A comparison test of AI real-time object detection based on UE requests shows that the AI and communication services provided by a 6G network deliver higher identification accuracy and shorter E2E latency than local computing on the UE.

Along with the development from 1G to 5G, mobile communication technologies have been verified in commercial applications. However, challenges still lie in 6G AI and communication applications. Currently, high-end mobile phones can support virtual humans with up to 100,000 faces and large models with up to 7 billion parameters for inference. Therefore, new applications with a larger computing workload on the UE, especially applications with ultra-low latency and high computing workload requirements, can better demonstrate the strengths of 6G AI and communication services. Additionally, 6G AI and communication services are one of the ways to improve user experience of new types of UE (e.g., XR devices, robots, and unmanned vehicles). If new types of UE can deliver better user experience in the future, they will be widely adopted, like mobile phones. When new types of UE gain popularity, they will lay a solid ecosystem foundation for the extensive application of 6G AI and communication. Computing as a service will unlock boundless possibilities for UE in the future.

References

- [1] ITU-R, "Framework and overall objectives of the future development of IMT for 2030 and beyond," November 2023.
- [2] China Academy of Information and Communications Technology (CAICT), *Cloud Computing White Paper*, July 2023.
- [3] Dinh C. Nguyen, P. Cheng, M. Ding, D. Lopez-Perez, P. N. Pathirana, J. Li, A. Seneviratne, Y. Li, and H. V. Poor, "Enabling AI in future wireless networks: A data life cycle perspective," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 553-595, September 2020.
- [4] vivo Communications Research Institute, "6G Services, Capabilities and Enabling Technologies [R]," 2022.
- [5] Jiang Dajie, Yuan Yannan, Zhou Tong, *et al.*, "6G-oriented services integrating communication, sensing and computing, system architecture, and key technologies [J]," *Mobile Communications*, 2023, 47(03): 2-13.
- [6] vivo Communications Research Institute, "6G Network Architecture [R]," 2023.
- [7] 3GPP TS 23.548 V18.5.0, "5G system enhancements for edge computing," March 2024.
- [8] 3GPP TS 23.228 V18.5.0, "IP multimedia subsystem (IMS)," March 2024.



Ten Issues of NetGPT

Wen Tong¹, Chenghui Peng¹, Tingting Yang², Fei Wang¹, Juan Deng³, Rongpeng Li⁴, Lu Yang¹, Honggang Zhang⁴, Dong Wang⁵, Ming Ai⁶, Li Yang⁷, Guangyi Liu³, Yang Yang⁸, Yao Xiao¹, Liexiang Yue³, Wanfei Sun⁶, Zexu Li⁵, Wenwen Sun⁷

¹ Huawei Wireless Technology Lab

² Peng Cheng Laboratory

³ China Mobile Research Institute

⁴ Zhejiang University

⁵ Research Institute of China Telecom

⁶ CICT Mobile Communication Technology Co., Ltd.

⁷ ZTE Corporation

⁸ The Hong Kong University of Science and Technology (Guangzhou)

Abstract

With the rapid development and application of foundation models (FMs), it is becoming increasingly clear that FMs will be a cornerstone of future mobile communications. As current artificial intelligence (AI) algorithms applied in mobile networks are dedicated models that aim for different neural network architectures and objectives, we are faced with a series of challenges related to generality, performance gains, management, collaboration, etc. In this paper, we introduce NetGPT (Network Generative Pre-trained Transformer) — a group of foundation models for mobile communications, and present a comprehensive overview of ten issues regarding the design and application of NetGPT.

Keywords

NetGPT, foundation models, wireless, mobile communications, ten issues

1 Introduction

1.1 Background

The International Telecommunication Union – Radiocommunication Sector (ITU-R) has proposed "AI and Communication" as one of the most important usage scenarios for International Mobile Telecommunications towards 2030 and beyond (IMT-2030) [1]. In recent years, many researchers have focused on this field and achieved promising results that benefit various components of mobile networks, including radio access network (RAN), core network (CN), operation, administration, and management (OAM), and user equipment (UE). For example, deep deterministic policy gradient (DDPG) is used to generate policies for CN, and deep Q-learning (DQN) is applied to network OAM. However, the current paradigm of tailoring AI algorithms for each specific use case can pose many problems, such as low generality, limited performance gains, complicated management, and difficulty in collaboration [2–4].

Foundation models (FMs) are a breakthrough development in AI techniques, with applications across a wide range of fields. The most well-known FMs are large language models (LLMs), which demonstrate powerful capabilities in tasks such as chatting and programming. FMs with a relatively unified neural network architecture are expected to drive mobile networks toward high generality, ultimate performance, simplified management, convenient collaboration, multi-task processing, etc. To distinguish from other industrial applications, we have named the group of FMs for mobile communications as *NetGPT*. However,

the design and application of NetGPT is still in its infancy. To advance the design of models and mobile network architectures and support efficient NetGPT applications, this paper summarizes ten fundamental issues of NetGPT.

1.2 Definition of NetGPT

A mobile network consists of several technical components, including RAN, CN, and OAM. These components differ significantly in terms of functional features, data structures, performance requirements, etc., and therefore require different NetGPT models. For example, the OAM component deploys a type of NetGPT model that can be fine-tuned directly from LLMs via parameter-efficient fine-tuning (PEFT). When deployed at edges, another type of NetGPT model (which is smaller in size) can be generated by distilling or pruning LLMs. Additionally, a new NetGPT model can be trained from scratch with a neural network architecture similar to LLMs. Therefore, NetGPT is not a single model that covers all mobile communications scenarios, but rather a series of models that cater to different technical components and vendors.

In this paper, we define three layers of agents for NetGPT: Layer 0 (L0), Layer 1 (L1), and Layer 2 (L2). The NetGPT-L0 agent is a large network-wide model. The NetGPT-L1 agents refer to FMs specific to different technical components, such as RAN, CN, and OAM. The NetGPT-L2 agents are focused on more specialized scenarios. As depicted in Figure 1, for the RAN component, we have a NetGPT-L2 agent for the physical layer (PHY), and for the OAM component, we have a NetGPT-L2 agent focused on network optimization.

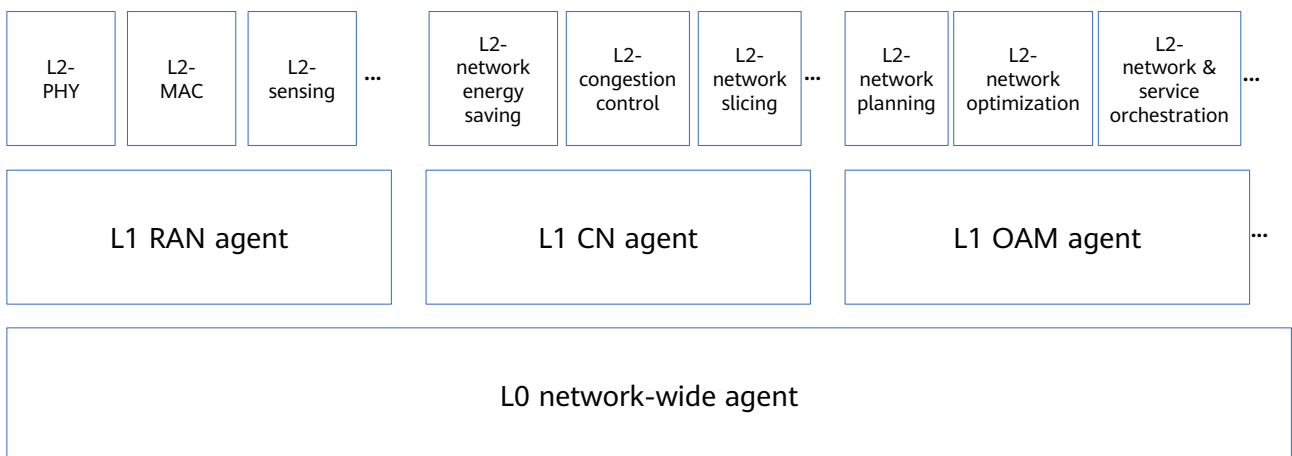


Figure 1 Three layers of agents for NetGPT

2 Ten Issues of NetGPT

In this section, we present NetGPT's issues in the context of 6G wireless communications. These issues can be divided into two categories. The first category pertains to NetGPT's design. The second category concerns the design of future mobile network architectures that will facilitate NetGPT's applications, including RAN, CN, and OAM.

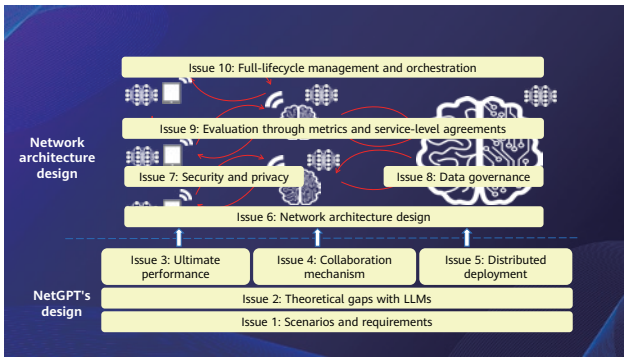


Figure 2 Ten issues of NetGPT

Issue 1: Scenarios and Requirements

Trained on enormous datasets and billions of parameters, NetGPT requires significant computing power. However, unlike the cloud, which is equipped with centralized and powerful computing capability, wireless networks are heterogeneous and distributed, comprising multiple edge devices and mobile terminals. Therefore, deploying cloud AI models/algorithms directly to wireless networks is not feasible. Instead, we need to redesign algorithms to adapt to the characteristics of wireless networks and redesign the network architecture to support NetGPT natively. It is important to note that lower layers in wireless networks require higher quality of service (QoS), including real-time performance and accuracy. However, LLMs' complications, such as hallucination and extra-large size, can impede the fulfillment of QoS. This raises a key question: is there a boundary for applying NetGPT in wireless networks? For instance, is NetGPT only applicable to the higher air interface layer and not the physical layer? Such boundary issues also influence the potential role that NetGPT may play in each specific application. For example, to what extent can NetGPT enable future OAM systems — fully or partially autonomous networks [5]? When studying NetGPT, researchers must clarify its scenarios and boundaries.

Issue 2: Theoretical Gaps with LLMs

As the most representative FMs, LLMs are considered the basis of NetGPT [6–8]. However, there are many essential differences between the communications and natural language processing (NLP) fields, leading to theoretical gaps between NetGPT and LLMs. The primary differences include:

- **Data features:** The datasets used by NetGPT contain communication information (e.g., channel information). Such information is represented in the form of high-dimensional tensors rather than tokens, as in LLMs.
- **Backend task:** Wireless networks deal with various types of tasks, which means that the outputs of NetGPT may take different forms instead of tokens that serve as both the input and output of LLMs.
- **Model size:** NetGPT defines a multi-layer hierarchy, with models of varying sizes deployed at each layer. Some NetGPT models, especially the NetGPT-L2 models deployed at network edges (e.g., base stations), may have only 0.1–1 billion parameters, which is much less than the 5–200 billion parameters in typical centralized LLMs.

Several questions require further investigation, such as whether NetGPT's neural network architecture is consistent with that of LLMs and whether NetGPT can trigger evolutionary innovations in AI theories and the associated neural network architecture.

Issue 3: Ultimate Performance

A prerequisite for applying NetGPT to wireless networks is meeting the ultimate performance requirements (e.g., real-time inference and reliability) of future wireless networks like 6G. These requirements will be much higher than what current FM applications can achieve.

Compared to 5G, future wireless networks must achieve a 10x improvement in communication performance and also support the new services assured by ultimate performance, such as autonomous driving and industrial robots. NetGPT can achieve these goals for future wireless networks by integrating AI and communication. Specifically, NetGPT must be able to derive inference results in a very short time and adapt to the highly dynamic environment of wireless networks in real time. However, due to complex computing

procedures and a large number of parameters involved, it is challenging for NetGPT to achieve 0.1 ms real-time inference in future wireless networks. More efficient model algorithms and inference acceleration methods are therefore desired.

Furthermore, future wireless networks will demand exceptionally high reliability [1]. Because FMs' hallucination problem can lead to incorrect network decisions and unpredictable risks, most of the current FMs only play an auxiliary role and lack direct application to network service functionalities. To meet the exceptionally high requirements for reliability, methods to enhance the reliability of NetGPT should be explored from the perspectives of data quality, model structure, knowledge graph, etc.

Issue 4: Collaboration Mechanism

Given the excessive computing, communication, and energy resources that FMs may consume, models with smaller sizes, such as NetGPT-L2, may be better suited for the edges of wireless networks (e.g., base stations). While NetGPT-L2 is advantageous for dealing with specific scenarios that require local knowledge, it may lack the necessary generality under certain circumstances. This issue can be resolved by collaborating with the centrally deployed NetGPT-L0/L1.

Therefore, the deployment of NetGPT in future wireless networks will emphasize collaboration between NetGPT models of various scales, which involves collaborative training and inference. Specifically, collaborative training scenarios include: (1) NetGPT-L0/L1 integrating local data to elaborate locally tailored NetGPT-L2 via transformation, distillation, pruning, etc; (2) NetGPT-L0/L1 aiding the training process of NetGPT-L2 as an auxiliary factor; and (3) feedback from NetGPT-L2 further boosting NetGPT-L0/L1. Collaborative inference scenarios include: (1) real-time inference being accomplished collaboratively with NetGPT-L0/L1 when NetGPT-L2 is unable to reach the desired confidence level independently and (2) real-time inference being accomplished collaboratively by multiple NetGPT-L2 models at each device.

In these scenarios, some key algorithms require further investigation, including parameter pruning according to relevance from NetGPT-L0 parameters and efficient fine-tuning to accommodate new tasks. In addition to the algorithmic challenges, a standardized collaboration mechanism needs to be defined to support cross-vendor collaboration of NetGPT. This mechanism should include standardization of collaboration content (e.g., functionalities

and procedures), generation methods of collaboration sets, and systematic control of collaboration incidents.

Issue 5: Distributed Deployment

Regarding device-edge-cloud collaboration, a full or partial version of NetGPT-L0, L1, or third-party FMs may need to be deployed at wireless network edges or mobile devices based on service requirements. Therefore, we need to determine how to split FMs to efficiently accommodate the heterogeneity of the dynamic device-edge-cloud environment in wireless networks. As each node may have different computing resources, storage capabilities, and transmission rates, the split should be determined based on actual availabilities to achieve optimal load balancing and resource utilization.

To ensure parameter consistency between wireless network edges, we need to explore a cross-node incremental training mechanism. During incremental training, an FM may encounter inconsistency if there is a delay in parameter updates for distributed subnets in the network.

Another core factor is efficient communication among distributed nodes. One approach is to optimize algorithms, such as model compression (e.g., via pruning and quantization), to reduce communication overhead. Another approach is to refine the interfaces and protocols to enhance the efficiency of data flow among nodes.

Issue 6: Network Architecture Design

Integrating NetGPT into future wireless networks will have an evolutionary impact on various aspects of network architectures:

- Network elements: NetGPT can fully or partially realize existing network functions. However, as NetGPT is widely applied to end-to-end scenarios, certain classic network functions will become outdated and require thorough reorganization or redesign, including network element classification for the current 5G core network, as well as the way they are organized (e.g., service-oriented architecture) [9].
- Network interfaces/protocols: Model-based collaborative interfaces (e.g., tokenized interfaces) among NetGPT will replace classic interface protocols built upon standardized signals and element strings.

- **New network capabilities:** Future wireless networks will natively support NetGPT and third-party FMs, allowing online updates and the evolution of various NetGPT models. To accommodate specific scenarios, FMs require incremental training via PEFT, such as low-rank adaptation (LoRA). In contrast, NetGPT requires incremental learning of new knowledge in addition to general FM capabilities, necessitating a life-long learning technique. Due to NetGPT's vast amount of parameters and data, data parallelism (e.g., federated learning) and model parallelism (e.g., split learning) may not be sufficient. Therefore, it is worthwhile to explore the emergent ability of both.
- **New network services:** Based on FMs' ability to interpret semantic information, future wireless networks will become exclusive to individual applications. This will allow services to be provided according to various perspectives, including business logic, network logic, and network resources.

Issue 7: Security and Privacy

When NetGPT is applied in end-to-end wireless networks, its security becomes a major concern. The involvement of a huge amount of data and parameters in FMs creates attack vulnerabilities for NetGPT, particularly the deliberate implant of backdoors, which can potentially confuse NetGPT with illegal instructions. Additionally, as NetGPT becomes larger, it may become more biased and untrustworthy. To prevent all these issues, NetGPT requires a tailored security design.

The lack of interpretability in NetGPT, like other AI algorithms, poses risks to its application in networks. Although some important research theories have been proposed, a well-established theoretical framework is still lacking, particularly regarding the mathematical and analytical approaches for quantitative analysis of FMs.

Furthermore, the use of NetGPT may expose users' privacy data, such as user account information, dialog history, and information obtained during interactions, to suppliers, service providers, and affiliates. Multiple data leakage incidents have drawn close attention from global regulatory agencies to the data privacy risks associated with FMs. Hence, there is an urgent need to develop data privacy principles and usage regulations for NetGPT.

Issue 8: Data Governance

As the performance of NetGPT highly depends on data quality, future wireless networks require data governance services specifically designed for NetGPT [10].

- The data fed into NetGPT comes from various technical components and parties. To efficiently process such heterogeneous data in massive volumes, future wireless networks need a comprehensive framework for data governance in terms of data proprietary rights, formats, qualities, privacy, etc.
- Future wireless networks need a large-scale distributed storage and real-time provisioning mechanism for NetGPT inference and online updates. Designing data services with guaranteed QoS is particularly challenging for NetGPT, which demands ultimate performance.
- To boost the reliability of NetGPT and reduce hallucination, future wireless networks need to establish a network knowledge graph.

In summary, a unified data governance framework needs to be designed for different types of NetGPT models in future wireless networks.

Issue 9: Evaluation Through Metrics and Service-Level Agreements

Evaluating NetGPT's performance comprehensively and objectively has become an urgent problem. The evaluation result can provide a strong basis for optimizing and improving NetGPT, thereby enhancing its application effect and commercial value. Additionally, the evaluation can serve as a benchmark for understanding the performance and applicability of NetGPT models provided by different vendors.

In addition to existing evaluation metrics like accuracy, F1 score [11], and bilingual evaluation understudy (BLEU) [12], more metrics (e.g., network function correctness and communication task success rate) should be formulated based on network characteristics. The combination of these specific metrics can help evaluate NetGPT's performance in specific scenarios in a more refined manner. For in-network tasks, understanding and applying domain-specific terms, concepts, and rules are crucial to ensure reliable evaluation results.

Unlike the language domain, the network is relatively closed, and only limited annotated data can be collected publicly. The difficulty in collecting small sample data results in an incomplete scope of possible scenarios encountered during model training. Therefore, it is crucial to evaluate NetGPT's generality to different network scenarios.

Issue 10: Full-Lifecycle Management and Orchestration

NetGPT models from various vendors can be deployed in the same network. However, to ensure efficient orchestration and management of these models throughout their entire lifecycle, a unified mechanism must be established. This mechanism must enable tasks such as adding, updating, transferring, and removing NetGPT models while ensuring that the intellectual property rights of NetGPT models are well-preserved during this process. Because the owners of NetGPT models may not necessarily be network operators, the owners may not be willing to relinquish control of their models to the operators. Therefore, a well-balanced scheme is required to protect the interests of both parties.

Another challenge is organizing and scheduling these NetGPT models. To address this, we must standardize the model language across vendors to establish consistent interfaces and interaction methods. Additionally, deploying NetGPT requires a multidimensional approach that considers various resources, including connections, computing power, and storage. Therefore, it is a delicate task to orchestrate NetGPT models and network resources based on scenario characteristics and requirements to improve system performance and resource utilization.

3 Conclusion and Prospect

While FMs have the potential to revolutionize future wireless networks, many related research directions and FM standards demand further investigation. NetGPT represents the deep bidirectional convergence between wireless networks and FMs. Our paper proposed ten issues of NetGPT, including fundamental theories, scenario requirements, network architectures, deployment management and control, and data governance. In-depth research on NetGPT is essential to advancing the integration of FMs in future wireless networks.

References

- [1] ITU-R, "Framework and overall objectives of the future development of IMT for 2030 and beyond," 2023.
- [2] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.
- [3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.
- [4] Y. Yang, M. Ma, and H. Wu, "6G network AI architecture for everyone-centric customized services," *IEEE Network*, pp. 1–10, 2022.
- [5] T. McElligott, "Network automation using machine learning and AI," TMForum, 2020.
- [6] Y. Chen, R. Li, Z. Zhao, C. Peng, J. Wu, E. Hossain, and H. Zhang, "NetGPT: A native-AI network architecture beyond provisioning personalized generative services," arXiv e-prints, p. arXiv:2307.06148, July 2023.
- [7] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large language models for telecom: The next big thing?," arXiv e-prints, p. arXiv:2306.10249, June 2023.
- [8] Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, J. Zhang, and K. B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," arXiv e-prints, p. arXiv:2307.02779, July 2023.
- [9] 6GANA, "Ten questions of 6G native AI network architecture," <https://www.6g-ana.com/>, 2022.
- [10] 6GANA, "6G data service - concept and requirements," <https://www.6g-ana.com/>, 2022.
- [11] C. J. Van Rijsbergen, "Information retrieval (2nd ed.)," Butterworth-Heinemann, 1979.
- [12] K. Papineni, S. Roukos, and T. Ward, "BLEU: A method for automatic evaluation of machine translation[C]," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318, 2002.



Key Issues and Technological Explorations of Large Models in 6G Network Communications

Tingting Yang, Ping Zhang, Mengfan Zheng, Nan Li, Shuai Ma
Peng Cheng Laboratory

Abstract

This paper describes the technical challenges and future development direction of large models in 6G network communications. We start by outlining the research background and current research status of large models in 6G network communications and discussing the key issues and challenges of the models. We then explore generative artificial intelligence (GenAI)-enabled integrated sensing and communication (ISAC) and GenAI-enabled semantic communication, and introduce the development progress of 6G large models. Finally, we introduce the progress of the "Large Generative AI Models in Telecom Emerging Technology Initiative (GenAINet ETI)" founded at the IEEE Communications Society (ComSoc).

Keywords

6G, large models in network communications, GenAI

1 Introduction

The IMT-2030 6G vision recently announced by the International Telecommunication Union – Radiocommunication Sector (ITU-R) has defined a new usage scenario — "AI and Communication." Future wireless networks will undergo a fundamental transformation, evolving from a simple infrastructure that provides only connectivity services to an intelligent system that natively supports computing, data, artificial intelligence (AI) functions, and integrated communication and computing. Figure 1 illustrates the development history and future prospects of "Communication + AI" from 2020 to 2030. In the "Communication + AI 1.0" era, which integrated AI with communication, most research was conducted during the 5G-A period between Release 17 and Release 18. It was during this period that the trends and vision of 6G were initially formulated. Starting from 2023, we have entered the era of "Communication + AI 2.0", the stage of integrating large models into communication. The 6G research and proposal phase will continue from Release 19 until 2026. Following this, 6G technology research will commence in Release 20, with 6G standards being formulated in Release 21 around 2028. Ultimately, 6G will be commercially available in 2030. Candidate technologies will be submitted throughout this process, with technological standards and specifications being developed simultaneously. This will comprehensively transform the communication network into an intelligent system.

Research on the integration of AI and wireless communication has been ongoing for over five years since the "Communication + AI 1.0" era [1]. During this period, significant progress was made in integrating AI with wireless network architecture, such as "N-layer, N-plane", and related algorithms. However, this stage gradually stagnated. Recently, with the emergence of large model technologies, the era of "Communication + AI 2.0" has kicked off. The rapidly developing large model technologies [2-4] are expected to become an integral part of the information

and communications technology (ICT) infrastructure, along with 5G/6G. They have the potential to address pain points from the "Communication + AI 1.0" era, such as low generalizability of models, high cost of customized samples, weak commonsense reasoning, and inability to process complex tasks.

2 Key Issues of Large Models in 6G Network Communications

As a cutting-edge technology in the AI field, generative pre-trained large models will play an important role in enhancing wireless network serviceability, optimizing resource configuration, enabling ubiquitous intelligent connectivity, etc. This technology has the potential to transform the design paradigm of the next-generation wireless communication network. In addition, future wireless networks will also undergo revolutionary changes or enhancements in various aspects (e.g., network architecture, wireless air interface, access network, and core network) to empower a variety of applications based on generative pre-trained large models.

However, the current research of large models faces numerous challenges. Experts from different industries, including communication, network, computing, mathematics, and AI, engaged in extensive discussions and summaries during the 6GANA NetGPT Symposium and the Third Symposium on the Theories and Technologies for Communication and Computing Integrated Networks. This culminated in 6GANA publishing a white paper titled "Ten issues of NetGPT" [5]. The white paper elaborates on the ten crucial research issues of Network Generative Pre-trained Transformer (NetGPT). The issues are: NetGPT's scenarios and requirements, theoretical gaps, ultimate performance, collaboration mechanism, distributed deployment, network architecture design, security and privacy, data governance, evaluation through metrics and service-level agreements, and full-lifecycle management and orchestration. In addition to explaining these issues systematically, the white paper

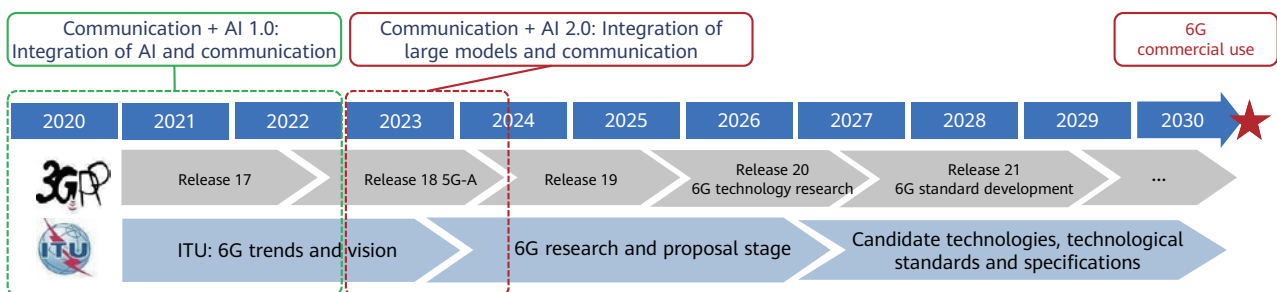


Figure 1 Evolution history of the integration of AI and communication technologies

also delves into the specific connotations and challenges of each issue. For example, in terms of scenarios and requirements, the white paper highlights that NetGPT must be adaptable to various complex network environments in order to fulfill the demands of different application scenarios, such as enhancing network performance and optimizing intelligent management and control. Regarding theoretical gaps, the white paper underlines the inherent differences between NetGPT and large language models (LLMs) and suggests developing a dedicated model architecture for the communication field to improve the generalizability and processing efficiency of NetGPT.

Furthermore, the white paper also pays special attention to issues such as performance requirements, collaboration, distributed deployment, network architecture design, security and privacy, data service, evaluation, and full-lifecycle management. Specifically, in terms of performance requirements, it emphasizes NetGPT's rigorous demands for real-time performance, reliability, availability, flexibility, and scalability, and proposes corresponding optimization strategies, such as enhancing inference efficiency through efficient hardware acceleration and model compression optimization. In terms of collaboration, the white paper discusses the collaborative evolution between large and small models, and proposes an effective mechanism for model capability transfer and collaboration between edges and cloud. For distributed deployment, it proposes specific solutions such as model splitting, distributed training, and efficient inter-node communication mechanism. Concerning network architecture design, it recommends deep integration of communication, computing, data, and AI algorithm models at the architecture level in order to improve the efficiency of

data processing, decision-making, and inference. Regarding security and privacy, it explores the specific challenges and solutions related to model reliability, explainability, and privacy protection. With respect to the data service, the white paper emphasizes the necessity of providing efficient data support for NetGPT, which includes processing a vast amount of heterogeneous data, implementing distributed deployment and real-time provisioning of large-scale data, and building a network knowledge graph to enhance NetGPT's inference capability. As for evaluation, it proposes new evaluation metrics and methods for the network field and highlights the importance of comprehensive evaluation of NetGPT's performance and security. Finally, the white paper discusses the full-lifecycle management and orchestration issue of NetGPT, and proposes a specific framework for task decomposition, task orchestration, and task execution to enable efficient application of NetGPT in various scenarios. In summary, the white paper offers systematic guidance for the research and application of NetGPT, with the goal of comprehensively empowering 6G communication networks with NetGPT.

3 Technological Explorations of Large Models in 6G Network Communications

Peng Cheng Cloud Brain is a major network intelligence infrastructure situated in China. It is dedicated to delivering a high-performance computing platform (as illustrated in Figure 2). In 2019, Peng Cheng Cloud Brain I was launched as

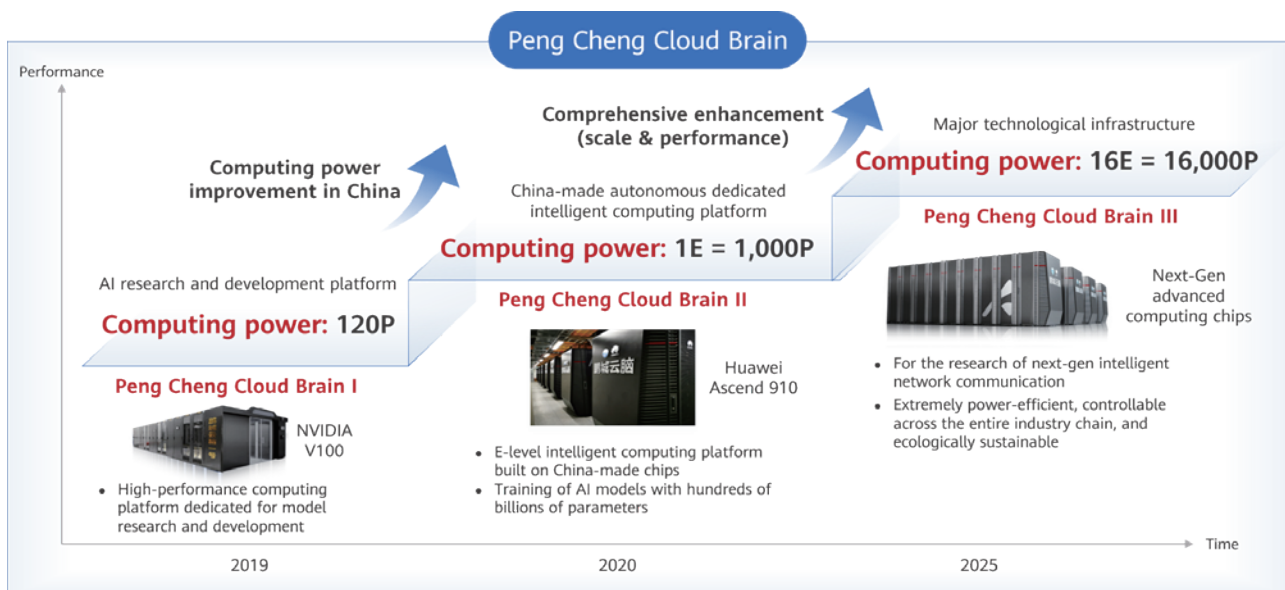


Figure 2 Peng Cheng Cloud Brain, a major network intelligence infrastructure

a high-performance computing platform with 120P FLOPS, specifically designed for model research and development. In 2020, Peng Cheng Cloud Brain II was introduced, an E-level intelligent computing platform built on China-made chips, boasting 1E FLOPS of computing power and supporting the training of AI models with hundreds of billions of parameters. Peng Cheng Cloud Brain III, scheduled for release in 2025, will further increase computing power to 16E FLOPS. It will provide computing capabilities that are extremely power-efficient, operable across the entire industry chain, and ecologically sustainable. Thanks to these strengths, it will facilitate the research of next-generation intelligent network communication. Serving as a model platform in Peng Cheng Cloud Brain, *Peng Cheng Mind* covers various applications such as 200B-parameter Chinese models and 33B-parameter general models. It significantly enhances the performance of AI models in different fields. Peng Cheng Mind offers cross-domain AI computing power and supports both general intelligence and customized applications, enabling itself to process complex tasks and meet high-performance computing requirements.

In this section, we begin with an overview of the Peng Cheng Mind technology, which is also a type of large model used in 6G network communications. We then showcase the research progress of large models in 6G network communications with two prominent examples — generative artificial intelligence (GenAI)-enabled integrated sensing and communication (ISAC) and GenAI-enabled semantic communication.

3.1 Overview of Peng Cheng Mind

Utilizing Peng Cheng Cloud Brain and Peng Cheng Mind, our research team is developing large models for the next-generation 6G network. To address the pain points associated with small models, such as low generalizability, high cost of customized samples, weak commonsense reasoning, and inability to process complex tasks, we propose using Peng Cheng Mind to fine-tune domain applications based on operator cooperation data, open-source data, and communication expert instruction data. Furthermore, we explore the intelligent emergence capabilities of large models by scaling models and data from small-scale communication simulation data (10M parameters and 10B tokens) to large-scale simulation and measured communication data (100B parameters and 10T tokens). The purpose is to transition from general to specific and from small to large models, with the ultimate goal of advancing the design and development of large models in 6G network communications.

3.2 Technological Explorations for GenAI-enabled ISAC

The authors' team has conducted mathematical and theoretical research on the unified representation of multimodal data, including 6G network communication data and natural language corpus. Based on the NVIDIA Sienna Neural Radio Framework [6], we have built a simulation platform for generating multipath channels, collecting wireless channel data, and extracting the scattering point information necessary for sensing and imaging. Additionally, we have constructed generative pre-trained large models with 30M and 3.5B parameters, using the Transformer architecture. These models can be used to train and infer sensing and imaging tasks. Moreover, the models support open-source wireless channel datasets such as Sensiverse [7], with the aim of providing the industry with a comprehensive ISAC model testing platform and evaluation benchmark. The models are currently available in our Open Intelligence community, with the open-source code accessible at <https://openi.pcl.ac.cn/Foundation-Model-of-6G-Communication?lang=en-US>.

3.3 Technological Explorations for GenAI-enabled Semantic Communication

GenAI-enabled semantic communication involves extracting features from, encoding, transmitting, and decoding source data from a semantic perspective. Essentially, it is about utilizing computing resources to offset physical resource overheads (bandwidth, power consumption, latency, etc.), with a particular focus on accurate transmission at the semantic level. Its advantage lies in understanding before transmission, which significantly enhances the transmission efficiency of communication systems. Peng Cheng Mind's powerful computing capabilities will help with the design and implementation of GenAI-enabled semantic communication. Compared to traditional syntactic communication, GenAI-enabled semantic communication offers several advantages: 1) Ultra-high compression ratio: Traditionally, data compression performance is restricted by the Shannon entropy limit. In contrast, semantic encoding focuses on extracting and encoding task-related semantic information while ignoring irrelevant information. As a result, its compression performance can exceed the Shannon entropy limit. 2) Strong noise resistance: Semantic communication employs joint source-channel coding, which exhibits greater

noise resistance than current channel coding methods such as low-density parity check (LDPC). This is particularly evident in low signal-to-noise ratio (SNR) conditions, where it can prevent the "cliff effect" commonly associated with conventional channel coding methods. 3) Large network capacity: Unlike the conventional approach of increasing transmission capacity by using more bandwidth, antennas, or power, semantic communication reduces the amount of transmitted data by capitalizing on its ultra-high compression ratio. This leads to a significant enhancement in end-to-end data transmission capacity. GenAI-enabled semantic communication transforms conventional syntactic communication into content-oriented semantic communication, signifying a paradigm shift in the evolution of mobile communication networks and establishing itself as a fundamental theory of 6G mobile communication.

Our team has extensively researched GenAI-enabled semantic communication over our Peng Cheng Mind models. Through this research, we have discovered new semantic dimensions of data at the transmitter end, thus achieving highly abstract representation and intelligent, simplified transmission of information. At the receiver end, we utilize GenAI's powerful content generation capability to restore high-quality source data content, establishing a new technological path to overcome the bottleneck of syntactic communication.

4 About IEEE ComSoc's GenAINet ETI

To promote the ecological innovation and technological development of large models in 6G network communications and establish an international alliance for network communication technology cooperation, the authors' team, as a founding member and academic chair, joined forces with Huawei Technologies Co., Ltd., Khalifa University, China

Mobile, and 6GANA to set up the "Large Generative AI Models in Telecom Emerging Technology Initiative (GenAINet ETI)" [8] at the IEEE Communications Society (ComSoc). As the sole academic platform for technological exchanges and cooperation on network communication models within IEEE ComSoc, GenAINet ETI aims to advance the research on network communication models through collaborative efforts across various disciplines, including mathematics, information theory, wireless communication, AI, signal processing, networking, information security, and more. At present, GenAINet ETI has attracted nearly 200 institutions worldwide, establishing itself as an interdisciplinary cutting-edge innovation platform for academics, researchers, and industry leaders. Moving forward, GenAINet ETI will host international seminars and publish research reports and white papers to promote the technology standard development and adoption for large models in 6G network communications.

5 Conclusion

This paper systematically described the key issues and technological explorations of large models in 6G network communications. It started by highlighting the importance of such models in future wireless networks and providing an overview of the current research status. Then, drawing on the "Ten issues of NetGPT" white paper published by 6GANA, it delved into the ten crucial research issues (including scenarios and requirements, theoretical gaps, network architecture, deployment, and data governance) and analyzed their connotations and challenges. Additionally, this paper demonstrated the explorations and achievements of the authors' team with respect to large models in 6G network communications, and provided a detailed account of the construction of IEEE ComSoc's GenAINet ETI.



References

- [1] Khaled B. Letaief, Yuanming Shi, Jianmin Lu, and Jianhua Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications* 40, no. 1 (2021): 5–36.
- [2] Hang Zou, Qiyang Zhao, Lina Bariah, Yu Tian, Mehdi Bennis, Samson Lasaulce, Merouane Debbah, and Faouzi Bader, "GenAINet: Enabling wireless collective intelligence via knowledge transfer and reasoning," *arXiv preprint arXiv:2402.16631* (2024).
- [3] Hongyang Du, Guangyuan Liu, Dusit Niyato, Jiayi Zhang, Jiawen Kang, Zehui Xiong, Bo Ai, and Dong In Kim, "Generative AI-aided joint training-free secure semantic communications via multi-modal prompts," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12896–12900. IEEE, 2024.
- [4] Yifei Shen, Jiawei Shao, Xinjie Zhang, Zehong Lin, Hao Pan, Dongsheng Li, Jun Zhang, and Khaled B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Communications Magazine* (2024).
- [5] Wen Tong, Chenghui Peng, Tingting Yang, Fei Wang, Juan Deng, Rongpeng Li, Lu Yang, *et al.*, "Ten issues of NetGPT," *arXiv preprint arXiv:2311.13106* (2023).
- [6] Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller, "Sionna: An open-source library for next-generation physical layer research," *arXiv preprint arXiv:2203.11854* (2022).
- [7] Jiajin Luo, Baojian Zhou, Yang Yu, Ping Zhang, Xiaohui Peng, Jianglei Ma, Peiying Zhu, Jianmin Lu, and Wen Tong, "Sensiverse: A dataset for ISAC study," *arXiv preprint arXiv:2308.13789* (2023).
- [8] <https://www.comsoc.org/about/committees/emerging-technologies-initiatives/large-generative-ai-models-telecom-genainet>



Performance Requirements and Evaluation Methodology for AI and Communication in 6G

Gongzheng Zhang ¹, Jian Wang ¹, Rong Li ², Yan Chen ³, Jiafeng Shao ³, Hui Lin ³, Jun Wang ¹, Jianglei Ma ⁴, Peiying Zhu ⁴

¹ Wireless Technology Lab

² Advanced Wireless Technology Lab

³ Research Dept, WN

⁴ Ottawa Advanced Wireless Technology Lab

Abstract

The International Telecommunication Union – Radiocommunication Sector (ITU-R) has defined "AI and Communication" as one of the six usage scenarios for next-generation mobile communication systems. As a new scenario involving new capabilities, its key performance indicators (KPIs) and minimum requirements are yet to be defined. This paper starts by describing the "AI and Communication" scenario and the typical artificial intelligence (AI) services provided in the scenario in next-generation mobile communication systems. Then, it introduces the general principles for performance definition, and proposes detailed performance indicators and requirements. It also provides an evaluation methodology for the proposed performance indicators, along with an example of evaluation procedures.

1 Introduction

With the rapid development of artificial intelligence (AI) technologies, especially deep learning and large pre-trained models, AI will become an essential part of almost all systems in industries and daily life. Mobile communication systems, which are already deployed on a massive scale, could be the best choice as a unified infrastructure that integrates communication and AI and that delivers ubiquitous AI services to all connected people and machines. This will drive the revolution of mobile communication systems.

To promote the development of next-generation mobile communication systems, the International Telecommunication Union – Radiocommunication Sector (ITU-R) has identified six typical usage scenarios for IMT-2030 and beyond. In addition to the three usage scenarios enhanced from IMT-2020, AI and sensing are newly included as two beyond-communication services that are expected to be provided by 6G networks. This will boost new capabilities and corresponding performance indicators. As such, it is necessary to study the corresponding technologies, performance requirements, and evaluation methodology. While most recent works focus on the technologies, further study is needed for the performance requirements and evaluation methodology.

This paper aims to provide guidelines for designing next-generation mobile communication systems and ensure users receive guaranteed AI services. To achieve this, the paper will introduce the performance requirements and evaluation methodology for AI and communication in 6G. Specifically, this paper will first describe the "AI and Communication" scenario defined in the IMT-2030 framework, with a particular focus on the typical AI services and capability requirements of this new usage scenario for 6G. Next, the paper summarizes the current status of the performance indicators, and then introduces the design principles and the proposed qualitative and quantitative performance requirement definitions. Finally, the paper provides the corresponding evaluation methodology and an example, followed by the conclusions.

2 AI and Communication

6G, a next-generation mobile communication system, aims to make intelligence inclusive by providing artificial intelligence as a service (AlaaS). This will enable easy training, fast distribution, and accurate inference of large

AI models distributed in wireless networks. And by utilizing the data and resources of distributed intelligent terminals, 6G will be able to provide AI model training services, made possible through local training at distributed terminals and model interaction between them over the network — this can also effectively protect users' data privacy. Furthermore, 6G can provide high-accuracy inference services for resource-constrained terminals by joint scheduling of communication and AI resources. This will drive AlaaS to become a typical application scenario of 6G. This section will describe the corresponding standardization progress in ITU and typical services.

2.1 AI and Communication in the IMT-2030 Framework

To facilitate the development of IMT-2030 and beyond, the Working Party (WP) 5D in the ITU-R has approved a new framework and overall objectives [1]. This includes identifying motivations, applications, technology trends, spectrum, usage scenarios, and capabilities for next-generation mobile systems. A key application trend and enabling technology is ubiquitous intelligence — while AI can enhance the performance of wireless interfaces and enable automation of wireless networks and intelligent network services, one of the key objectives of the IMT system design is to efficiently support AI services within wireless networks.

Among the six usage scenarios identified by the ITU-R, as shown in Figure 1, "AI and Communication" is one that IMT-2030 specifies as providing beyond-communication services. This usage scenario will support distributed computing and AI applications, including data collection, local and

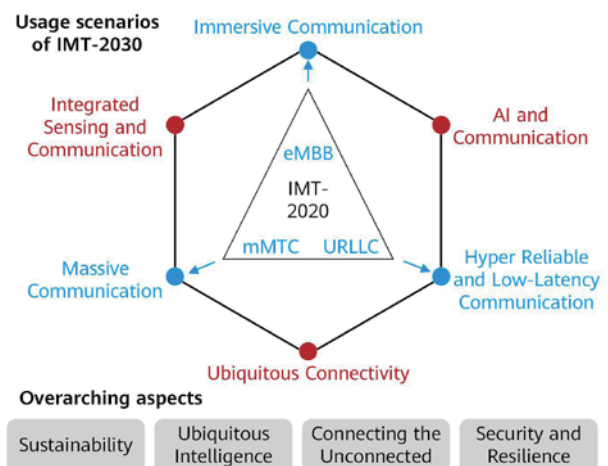


Figure 1 Usage scenarios of IMT-2030 [1]

Table 1 Capabilities of IMT-2030 [1]

Enhanced Capabilities	IMT-2020	IMT-2030
Peak data rate (Gbps)	20/10 for DL/UL	e.g., 50, 100, 200
User experienced data rate (Mbps)	100/50 for DL/UL	e.g., 300, 500
Spectrum efficiency (bps/Hz)	(Peak) 30/15 for DL/UL	e.g., x1.5, x3
Area traffic capacity (Mbps/m ²)	10	e.g., 30, 50
Connection density (devices/km ²)	10 ⁶	10 ⁶ –10 ⁸
Mobility (km/h)	500	500–1000
Latency (ms)	1	0.1–1
Reliability	1 – 10 ⁻⁵	1 – 10 ⁻⁵ to 1 – 10 ⁻⁷
New Capabilities of IMT-2030		Value
Coverage		TBD
Sensing-related capabilities		TBD
AI-related capabilities		TBD
Sustainability		TBD
Positioning (cm)		1–10

distributed computation offloading, and distributed AI model training and inference. Typical use cases include IMT-2030 assisted automated driving, autonomous collaboration between devices for medical assistance applications, offloading of heavy computation operations across devices and networks, and the creation of and prediction with digital twins.

To support the new usage scenarios, IMT-2030 should include AI- and sensing-related capabilities in addition to traditional communication capabilities, as listed in Table 1. The "AI and Communication" usage scenario would require high area traffic capacity and user experienced data rates, as well as low latency and high reliability, depending on the specific use case. In addition to the communication aspects, this usage scenario is expected to include a set of new capabilities related to the integration of AI functionalities into IMT-2030. Such capabilities include data acquisition, preparation and processing from different sources, distributed AI model training, model sharing and distributed inference across IMT systems, and computing resource orchestration and chaining. The following subsections will describe the AI capabilities and performance requirements based on the typical AI services.

2.2 Typical Services in the "AI and Communication" Scenario

IMT-2030 will efficiently support AI applications in an end-to-end manner, connecting distributed intelligence to provide ubiquitous AI services (e.g., AI model training, inference, deployment, and more). To achieve this goal, IMT-

2030 can build a distributed and efficient AI service platform by utilizing the connection, data, and model resources and capabilities within the network. AI applications include providing intelligent capabilities for network optimization, i.e., using end-to-end AI algorithms in customized optimization and automated operation and maintenance (O&M) for wireless interfaces and networks. Such applications also include providing intelligent capabilities to the users, i.e., serving as a distributed learning infrastructure and moving the centralized intelligence on the cloud to deep edge ubiquitous intelligence through the network's native integration of communication and AI capabilities.

2.2.1 An Exemplary AI Application Served by IMT-2030

Collaborative robots are widely recognized as a future 6G application scenario that requires AI services with low latency and high learning and inference accuracy. In this exemplary use case, multiple robots work together to accomplish complex tasks in an industrial environment. Each robot is equipped with cameras and other sensors and powered by AI capabilities to achieve partial autonomy. For full autonomy and complex tasks, the collaborative robot system should sense, perceive, plan, and control towards the ultimate goal of the task. For example, when someone issues a new voice request for some goods, the command in natural language should first be understood, and the subtasks for each robot should be planned. Both aspects need to be achieved through efficiently trained large (language) models that require huge computing and

memory resources. And through local vision or control models, the robots will be able to detect objects from the sensed images and plan the path trajectory with corresponding control decisions for the subtasks. This makes it possible for the AI-enabled robots to collaborate with the network in order to utilize its connected super AI capabilities for planning in complex tasks. These robots can also cooperate with each other over the network to improve the performance of local models via collaborative training, sharing and learning from the experience of each other. Two typical services in this exemplary use case are model inference and training, which are described in following sections.

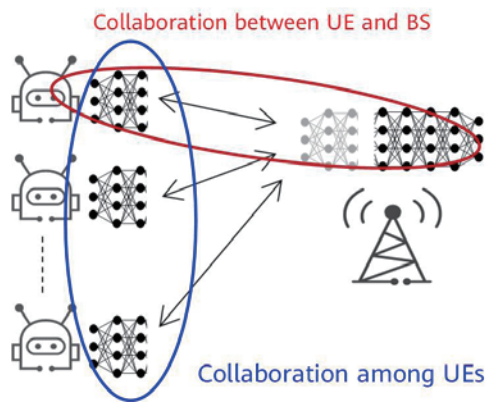


Figure 2 AI applications for collaborative robots

2.2.2 Model Inference Service

AI model inference is a fundamental function for AI applications. It takes inputs, runs the AI models, and produces the expected outputs. Through ubiquitous connectivity, the 6G network with native intelligence could provide real-time model inference capabilities that meet different requirements. In the distributed AI model inference service, the 6G network jointly utilizes the communication and AI capabilities to provide high-accuracy model inference services in real time for users with limited capabilities through model collaboration. Figure 3 illustrates a typical AI model inference service. In this service, a large model may be split into two parts, which are deployed on the network and user sides and work together. The part with

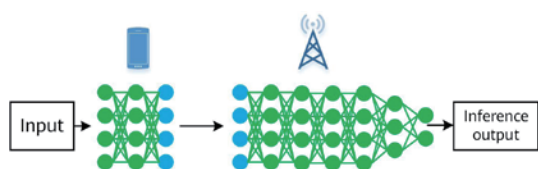


Figure 3 AI model inference service

high resource requirements is deployed on the network side, enabling powerful network AI capabilities to provide the joint model inference service for end users.

2.2.3 Model Training Service

AI model training is key for obtaining a model with high accuracy. The 6G network with native intelligence could provide suitable algorithms and resources for model training orchestration based on different user and network characteristics, thereby improving the speed and accuracy of model training. In the large-scale distributed AI model training service, the network serves as a management platform to provide high-speed data channels and efficient scheduling mechanisms for exchanging data or model parameters between distributed terminals. This supports fast model aggregation and distribution while also ensuring user privacy protection. Figure 4 illustrates a typical distributed training service. In each round, the distributed terminals use local data to train models locally and upload the updated models to the network. The network aggregates these updated local models to obtain a global model, which it then distributes to the terminals. The aggregation and distribution procedures are iterated, enabling joint learning while also protecting users' raw data.

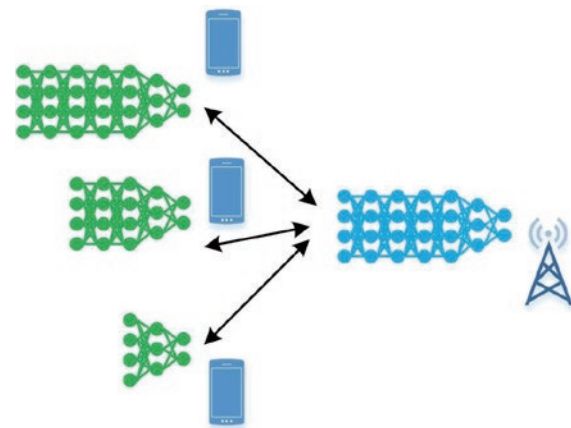


Figure 4 Distributed AI model training service

3 Performance Requirements for the "AI and Communication" Scenario

System design is driven primarily by performance requirements, which evolve or revolutionize each generation of mobile communication systems. Existing mobile communication systems are mainly designed

for connection-oriented data transmissions — the key performance indicators (KPIs) of such networks mainly include the transmission rate and latency of connections. However, AI services not only involve transmissions, but also include AI-related resources, meaning that AI model learning/inference accuracy and latency are the KPIs. From a communication perspective, the 6G network should provide high traffic capacity, especially in the uplink direction, in order to meet the requirements of data or model exchange in model training and inference. And from an AI perspective, the 6G network should support large-scale distributed learning and real-time inference. This is why the design of the 6G network should take into account both AI and communication in an integrated manner from the beginning. The following subsections will describe the current status and then detail the principles and architectures for performance requirement definitions in the "AI and Communication" usage scenario.

3.1 Current Status

Previous mobile communication systems, from 2G to 5G, have mainly provided communication services, with data transmission being almost the only objective. Support for AI/machine learning (ML) operations has been studied in Release 18 of 5G. Three typical AI/ML operations have been identified and reported in 3GPP TR 22.874 [2], namely, split inference, model/data distribution, and distributed/federated learning. Various applications have also been defined, for example, image recognition, real-time media editing, split inference and control among robots, and collaborative learning among multiple agents. However, all the AI/ML operations are expected to be executed in cloud servers, and the 5G system still provides only communication services to transmit the data between users and cloud servers, potentially leading to higher data rate requirements.

For the next-generation mobile communication systems that introduce new capabilities beyond communication (e.g., AI-related capabilities), supporting AI services is commonly considered in 6G-related research groups. In their published white papers [3, 4], the China IMT-2030 Promotion Group and Hexa-X, two main 6G research organizations in China and Europe respectively, both identify AI services provided by 6G as a key factor in the design of next-generation networks. They also suggest including new capabilities to identify the performance requirements of AI services. For both AI-enabled air interfaces and AI services, the two organizations propose AI-related performance indicators

in addition to the traditional communication performance indicators, including AI model inference accuracy and latency. However, the performance indicators are not illustrated clearly, and no details are defined for the requirements and evaluation methodology in 6G.

The computer science community has defined some training and inference KPIs to evaluate the capabilities of the AI hardware and software systems. For example, the MLPerf benchmark [5] defined by MLCommons builds a method to measure the AI performance for both model training and inference via reference applications, models, and datasets. However, because these KPIs are used to measure the hardware or software capabilities in a centralized way, they cannot be used to measure the capabilities of distributed AI services in 6G networks that involve communication.

3.2 Principles for Performance Definition for AI and Communication in 6G

Leveraging resources such as connections, models, and data within 6G networks, 6G AlaaS provides AI capabilities that adapt to different application scenarios. Unlike conventional mobile networks, 6G networks need resources beyond only connections in order to provide high-performance AI services for users. Accordingly, 6G AlaaS needs to consider integrating communications capabilities and AI capabilities in order to build comprehensive performance indicators and evaluation methods oriented to AI services. This will provide guidance for the 6G network design and network resource configuration.

The main principles of performance definition for AI-related capabilities are as follows.

- **End-to-end AI capabilities.** AI services should use end-to-end performance as indicators in order to guarantee user-experienced service quality. The AI service quality depends on both communication and AI capabilities. However, the existing performance indicators and evaluation methods only focus on the communication capabilities and therefore cannot guarantee the AI service quality. This is why the IMT-2030 system needs to consider how to integrate communication capabilities with AI capabilities.
- **Typical services.** The IMT-2030 system is the key to realizing ubiquitous intelligence. By utilizing the AI capabilities within the network, this system should

provide a platform for large-scale distributed model training and unified high-accuracy model inference to diverse users via collaboration. Accordingly, AI-related capability indicators need to be defined based on typical services, such as training and inference.

- **Core performance.** The goal of AI and communication integration is to enable AI services efficiently, including model training and real-time high-accuracy model inference. To ensure that AIaaS is acceptable to billions of users, it is crucial to focus on the key factors that impact user experience. Amidst the plethora of performance indicators available for AI services, the IMT-2030 system must prioritize the core indicators to accomplish this goal.

3.3 Proposed Performance Requirements for AI and Communication in 6G

The KPIs for AI and communication are defined from the perspective of services (including AI model training and inference) provided by 6G networks. The performance of such services depends on the AI model capabilities provided by the system's AI resources and the communication capabilities connecting users and networks. The proposed AI service performance requirements include a group of functionality requirements and three quantitative requirements, described below. The functionality requirements can be evaluated via inspection, and the quantitative requirements need to be evaluated via simulation.

- **AI service functionality requirements**

The functionality requirements for AI-related capabilities are that the candidate radio interface technologies (RITs) or sets of radio interface technologies (SRITs) shall have mechanisms and/or signaling related to the functionalities (e.g., distributed data processing, distributed learning, AI computing, AI model execution, and AI model inference) that are exposed as capabilities to external applications, or any other functionalities that the proponent(s) of candidate RITs/SRITs consider relevant to better support AI-enabled applications.

- **AI service accuracy (or AI service quality)**

AI service accuracy is defined as the accuracy of the AI inference/learning service. Specifically, it is the degree to which the outputs from the AI service are the same as the true values for the given inputs within the given service latency requirements, or relative to the reference

case. For a given AI task, the AI service accuracy depends on the task characteristics, AI model deployment method, and AI-related data transmissions. Different applications may have different requirements on AI service accuracy. For example, the accuracy requirement of object recognition in autonomous driving applications is much higher than that of ordinary consumers who want to identify flowers. Consequently, the minimum performance requirements for the 6G network can be defined for specific applications with certain accuracy requirements. If the requirements can be met, all applications with lower requirements can be supported by the network. The minimum requirement for AI service accuracy in 6G can therefore be defined as higher than a certain accuracy within a given time for the deployment environment, assuming a given AI inference/learning task.

- **AI service latency**

AI service latency is defined as the time taken from the start to the end of the AI inference/learning service. It is the sum of the communication time for AI-related data transmissions and the processing time of the AI model, where the processing time depends on the devices and implementations. Similar to AI service accuracy, different applications may have different requirements on AI service latency. As such, the minimum performance requirements for the 6G network can be defined for specific applications with certain latency requirements. If the requirements can be met, all applications with lower requirements can be supported by the network. The minimum requirement for AI service latency in 6G can therefore be defined as less than a certain time with a given inference accuracy for the deployment environment, assuming a given AI inference/learning task.

- **AI service density**

AI service density is defined as the number of AI services that meet given AI service accuracy and AI service latency requirements supported by the network simultaneously per unit area. It is a system capacity indicator of the IMT-2030 system. For different application requirements (i.e., accuracy or latency), the system can support different AI service densities. This means that the minimum performance requirements for the 6G network can be defined for specific applications or combinations of applications with certain accuracy and latency requirements. The minimum requirement for AI service density in the 6G network can therefore be defined as the number of services per km² for the deployment environment, assuming a given AI inference/learning task.

4 Evaluation Methodology and Example

The previous section defines quantitative performance requirements for the typical distributed AI model training and inference services. Service performance is determined by both communication and AI resources and should therefore be evaluated with certain communication and AI assumptions. This section will describe the evaluation methodology first and then present an example with detailed assumptions and results.

4.1 Evaluation Methodology

The performance requirements can be derived from two essential KPIs, namely, AI service accuracy and latency. AI service accuracy is defined as the degree to which the outputs from the AI service are the same as the true values for the given inputs, which depends on both the AI model and AI-related data or model transmissions. AI service latency is defined as the sum of AI model processing time and data transmission time, which also depends on both the AI model and AI-related data or model transmissions. These two definitions hold true for both model training and inference services, sharing similar radio resources but with model exchange and data exchange respectively.

The performance evaluation can follow the service procedures. Figure 5 shows the AI service performance evaluation system. The performance evaluation includes the following key components:

- **Resource assumptions:** The evaluation should be done in a test environment similar to the definition in communication performance evaluations [6]. Within the test environment, the radio configurations should be

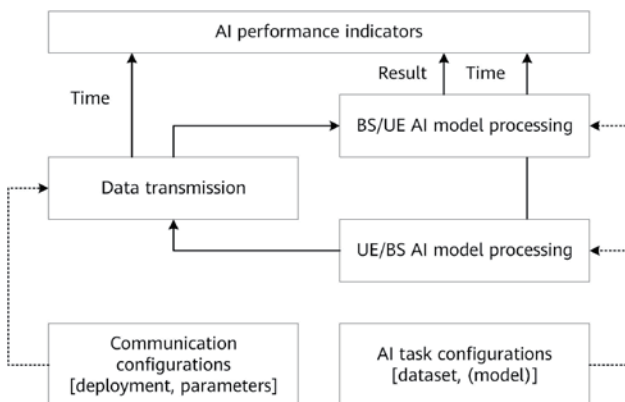


Figure 5 AI service performance evaluation system

provided, including the bandwidth, number of antennas at the user equipment (UE) and base station (BS), and so on. This is necessary to reflect the radio resources. In addition, AI tasks should be defined with AI-related configurations, including datasets consisting of inputs and corresponding target outputs with accuracy calculation methods.

- **AI service procedures:** The entire procedures can start from AI model processing at the UE where the intermediate data (model output or model weights) is generated. Then, this data is transmitted from the UE to the BS under the assumed radio configurations. Next, the BS receives the intermediate data and uses the AI model to process it in order to get the service results, which are then used to calculate performance indicators.
- **AI service performance calculation:** The AI service accuracy and latency are calculated based on the service results, AI model processing time, and transmission time. As mentioned earlier, the AI service accuracy is calculated as the degree to which the output of the AI model processing is the same as the target value for each input in the dataset. The degree is defined according to the AI task. The AI service latency is the sum of the AI model processing time at the UE and BS and the transmission time of intermediate data.

For AI service density evaluation, AI service density is defined as the number of AI services that meet given AI service accuracy and latency requirements. This can be evaluated through AI service accuracy and latency simulation. For example, we can first set the number of served UEs N to a minimum value, and generate service requests from the UEs. Then, we use the evaluation parameters of the test environment to perform system simulation and collect statistics on the AI service accuracy within the service latency. We can gradually increase N and repeat the simulation until the AI service accuracy falls below requirements, with the value of N to be N_{max} . The AI service density is calculated as $C = N_{max}/\text{Coverage area}$.

4.2 Evaluation Example

In this subsection, we use the distributed AI inference service as an example to illustrate the performance evaluation methodology presented earlier. This methodology can also be used for collaborative training and inference services after the procedures are modified according to the corresponding service procedures (this is left for future work). In a typical future smart factory, AI-enabled robots

need to perceive the environment (to detect objects in real time, for example) through cameras. The images these robots collect can be further used to achieve real-time high-accuracy AI model inference.

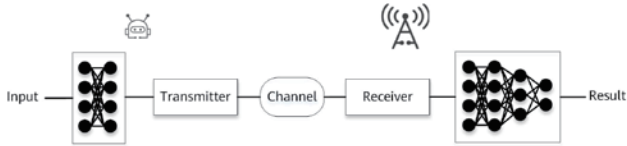


Figure 6 Distributed AI inference service example

Figure 6 illustrates the procedure involved in the AI inference service for a user. It includes not only the AI model processing on the user and network sides, but also the transmission between the user and the network. To be specific, the AI inference service consists of three steps: 1) the UE uses the UE-side AI model to process the input data in order to obtain intermediate data; 2) the UE transmits this data to the BS; 3) the BS uses the BS-side AI model to process the received intermediate data and obtain the inference results. Note that this example procedure represents a service starting from the UE with the input data and uploading the intermediate data to the BS to get the results. A similar procedure can also be applied in the downlink direction, where the BS first processes the input data and transmits the intermediate data to the UE to get the results. The following evaluation methodology can also be applied for this downlink case.

- **Evaluation configurations:** The evaluation configurations are defined as follows, with examples given in brackets.
 - Test environment: [Dense Urban]
 - Radio configurations: [same as immersive communication (user experienced data rate: 500 Mbps)]
 - AI task: [image recognition]
 - AI dataset: [ImageNet-1k validation dataset [7]]
 - AI model: [AlexNet [8], the left part is processed by the UE, and the right part is processed by the BS, as shown in Figure 7]
 - AI model processing time: [UE: 0.75 ms; BS: 0.45 ms]
- **Evaluation procedures**
 - AI service accuracy: AI service accuracy can be evaluated by simulation. The UE processes each

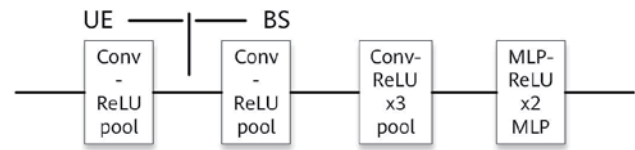


Figure 7 AlexNet model deployment example

input of sample $S_i, i = 1, \dots, n$ in the dataset based on the UE-side AI model, and obtains the intermediate data Z_i . According to the test environment and transmission configurations, the UE sends the intermediate data and the BS receives it. Taking a classical transmission scheme as an example, the intermediate data is first quantized and represented as bits, which are then encoded and modulated to symbols for wireless transmission. The BS processes the received intermediate data \tilde{z}_i based on the AI model on the BS side, and obtains the inference result \tilde{Y}_i corresponding to each sample. We can then compare or calculate the inference results with the target output or label Y_i of each sample in order to obtain the degree to which the output is the same as the true value $\text{acc} = \frac{1}{n} \sum_{i=1}^n 1_{\{\tilde{Y}_i = Y_i\}}$, that is, the AI service accuracy.

For the accuracy of the reference case, we can process each sample s_i in the dataset based on the whole AI model in order to obtain the inference result \tilde{Y}'_i . We can then compare the inference result with the label Y_i of each sample to obtain the output of the reference case. The degree to which the output is the same as the true value of reference case is $\text{acc}_{\text{ref}} = \frac{1}{n} \sum_{i=1}^n 1_{\{\tilde{Y}'_i = Y_i\}}$. The relative AI service accuracy is calculated as $\text{acc}/\text{acc}_{\text{ref}}$.

- AI service latency: The AI service latency is the sum of the time used for intermediate data transmission, t_{comm} , and the UE- and BS-side AI model processing time, $t_{\text{proc,UE}}, t_{\text{proc,BS}}$. Therefore, the AI service latency is given by $t_{\text{service}} = t_{\text{comm}} + t_{\text{proc,UE}} + t_{\text{proc,BS}}$. In this example, we use the time calculated as the number of payload bits divided by the data rate as the data transmission time. The number of payload bits is determined by the number of elements in the intermediate data and the number of quantized bits per element. Other schemes can also be used taking new technologies into consideration.

Table 2 AI service performance evaluation results

Number of bits per element	2	4	6	8	10	12	16	32
AI service accuracy (%)	0.14	10.35	52.94	56.47	56.53	56.55	56.55	56.56
Relative AI service accuracy (%)	0.24	18.30	93.61	99.84	99.95	99.99	99.99	100
AI service latency (ms)	1.4	1.6	1.8	1.9	2.1	2.3	2.7	4.2

• Evaluation results

The AI service accuracy and latency under different transmission setups (i.e., number of quantized bits per element) are provided in Table 2. As can be seen from the table, there is a trade-off between AI service latency and AI service accuracy due to the intermediate data transmission. If a minimum AI service accuracy (e.g., 56%) and maximum AI service latency (e.g., 2 ms) are required, we need to optimize the transmission configurations (8 bits per element in the table for this example) or improve the transmission technology to meet both requirements.

5 Conclusions

In this paper, we have illustrated the motivations, typical AI services, and performance requirements of the "AI and Communication" usage scenario — a new scenario defined in IMT-2030 for 6G. To provide guidelines for the system design and better support AI services, we proposed new performance indicators that integrate AI and communication capabilities and resources in the network, from both the user experience and network capacity perspectives. We also provided the corresponding evaluation methodology with a detailed example. This is a first step towards 6G moving from vision to technical designs.

References

- [1] ITU-R, Recommendation ITU-R M.2160-0, "Framework and overall objectives of the future development of IMT for 2030 and beyond," Nov. 2023.
- [2] 3GPP TR 22.874, "Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," Release 18, 2021.
- [3] IMT-2030 (6G) Promotion Group, "White paper on typical usage scenarios and key capabilities in 6G," July 2022.
- [4] Hexa-X Deliverable D1.3, "Targets and requirements for 6G – Initial E2E architecture," Feb. 2022.
- [5] <https://mlcommons.org/en/>, accessed on Aug. 10, 2022.
- [6] ITU-R M.2412-0, "Guidelines for evaluation of radio interface technologies for IMT-2020," 2017.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," IEEE CVPR, 2009.
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS, 2012.



Data Plane Design for AI-Native 6G Networks

Xueqiang Yan ¹, Xinran Zhang ², Junfan Wang ¹, Yi Zhang ³

¹ Wireless Technology Lab, Huawei

² Beijing University of Posts and Telecommunications

³ Zhejiang University

Abstract

The sixth generation (6G) networks are set to revolutionize connectivity and enable groundbreaking intelligent applications by integrating artificial intelligence (AI) and radio frequency (RF) sensing capabilities. With RF sensing, 6G networks will collect massive volumes of data from the physical world, accelerating AI-driven innovations across multiple domains. To manage the massive data generated by integrated sensing and communication (ISAC) and network AI, we introduce the concept of data as a service (DaaS) and propose the inclusion of a dedicated data plane (DP) within the 6G architecture for managing data flows across distributed sensing and AI applications. The DP comprises two key components: data orchestration (DO) and data agent (DA). The DO translates service requirements into network configurations, while the DAs handle data collection, processing, storage, and sharing. Additionally, we propose a stateless, distributed data communication proxy (DCP), which decouples data producers from data consumers through asynchronous communication, specifically by using the publish/subscribe (Pub/Sub) model. The DCP brings several key advantages to the DP, including ultra-low latency in data forwarding, on-path data processing, and a flexible data processing topology.

Keywords

DP, DO, data service, DCP, AI, ISAC

1 Introduction

Data plays a pivotal role in the deployment of artificial intelligence (AI) in 6G networks, necessitating massive amounts of high-quality training and test datasets. Incomplete or erroneous datasets can result in unreliable models, which will generate suboptimal output. However, the process of massive data collection is often labor-intensive and time-consuming. These challenges critically hinder access to sufficient quantities of high-quality datasets for AI applications.

Unprecedented volumes of data will be generated, transmitted, processed, and stored by an enormous number of data-driven applications that leverage the radio frequency (RF) sensing capability inherent in 6G networks. By leveraging this capability, a 6G network can effectively transform itself into a "sensor" that reads data from the physical world. Consequently, 6G sensing data will emerge as a vital data source for training AI models, driving the development of generative and interactive AI services. Specifically, if 1% of the capacity of 6G networks is allocated for sensing, the 6G sensing data volume is expected to reach the quettabyte (10^{30}) level per day for on-device AI models and the zettabyte (10^{21}) level per day for cloud-based AI models at base stations (BSs). If this data is efficiently used for AI model training, the models can better adapt to real-world scenarios and deliver customized AI services.

As a fundamental component of 6G, network AI will employ a deep-edge architecture to facilitate extensive machine learning (ML) in a distributed and collaborative manner. Its primary objective is to intelligently connect a multitude of distributed intelligent agents for large-scale deployment of AI. This requires efficient, high-capacity, and low-latency transmission of massive amounts of data, including model parameters across a vast array of intelligent agents, which are crucial elements in the data as a service (DaaS) ecosystem. From the DaaS perspective, the network must provide flexible support for machine learning operations (MLOps), that is, to automate and streamline ML workflows and deployments.

1.1 Current Network Architecture for Connectivity Services

In today's typical mobile network architecture, the control plane (CP) and user plane (UP) play a key role in providing wireless access services for mobile users. The CP handles tasks such as establishing connections and controlling the

forwarding of UP data, commonly referred to as session management within the 5G core. The UP is designed to forward network packets to their intended destinations efficiently. The following sections describe the CP and UP in more details.

1.1.1 Control Plane (CP): Service-based Interface

Figure 1 illustrates the service-based architecture (SBA) defined by the 3rd Generation Partnership Project (3GPP). In the SBA, the CP functions and common data repositories of a 5G network are distributed across a set of interconnected network functions (NFs). Each NF is granted permission to access the services provided by other NFs [1]. The SBA has a service-based interface (SBI) between the interconnected NFs on the 5G core (5GC) CP. The SBI acts as a communication bus, using the HTTP/2 protocol for communication across all the CP functions.

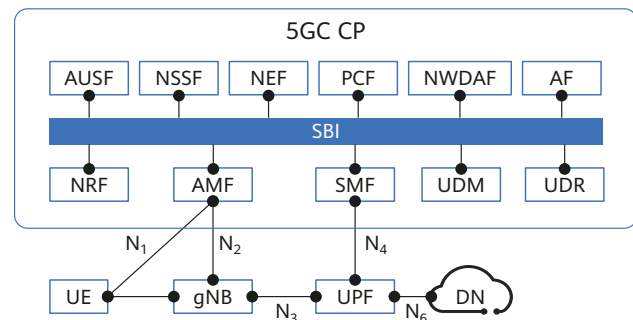


Figure 1 SBA of 5GC

In 5G networks, HTTP/2 connections are peer-to-peer connections, with messages flowing in both directions between NFs. An NF establishes a connection to another NF before sending request messages to it, and the receiving NF responds via the same connection. If the receiving NF needs to send messages to the originating NF, it sets up another connection in the opposite direction. These signaling messages contain small data packets of predefined sizes, and therefore, the overall traffic volume scales proportionally with the number of active user equipments (UEs).

1.1.2 User Plane (UP): GTP-U Tunnel

As illustrated in Figure 2, data traffic on the UP is carried by protocol data unit (PDU) sessions established between

the UEs and the user plane function (UPF) deployed in the core network (CN) across the radio access network (RAN). To guarantee data transmission quality, data packets are first mapped to appropriate quality of service (QoS) flows and then transmitted through a specific data radio bearer (DRB) between the UE and the RAN node and then through an N3 GTP-U tunnel between the RAN node and the UPF in the CN. This paradigm can be considered session-oriented because data flows in dedicated tunnels.

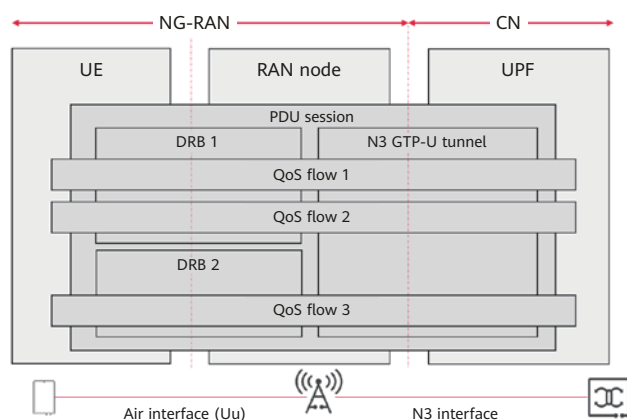


Figure 2 5G PDU session

Both the CP and UP operate within a communication session-centric framework. The signaling procedures and user data packets are inherently associated with the UEs. Specifically, the CP sets up UP data tunnels for each session, facilitating signaling message exchange among various entities, including UEs, BSs, and CN functions, to maintain state synchronization.

1.2 Challenges

In addition to its primary function of providing connectivity services for UEs, 6G will expand into the realm of AI and sensing services, collectively referred to as "beyond-connectivity services." These services differ from traditional connectivity services, necessitating a shift from a connectivity-centric network design to a data-centric one. Consequently, the two planes of CP and UP are no longer applicable due to the following reasons.

- **Massive data volume:** 6G networks are expected to handle unprecedented amounts of data, a significant part of which will be generated by "beyond-connectivity services." The actual data volume in 6G networks is expected to significantly surpass current ITU-R forecasts, which are only based on mobile subscribers'

data consumption [2]. With 6G's ubiquitous sensing capabilities, massive amounts of data will be generated and fed into algorithms designed to construct digital twins of the physical world. At the same time, with intelligence enabled, the distributed UEs and BSs, the CN, and the cloud will generate "AI data," including gradients, parameters, models, and tokens, which constitute yet another massive volume of non-connectivity data. Such enormous amounts of data cannot be handled by the CP (e.g., the SBI), which features reliable transmission of small data packets and short-term connections.

- **Complex data topologies:** Emerging 6G services involve complex data topologies across various data producers and consumers. The RAN nodes, UEs, and NFs are not connected in a simple and linear point-to-point manner. Unlike connectivity services, which are delivered only to UEs, "beyond-connectivity services" can reach any authenticated and authorized subscriber with network access. The current UP (e.g., the PDU session) establishes only one-to-one communication tunnels between the UEs and the UPF. This connection-oriented paradigm is incapable of handling the complex data topologies in 6G "beyond-connectivity services."
- **Data orchestration requirement:** Different from wireless access, the AI or sensing services run on demand, meaning that different AI or sensing services require the collaboration of different UEs, BSs, and NFs for service provisioning. From the service orchestration viewpoint, multiple UEs, BSs, and NFs with sensing capability/computing power need to be orchestrated to deliver a service such as distributed learning. This requires automatic translation of service requirements into network configurations, which cannot be achieved with the current network structure.

With the advent of 5G and beyond, the ecosystem of connected intelligent devices is expanding rapidly, resulting in a significant increase in the amounts of data requiring real-time processing. Conventional session-centric approaches, which heavily rely on centralized brokers, are insufficient to meet the communication flow management requirements of large-scale distributed AI systems and collaborative sensing networks. Additionally, the conventional approaches face challenges in scaling up and meeting the latency and privacy requirements of the device-edge-cloud computing continuum. Thus, there is a pressing need for an innovative network architecture tailored for data management and processing in AI-capable 6G networks.

2 Data Plane (DP) for DaaS

Inspired by the decoupling of the CP and design of the UP in the CN, we propose a dedicated data plane (DP) in the 6G network architecture to improve the overall network flexibility and efficiency. The DP is capable of facilitating data flow on any topology, enabling flexible data services and efficient on-path data processing. Specifically, it breaks down service requirements into distinct data processing functions, including data collection, pre-processing, and analytics. It coordinates various network components involved in meeting service requirements [3]. The primary objective of the DP, which transmits non-connectivity data, is to streamline management and control. The non-connectivity data will be fed into AI algorithms for iterative optimization. The DP serves as the foundation for providing data services for both internal and external applications and services. Within the DP, two key components are defined: DO and DA, which are elaborated in the following sections.

2.1 Data Orchestration (DO)

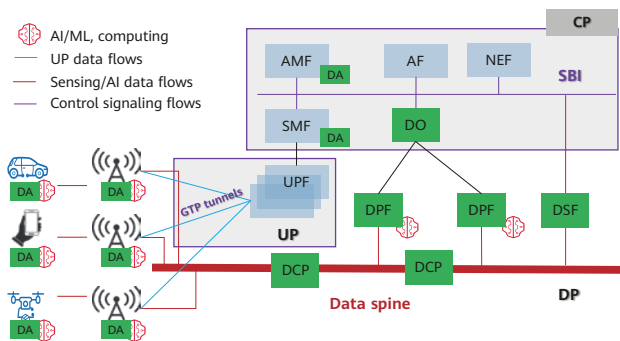


Figure 3 Architecture of the 6G DP

In contrast to connectivity services, for which the network establishes tunnels from the UEs to the UPF, non-connectivity services require the involvement of multiple entities in service provisioning and delivery. These entities must collaborate effectively through proper orchestration. As illustrated in Figure 3, the DO serves as a conductor of the DP, overseeing data flows across network elements. By analyzing service requirements and translating them into network configurations, the DO ensures precise data flow within the DP — the data is delivered to the right processing units and applications at the right time.

The DO includes four essential components: DA registration, data flow engine, data flow management, and policy control. DA registration collects the capabilities of all DAs under the DO's management and uses the collected information to ensure efficient resource utilization. The data flow engine determines the sequence of actions and data transformations that need to be performed (e.g., filtering, aggregation, and forwarding) based on the service requirements and, by doing so, ensures that the data flows effectively across the system. Data flow management is responsible for overseeing the actual routing of data packets across the orchestrated data topology, considering factors such as network congestion, latency requirement, and resource availability. Policy control decides on policies regarding data security, access control, and privacy, ensuring that sensitive data is handled appropriately.

2.2 Data Agent (DA)

The DAs act as intelligent intermediaries within the DP throughout the data service lifecycle. They perform a range of functions, including data preprocessing, data transformation, local decision-making, data caching, and security enforcement. Specifically, DAs filter, aggregate, and compress raw data — for example, the in-phase and quadrature (I/Q) signals of ISAC data — to reduce network traffic and processing overhead. DAs convert sensing data into formats suitable for network AI algorithms or applications to facilitate data transmission. In distributed data processing scenarios, DAs perform basic analysis and make decisions based on locally collected data, thus eliminating the need for constant communication with centralized entities. Additionally, DAs support data caching to facilitate frequent data access and intermediate processing results, resulting in faster retrieval and enhanced efficiency. DAs also strengthen data security by implementing access control mechanisms and encryption protocols.

DAs can be deployed on various network nodes, such as UEs, BSs, network edge nodes, and CN nodes. DAs deployed on UEs or BSs perform preliminary data acquisition and local processing, including filtering and compression, before transmission. Deployed on edge computing nodes, they further process and analyze data, and potentially trigger real-time actions, feed the outcome to downstream DAs equipped with AI algorithms, or perform other actions. Furthermore, centralized DAs located in the CN can

potentially handle more complex tasks like data aggregation from multiple data sources or feed data to centralized AI algorithms.

Based on the implementation mode, DAs can be classified into two types: embedded DAs and standalone DAs. Embedded DAs function as part of a network entity and are generally responsible for data collection and preprocessing. Standalone DAs play dedicated roles assigned by the DO. The data processing function (DPF) and data storage function (DSF) are two examples of standalone DAs that can be deployed in the RAN or CN. The DPF is responsible for processing data, including ISAC and AI data, while the DSF acts as a repository for long-term data storage.

2.3 Data Communication Proxy (DCP)

The implementation of DOs and DAs presents another challenge — the traditional data traffic pattern cannot meet the requirements of new data topologies. Particularly, a typical "beyond-connectivity service" powered by either network AI or wireless sensing involves a data topology composed of multiple data sources, data processing functions, and data sinks. Network AI and wireless sensing are transforming the mobile network into a data-intensive computing and big data analytics platform. However, traditional session-centric data communication methods for data transmission are insufficient to handle the data traffic generated by the "multiple-input-multiple-output" pattern. To address this issue, we propose a new logical function, namely the DCP, to handle the massive sensing and AI data featuring complex topologies and on-path data processing in 6G networks. As depicted in Figure 4, this function can provide efficient data distribution, overcoming the limitations of the traditional "connect-then-communicate" model.

In the current network architecture, data is ingested into the NFs via the UPF. In contrast, in our proposed architecture, the DCP serves as an intermediary data collector. It receives data and then efficiently distributes it to the corresponding NFs. By decoupling data producers from data consumers, the DCP enables asynchronous data exchange, allowing data producers to continue operating without waiting for responses from the consumers. The DCP can be implemented as a distributed system that comprises a message broker and multiple message queues. The message broker acts as an intermediary between the data sender and the receiver. It performs topic-based message routing from a publisher to a subscriber. The message queue is a data structure or container that facilitates communication between applications by sending, receiving, and storing messages. The Publish/Subscribe (Pub/Sub) communication pattern is used as the asynchronous messaging architecture, enabling messages to flow between entities without the need for the sender and the recipient to know each other's identity. The data producers publish messages of specific topics (that is, data categories into which messages can be organized) to the broker, and subscribers register their interest in specific topics with the broker.

There are two implementation options for the DCP: stateful DCP and stateless DCP. A stateful DCP stores the subscription and topic information in a table, and message routing is performed based on table lookup. A stateless DCP does not store the subscription and topic information locally, and message routing is performed based on the in-band information contained within the messages. The in-band information is a numerical representation of the data topology orchestrated and configured by the DO. In our implementation, which is described in detail in Section 3, data forwarding is realized as a modulo operation with the numerical representation of data topology as the input. The

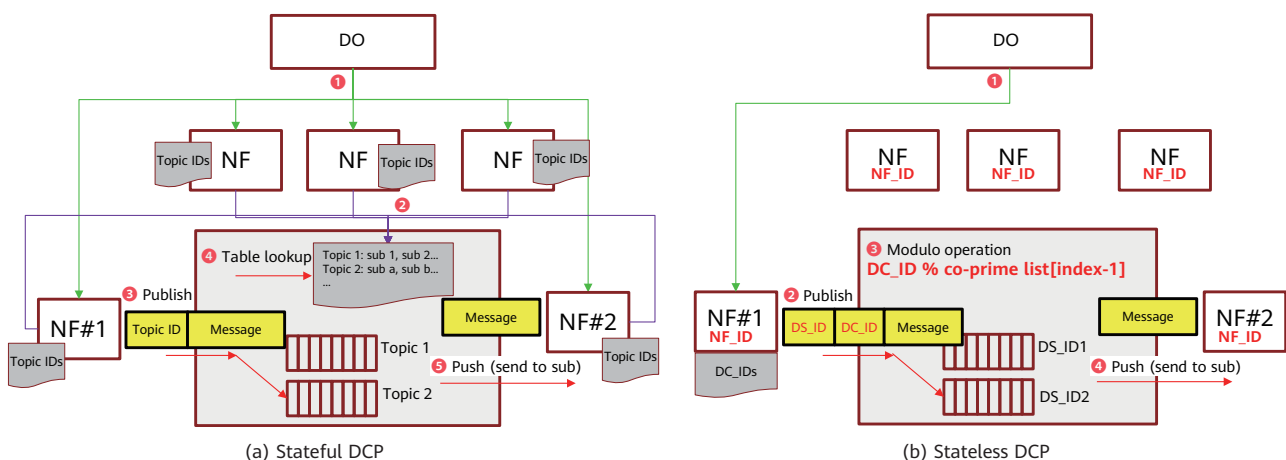


Figure 4 DCP implementation options

stateless DCP implementation has advantages, such as ultra-low latency data forwarding and high scalability, compared to the stateful implementation. Ultra-low latency is achieved by modulo-based data forwarding, which is much faster than the table lookup approach in the stateful implementation. High scalability is achieved by eliminating the massive configuration interactions between the DO and NFs involved in data service provisioning. In the stateless implementation, the DO needs to communicate only with the data source nodes for configuring the numerical representation of data topology contained within the data packets.

The DCPs can be deployed as independent NFs or reside with existing NFs. Figure 5 is an example of a computer vision task involving AI models for age and gender classification with human face detection and dog breed detection. In the example, the DO receives service requests and forms a data processing topology based on the capabilities of the registered DAs/DPFs. The data source is also considered a DA and assigned a unique topic ID, just like the participating DPFs. Here, the live video camera acts as the data source DA, publishing the video frame data with topic ID "Video" to the DCP. The DCP then pushes the video frame data to the queues and then to the data consumers who subscribed to the topic "Video." In this example, the data consumers are two DAs/DPFs, one powered by a human face detection model and the other with a dog

breed detection model. After performing inference, the two DAs/DPFs publish new topics with IDs of "Human face" and "Dog" back to the DCP. The DCP then pushes the age/gender detection results to the DAs/DPFs. In this process, data collection occurs asynchronously with model inference, enabling more sensing data to be collected during the prediction phase.

This DCP-based ML model orchestration is advantageous because the DCP allows an arbitrary number of queues to be bound to the broker. The same data can be reused across ML applications and the output of a model/epoch can be used as input in the upcoming model/epoch. The implementation of the DCP offers several advantages:

- The DCP introduces buffers between the data producers and data consumers, ensuring a seamless and efficient data flow and preventing potential bottlenecks. This is particularly critical when handling large volumes of data, as any delay or interruption in data flows can significantly impact the performance of ML models.
- The DCP accommodates variations in processing speed, preventing data loss while allowing all DPFs to operate at their own pace.
- In collaboration with the DO, the DCP facilitates orderly task execution, which is crucial for successful ML model inference.

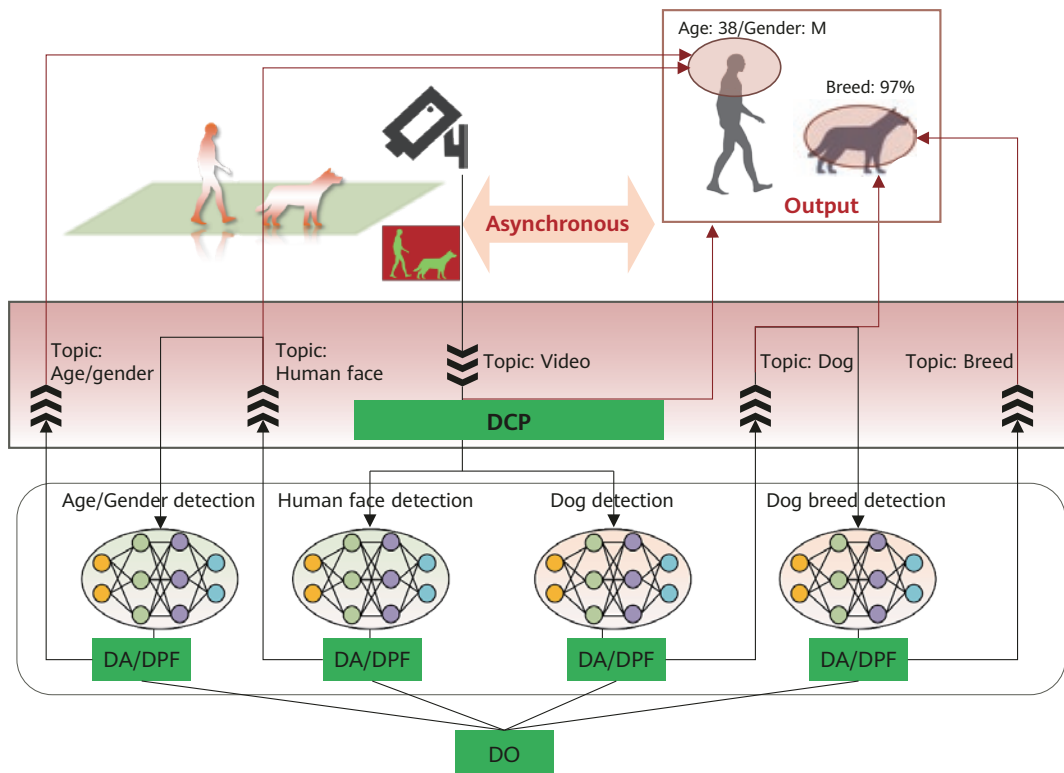


Figure 5 DCP-based ML model orchestration

3 Distributed Message Broker with Stateless Pub/Sub Messaging

3.1 Pub/Sub Paradigm

The advent of AI-powered applications, particularly those adopting advanced neural network architectures, requires a new approach to orchestrate and schedule AI processes within the in-network computing of 6G. We propose a stateless Pub/Sub messaging system to address the critical need for effective information flow management — from data sources to subscribers — on a large-scale distributed AI system.

The Pub/Sub paradigm decouples communication endpoints from one another and enables many-to-many data dissemination, such as in distributed learning, inferencing, and cooperative wireless sensing scenarios. This unique paradigm, featuring efficient data flow management, supports applications involving multiple data producers and consumers and can potentially meet the ML workflow management requirements, leading to substantial enhancement of network AI and MLOps capabilities. Additionally, the Pub/Sub paradigm infuses more dynamism into ML workflow configuration, particularly in terms of response to real-time data or changes. Unlike traditional control flows, which are deterministic and pre-defined, the Pub/Sub paradigm allows for a more reactive and event-driven process, where various nodes on the data pipeline act as subscribers that listen to specific topics published by other nodes or publishers. The decoupling of data producers and consumers inherent in the Pub/Sub paradigm leads to better scalability and improved fault tolerance, as the failure of a single component does not directly impact other components.

3.2 CRT-based Stateless Message Broker

"Beyond-connectivity services" adopt the data-driven paradigm, in which the hidden value of data needs to be leveraged using statistical or ML methods. Considering that all nodes in a data topology need to process data and then send the data to downstream nodes, we propose an on-path data processing approach for the transmission of non-connectivity data, which reduces network bandwidth and ensures data privacy.

The traditional stateful message broker maintains a data forwarding table containing the destination addresses of all subscribers for each topic. Such a broker is required to look up the data forwarding table every time a message arrives, leading to more latency in data transmission. Another issue with the traditional broker is limited scalability, which results from all nodes having to maintain the data forwarding table.

To address these challenges, we propose to use a stateless approach to free the broker of "states." One of the solutions involves having the message data carry a coded "state" as in-band information. In this sense, the design goal of a stateless message broker is to find a function f that is capable of packetizing or transforming a data topology into numerical information suitable for encapsulation in data packets. In the meantime, its reverse function f^{-1} must be computationally simple. Considering these requirements, we design a function f and its reverse function f^{-1} based on the residue number system (RNS) and Chinese remainder theorem (CRT). According to the conventions of RNS literature, we refer to f as the reverse converter and f^{-1} as the forward converter.

- **RNS**

The RNS is a method for representing numbers and executing rapid calculations. It represents an integer as residues obtained by using a set of moduli. Because RNS arithmetic is executed digit-wise without carry propagation, a large number can be decomposed into smaller ones that can be operated on in parallel [4], leading to more efficient computation. The forward converter, shown on the right of Figure 6, converts a large integer X into a set of smaller integer residues (x_1, x_2, x_3) based on the chosen moduli $M = (m_1, m_2, m_3)$. These moduli are carefully selected to be coprime (mutually prime). Converting the number X to an RNS representation is a straightforward process — take the modulo X using each number in the moduli set, and the resulting set of numbers is the RNS representation of X . The formula is as follows:

$$x_i = X \bmod m_i \quad (1)$$

For example, if $X = 11$ and the moduli set is $(3, 4, 5)$, the RNS representation of 11 is $(2, 3, 1)$, represented as follows:

$$11 \equiv (2, 3, 1)_{RNS(3,4,5)} \quad (2)$$

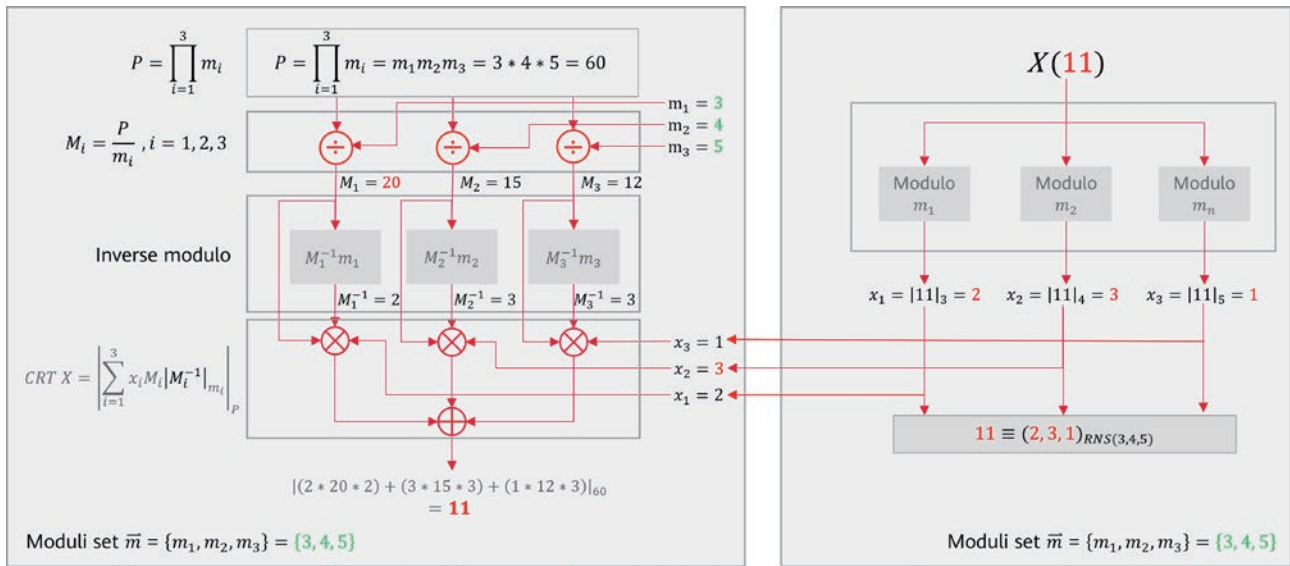


Figure 6 RNS and CRT [4]

- CRT

The CRT can be used as the reverse converter, since it can uniquely reconstruct an integer n from the residues of the Euclidean division of the integer by several integers that are pairwise coprime, as shown on the left of Figure 6. The formula is as follows:

$$P = \prod_{i=1}^3 m_i = m_1 m_2 m_3 = 3 * 4 * 5 = 60, \quad (3)$$

$$M_i = \frac{P}{m_i}, i = 1, 2, 3, \quad (4)$$

$$CRT X = \left| \sum_{i=1}^3 x_i M_i |M_i^{-1}|_{m_i} \right|_P \quad (5)$$

In these formulas, $|M_i^{-1}|_{m_i}$ is the inverse modulo of M_i .

In the context of the DCP, a data topology is represented as a tuple consisting of a moduli set and its residue representation, denoted as (\vec{m}, \vec{r}) . The numerical representation X of a data topology is contained within the message data packets. The forward converter, which is computationally intensive and can be pre-calculated, is integrated into the data orchestration process for data forwarding. In this process, the residues represent the downstream data processing node within the data topology.

3.3 Performance Assessment

One of the primary design objectives of our distributed message broker with stateless Pub/Sub messaging is to minimize the message forwarding latency, which is crucial

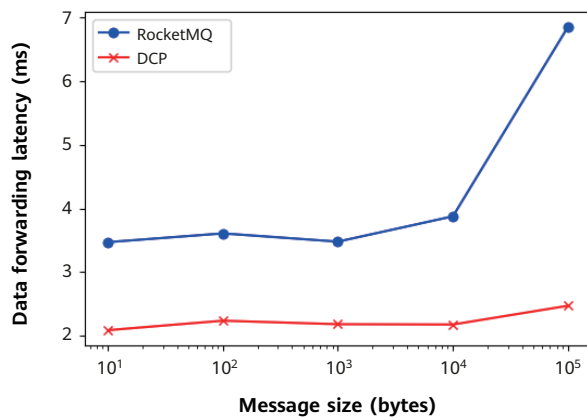
for "beyond-connectivity services" such as collaborative sensing and AI/ML-based autonomous driving. We define message forwarding latency as the time interval from when the first bit of an input message enters a queue to when the last bit exits the queue. Several factors, including the message size, message arrival rate, and processing capacity of the DCP, influence the end-to-end latency.

We conducted a performance comparison of our design against RocketMQ [5] — a distributed message queue system widely used in the information technology (IT) domain for its high-capacity, real-time, and zero-error transactions. RocketMQ operates in single replica mode, which means it is set up with only one broker and without replica nodes. We implemented the DCP in the same setup and conducted comparison experiments with the same hardware and software environment, as detailed in Table 1.

Figure 7 shows the message forwarding latency exhibited in the DCP and RocketMQ systems, respectively, for message sizes of 10^1 bytes to 10^5 bytes. In the test, five messages of different sizes were forwarded by the two systems, with each message forwarded by both systems five times repeatedly to calculate the average latency. As shown in the figure, the RocketMQ system exhibits a message-forwarding latency that increases with the message size. A noticeable surge occurs when the message size reaches 10^5 bytes. In contrast, the message-forwarding latency of our DCP design is stable at around 2 ms across all tested message sizes. Overall, our DCP design exhibits an average latency reduction of over 65% compared to RocketMQ for all message sizes, and the reduction is particularly pronounced for larger message sizes (e.g., 10^5 bytes).

Table 1 Experimental environment for DCP implementation

Category	Details
Operating system	Ubuntu 22.04.4 LTS (Jammy Jellyfish)
Kernel version	5.15.0-113-generic
CPU	Two-socket architecture with 40 cores per socket, 2 threads per core running at a base frequency of 2.30 GHz.
Memory	503 GiB memory
Storage	894.3 GB SSD 1.8 TB HDD
RocketMQ	5.1.4
Java	1.8.0_392
Docker	24.07

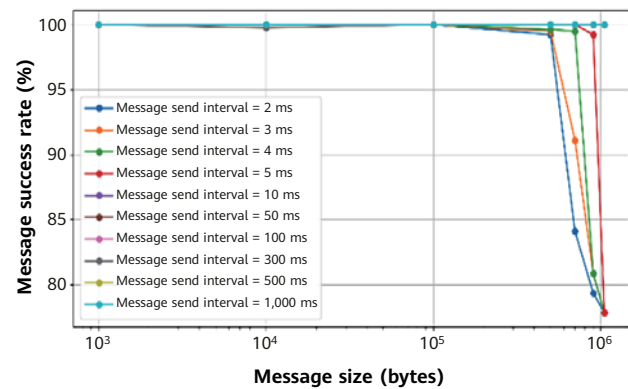
**Figure 7** Data forwarding latency of the DCP and RocketMQ systems

DCP's superiority over RocketMQ in message forwarding latency is mainly attributed to two factors:

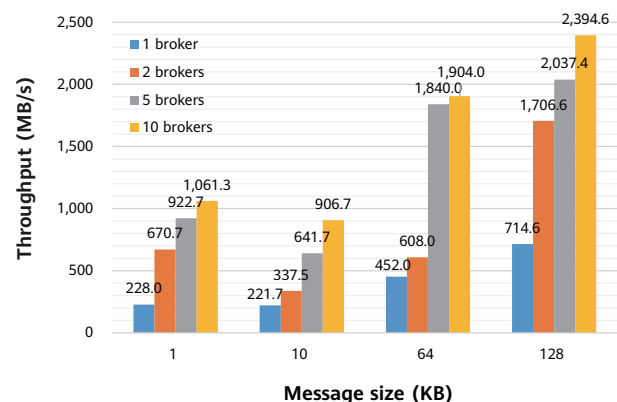
- Stateless design:** The DCP employs a stateless design, which eliminates the need to maintain and manage the complex state information for each message-forwarding operation. This design reduces the computational overhead associated with processing each message, thereby significantly lowering the message-forwarding latency.
- Elimination of message persistence:** In DCP application scenarios, message persistence is not required. Therefore, we removed persistence-related components from the system. This further simplifies I/O operations and reduces the overhead of writing data to the disk, thereby improving the data-forwarding efficiency.

In contrast, the RocketMQ system requires message persistence, resulting in significantly increased latency when processing large messages.

We also studied the message success rates for different message sizes at different message send intervals (2 ms, 3 ms, 4 ms, 5 ms, 10 ms, 50 ms, 100 ms, 300 ms, 500 ms, and 1,000 ms) in the DCP system. As shown in Figure 8, when the message size is within the range of 10³ bytes to 10⁵ bytes, the success rates at all send intervals are close to 100%. This indicates that the DCP system can effectively handle and successfully send messages of smaller sizes. However, when the message size increases to 10⁵ bytes and above, the success rates for all send intervals decrease.

**Figure 8** Message success rate of the DCP system

The result suggests that the message send interval has minimal impact on the success rate for smaller message sizes; however, as the message size increases, shorter message send intervals lead to lower success rates. This is mainly due to network congestion or insufficient system processing capacity. Therefore, when designing a message transmission strategy, it is crucial to balance the message size and send interval in order to ensure high success rates, even in heavy-load scenarios.

**Figure 9** Throughput of the DCP system

Additionally, we evaluated the DCP system's throughput for different message sizes (1 KB, 10 KB, 64 KB, and 128 KB) with configurations of 1, 2, 5, and 10 brokers, respectively. As shown in Figure 9, the DCP system's throughput increases as the number of brokers increases. For instance, when the message size is 1 KB, the throughput is 228.0 MB/s with 1 broker and 1,061.3 MB/s with 10 brokers. Similarly, when the message size is 128 KB, the throughput is 714.6 MB/s with 1 broker and 2,394.6 MB/s with 10 brokers.

The result also shows that the DCP system's throughput increases as the message size increases. This indicates that the DCP system is more efficient in handling larger messages as long as the send interval is appropriately managed to prevent the success rate from declining. Additionally, increasing the number of brokers has a more pronounced effect on enhancing the throughput for larger messages.

4 Conclusion

"Beyond-connectivity services," such as network AI and wireless sensing, generate and consume data that requires an efficient transmission mechanism capable of supporting complex data topologies and on-path data processing. However, the current network architecture employs a session-based model for managing UE communication data, which cannot meet the requirements of "beyond-connectivity services." To overcome this limitation, this research presents a dedicated DP for transmitting non-connectivity data. The incorporation of DO and DA enables the management of flexible data topologies, facilitating DaaS for network AI and wireless sensing through on-path processing of distributed data and ubiquitous computing. Additionally, we present an asynchronous data exchange mechanism, the DCP, which enhances data exchange efficiency among multiple data producers and consumers. As a stateless and distributed message broker, the DCP has shown to be particularly effective in scenarios requiring collaboration between multiple AI models and in ISAC sensing for environment reconstruction.

References

- [1] 3GPP, "3GPP TS 23.501 - System architecture for the 5G system (Rel 15)," 2021.
- [2] International Telecommunication Union (ITU), "IMT traffic estimates for the years 2020 to 2030," [Online]. Available: <https://www.itu.int/pub/r-rep-m.2370> (accessed on Oct. 25, 2023)
- [3] Z. Qin *et al.*, "6G data plane: A novel architecture enabling data collaboration with arbitrary topology," *Mobile Networks and Applications*, vol. 28, no. 1, pp. 394-405, 2023.
- [4] H. L. Garner, "The residue number system," *IRE Transactions on Electronic Computers*, pp. 140-147, 1959.
- [5] RocketMQ, [Online]. Available: <https://rocketmq.apache.org/docs/>



In-Network Learning for Distributed RAN AI

Distributed LLMs via Latent Structure Distillation

Abdellatif Zaidi¹, Romain Chor¹, Piotr Krasnowski¹, Milad Sefidgaran¹, Rong Li¹,
Fei Wang², Chenghui Peng², Shaoyun Wu², Jean-Claude Belfiore¹

¹Advanced Wireless Technology Lab

²Wireless Technology Lab

Abstract

In this paper, we propose a distributed learning algorithm, named In-Network Learning (INL), for inference over wireless radio access networks (RANs) without transmitting raw data. It is shown that this algorithm is particularly suitable for both *multimodal* and *heterogeneous* data settings where the fusion of features extracted in a distributed manner is necessary for making reliable decisions. In these cases, it offers substantial gains over state-of-the-art (SOTA) solutions such as Horizontal and Vertical Federated Learning (FL) and Horizontal and Vertical Split Learning (SL) in terms of both accuracy and bandwidth requirements. We also discuss how the algorithm can be extended to support the deployment of large language models (LLMs) and knowledge distillation in wireless networks.

1 Distributed Inference over Wireless RANs

Wireless radio access networks (RANs) have important intrinsic features that may pave the way for crossfertilization between machine learning (ML) and communication. This is in contrast to simply replacing one or more communication modules by applying ML algorithms as black boxes. For example, while relevant data is generally available at one point in areas such as computer vision and neuroscience, it is typically highly distributed across several sites in wireless networks. Examples include channel state information (CSI) or the so-called radio-signal strength indicator (RSSI) of a user's signal, which can be used for things like localization, precoding, or beam alignment [1].

A common approach for implementing ML solutions in wireless networks involves collecting all relevant data at one site (e.g., a cloud server or macrobase station) and then training a suitable ML model using all available data and processing power. However, this approach may not be suitable in many cases due to the large volumes of data and the scarcity of network resources, such as power and bandwidth. Additionally, some applications (e.g., automatic vehicle driving) might have stringent latency requirements that are incompatible with data sharing. In other cases, it might be desirable to not share the raw data in order to protect user privacy. Furthermore, edge devices such as small base stations, on-board sensors, and smartphones typically have limited memory and computational power. Also, the wireless environment is typically prone to rapid changes, for example, connectivity fluctuations or devices dynamically joining or leaving the network. Another critical aspect is that the data is largely multimodal, heterogeneous, or both across devices and users. Table 1 summarizes the main features of inference over wireless RANs.

1.1 AI at the Wireless Edge

The challenges discussed earlier have called for a new paradigm, referred to as "edge learning" or distributed learning, in which intelligence moves from the heart of the network to its edge. In this new paradigm, communication plays a central role in the design of efficient ML algorithms and architectures because both data and computational resources, which are the main ingredients of an efficient ML solution, are highly distributed. The goal of distributed inference over RAN is to make decisions on one or more tasks, at one or more sites, by exploiting the available distributed data. In this framework, multiple devices (e.g., BSs and UEs) are each equipped with a neural network (NN). Some of the devices possess data they have acquired through communication or sensing, whereas some only contribute to the collective intelligence through computational power. See Figure 1.

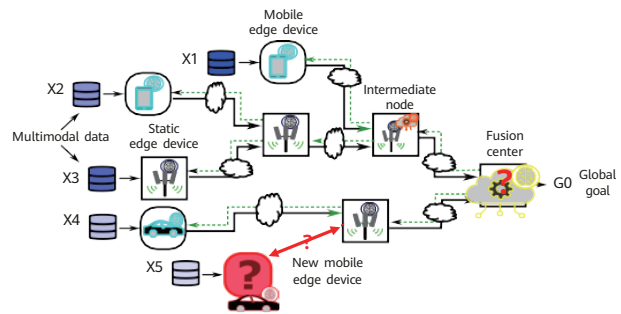


Figure 1 Distributed inference over wireless RAN

1.2 A Brief Review of SOTA Algorithms

AI solutions for RANs can be classified according to whether only the training phase is distributed (such as Horizontal Federated Learning and Horizontal Split Learning) or both the training and inference (or test) phases are distributed (such as Vertical Federated Learning).

Table 1 Summary of the main features of inference over wireless RAN

Data	Inference	Network connectivity/topology	Privacy	Compute resources
<ul style="list-style-type: none"> Distributed during training Distributed during inference Multimodal Heterogeneous 	<ul style="list-style-type: none"> Distributed Communication is the bottleneck Fusion is required Extremely short latency (≈ 0.1 ms) 	<ul style="list-style-type: none"> Prone to rapid changes: <ul style="list-style-type: none"> Users joining or leaving the network Link failure Channel quality drop-off 	<ul style="list-style-type: none"> Critical <ul style="list-style-type: none"> Raw data leaks information about users 	<ul style="list-style-type: none"> Small Distributed across sites

- Horizontal Federated Learning (HFL):** Perhaps the most popular distributed learning architecture is HFL of [2]. This architecture, as already mentioned, is most suitable for settings in which the training phase is performed in a distributed manner while the inference phase is performed centrally. During the training phase, each client is equipped with a distinct copy of a same NN model that the client trains on its local dataset. The learned weight parameters are then sent to and aggregated by (e.g., their average is computed) a cloud server or parameter server (PS). This process is repeated, each time using the obtained aggregated model for reinitialization, until convergence. The rationale is that this approach ensures the model is progressively adjusted to account for all variations in the data, not only those of the local dataset.
- Vertical Federated Learning (VFL):** In VFL [3], a variation of Federated Learning (FL), the data is partitioned vertically and both the training and inference phases are distributed. Figure 2 illustrates both HFL and VFL. In this case, every client holds data that is relevant for a possibly distinct feature. A prominent example is when the data is heterogeneous across clients or multimodal. In VFL, different clients may apply distinct NN models that are better tailored for their own data modalities. These models are trained jointly to extract features that collectively are enough to make a reliable decision at the fusion center. An illustration is given in Figure 3a. For recent advances on VFL and its applications in wireless settings, refer to [4, 5], and references therein.
- Split Learning (SL):** SL was introduced in [6]. Similar to FL, it has two variations: Horizontal SL (HSL) and Vertical SL (VSL). Although VSL was introduced earlier than VFL, VSL is now viewed as a special case of VFL. As such, we will not discuss it here. For HSL, a two-part

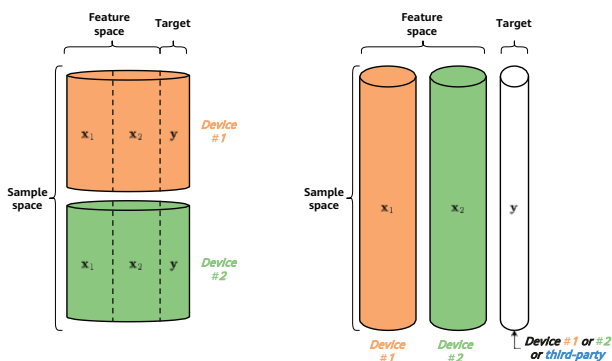


Figure 2 HFL (left) and VFL (right)

NN model is split into an encoder part and a decoder part. Each edge device possesses a copy of the encoder part and both NN parts are learned sequentially. The decoder does not have its own data, whereas in every training round, the NN encoder part is fed with the data of one device and its parameters are initialized using those learned from the previous round. Then, during the inference phase, the learned two-part model is applied to centralized data.

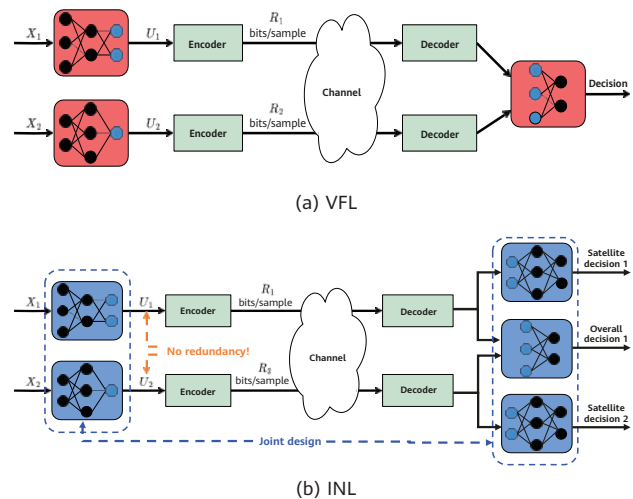


Figure 3 Feature redundancy removal by INL

2 In-Network Learning

The roots of In-Network Learning (INL) date back to [7, 8], with further development taking place in [9–11].

2.1 Overview

INL is most suited to distributed inference from heterogeneous and multimodal data. In this scheme, every device is equipped with an NN. During inference, each device independently extracts suitable features from its input data for a given inference task. These features are then transmitted over the network and fused at a given fusion center in order for a desired reliable decision to be made. The devices that hold useful data (these devices play the role of *encoders*) perform individual feature extraction independently from each other. Through training, the algorithm ensures that the encoders only extract *complementary* features and that, for instance, redundant inter-device features are removed, enabling substantial bandwidth savings. In summary, the key technical components of this algorithm are as follows:

- **Network Feature Fusion:** INL fuses features that are extracted in a distributed manner at a fusion center so that, collectively, they enable a desired decision to be made at the fusion center after being transmitted over the network.
- **Feature Redundancy Removal:** A distinguishing factor of INL is that, during inference, the encoders *only extract non-redundant features* and are trained to do so during training. Specifically, during inference, each encoder only extracts features that are useful for a given inference task from its input data while also taking into account the other features extracted by other encoders.
- **Feature Extraction Depends on Network Channel Quality:** Encoder feature extraction also takes into account the quality of the channel to the fusion center. That is, features are extracted only to the extent that it is possible to transmit them reliably to the decision maker.
- **Satellite Decoders:** The fusion center is equipped with a main decoder and satellite decoders, which are trained to make soft decisions based on the individual features transmitted by the encoders. The system is depicted in Figure 3b.

2.2 Formal Description

We model an N -node network by a directed acyclic graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{C})$, where $\mathcal{N} = [1 : N]$ is the set of nodes, $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ is the set of edges, and $\mathcal{C} = \{C_{jk} : (j, k) \in \mathcal{E}\}$ is the set of edge weights. Each node represents a device, and each edge represents a communication channel with capacity C_{jk} . The processing at the nodes of the set \mathcal{J} is such that each of them assigns an index or message $m_{jl} \in [1, M_{jl}]$ to each $x_j \in \mathcal{X}_j$ and each received index tuple $(m_{ij} : (i, j) \in \mathcal{E})$ for each edge $(j, l) \in \mathcal{E}$. Specifically, let for $j \in \mathcal{J}$ and l such that $(j, l) \in \mathcal{E}$ the set $\mathcal{M}_{jl} = [1 : M_{jl}]$. The encoding function at node j is

$$\phi_j : \mathcal{X}_j \times \left\{ \prod_{i:(i,j) \in \mathcal{E}} \mathcal{M}_{ij} \right\} \rightarrow \prod_{l:(j,l) \in \mathcal{E}} \mathcal{M}_{jl}, \quad (1)$$

where \times designates the Cartesian product of sets. Similarly, for $k \in [1 : N - 1]/\mathcal{J}$, node k assigns an index $m_{kl} \in [1, M_{kl}]$ to each index tuple $(m_{ik} : (i, k) \in \mathcal{E})$ for each edge $(k, l) \in \mathcal{E}$. That is,

$$\phi_k : \prod_{i:(i,k) \in \mathcal{E}} \mathcal{M}_{ik} \rightarrow \prod_{l:(k,l) \in \mathcal{E}} \mathcal{M}_{kl}. \quad (2)$$

The range of the encoding functions $\{\phi_i\}$ are restricted in size, as

$$\log |\mathcal{M}_{ij}| \leq C_{ij} \quad \forall i \in [1, N - 1] \quad \text{and} \quad \forall j : (i, j) \in \mathcal{E}. \quad (3)$$

Node N needs to infer on the random variable $Y \in \mathcal{Y}$ using all incoming messages, specifically

$$\psi : \prod_{i:(i,N) \in \mathcal{E}} \mathcal{M}_{iN} \rightarrow \hat{\mathcal{Y}}. \quad (4)$$

We choose the reconstruction set $\hat{\mathcal{Y}}$ to be the set of distributions on \mathcal{Y} , where $\hat{\mathcal{Y}} = \mathcal{P}(\mathcal{Y})$. We also measure discrepancies between true values of $Y \in \mathcal{Y}$ and their estimated fits in terms of average logarithmic loss, specifically for $(y, \hat{P}) \in \mathcal{Y} \times \mathcal{P}(\mathcal{Y})$

$$d(y, \hat{P}) = \log \frac{1}{\hat{P}(y)}. \quad (5)$$

As such, the performance for a given network topology and bandwidth budget $\{C_{ij}\}$ in INL is measured by the achieved *relevance* level, evaluated as

$$\Delta = H(Y) - \mathbb{E} [d(Y, \hat{Y})]. \quad (6)$$

For example, for classification problems, the relevance measure (6) is directly related to the miss-classification error.

In practice, in a supervised setting, the mappings given by (1), (2), and (4) need to be learned from a set of training data samples $\{(x_{1,i}, \dots, x_{J,i}, y_i)\}_{i=1}^n$. The data is distributed such that samples $\mathbf{x}_j := (x_{j,1}, \dots, x_{j,n})$ are available at node j for $j \in \mathcal{J}$ and the desired predictions $\mathbf{y} := (y_1, \dots, y_n)$ are available at the end decision node N . We parametrize the possibly stochastic mappings (1), (2), and (4) using NNs, which can be arbitrary and independent. For example, unlike in FL, these NNs do *not* need to be identical. It is only required that the following mild condition be met (this condition, as will become clearer from what follows, facilitates the back-propagation). For $j \in \mathcal{J}$ and $x_j \in \mathcal{X}_j$ ¹

$$\begin{aligned} &\text{Size of first layer of NN } (j) = \\ &\text{Dimension } (x_j) + \sum_{i:(i,j) \in \mathcal{E}} (\text{Size of last layer of NN } (i)). \end{aligned} \quad (7)$$

Similarly, for $k \in [1 : N]/\mathcal{J}$, we set

$$\begin{aligned} &\text{Size of first layer of NN } (k) = \\ &\sum_{i:(i,k) \in \mathcal{E}} (\text{Size of last layer of NN } (i)). \end{aligned} \quad (8)$$

In the following, without loss of generality, we illustrate the training and inference phases formally for the example network topology shown in Figure 4.

¹ We assume all the elements of \mathcal{X}_j have the same dimension.

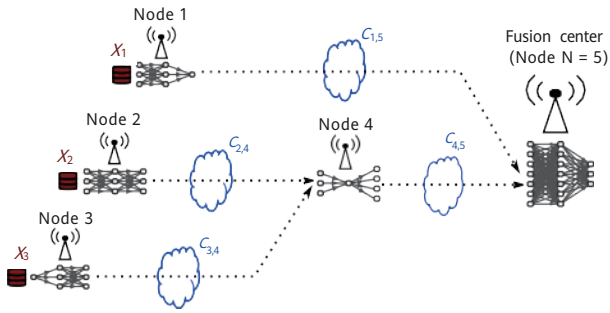


Figure 4 INL for an example RAN

1. **Training Phase:** During the forward pass, every node $j \in \{1, 2, 3\}$ processes mini-batches of its training dataset \mathbf{x}_j . The size of these mini-batches is b_j . Nodes 2 and 3 send their vectors formed of the activation values of the last layer of their NNs to node 4, where the vectors are concatenated vertically at the input layer of NN 4, due to (7). The forward pass continues on the NN at node 4 until its last layer. Next, nodes 1 and 4 send the activation values of their last layers to node 5. Again, as the sizes of the last layers of the NNs of nodes 1 and 4 satisfy (8), the sent activation vectors are concatenated vertically at the input layer of NN 5. Likewise, the forward pass continues until the last layer of NN 5. All nodes update their parameters using a standard back-propagation technique, as follows. For node $t \in \mathcal{N}$, let L_t denote the index of the last layer of the NN at node t . Also, let, for $\mathbf{w}_t^{[l]}$, $\mathbf{b}_t^{[l]}$, and $\mathbf{a}_t^{[l]}$ denote respectively the weights, biases, and activation values at layer $l \in [2 : L_t]$ for the NN at node t . σ is the activation function, and $\mathbf{a}_t^{[1]}$ denotes the input to the NN. Node t computes the error vectors

$$\delta_t^{[L_t]} = \nabla_{\mathbf{a}_t^{[L_t]}} \mathcal{L}_s^{NN}(b) \odot \sigma'(\mathbf{w}_t^{[L_t]} \mathbf{a}_t^{[L_t-1]} + \mathbf{b}_t^{[L_t]}) \quad (9a)$$

$$\delta_t^{[l]} = [(\mathbf{w}_t^{[l+1]})^T \delta_t^{[l+1]}] \odot \sigma'(\mathbf{w}_t^{[l]} \mathbf{a}_t^{[l-1]} + \mathbf{b}_t^{[l]}) \quad \forall l \in [2, L_t - 1] \quad (9b)$$

$$\delta_t^{[1]} = [(\mathbf{w}_t^{[2]})^T \delta_t^{[2]}] \quad (9c)$$

and then updates its weight- and bias-parameters as

$$\mathbf{w}_t^{[l]} \rightarrow \mathbf{w}_t^{[l]} - \eta \delta_t^{[l]} (\mathbf{a}_t^{[l-1]})^T, \quad (10a)$$

$$\mathbf{b}_t^{[l]} \rightarrow \mathbf{b}_t^{[l]} - \eta \delta_t^{[l]}, \quad (10b)$$

where η designates the learning parameter².

During the backward pass, each NN updates its parameters according to (9) and (10). Node 5 is the first to apply the back-propagation procedure in order to update the parameters of its NN. It applies (9) and (10) sequentially, starting from its last layer.

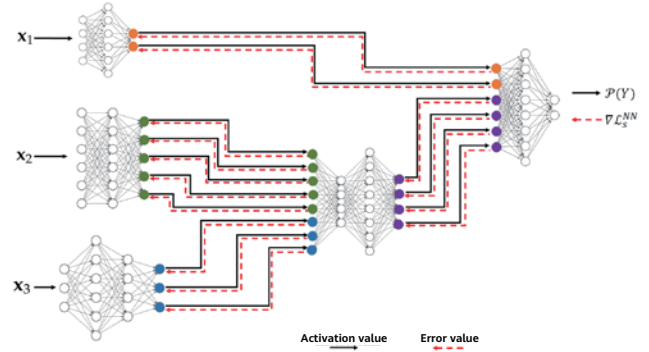


Figure 5 Forward and backward passes for the network topology of Figure 4

- The error propagates back until it reaches the first layer of the NN of node 5. Node 5 then horizontally splits the error vector of its input layer (9c) into 2 sub-vectors. The size of the top sub-error vector is that of the last layer of the NN of node 1, and the size of the bottom sub-error vector is that of the last layer of the NN of node 4 — see Figure 5. Similarly, the two nodes 1 and 4 continue the backward propagation during their turns simultaneously. Node 4 then horizontally splits the error vector of its input layer (9c) into 2 sub-vectors. Likewise, the size of the top sub-error vector is that of the last layer of the NN of node 2, and the size of the bottom sub-error vector is that of the last layer of the NN of node 3. Finally, the backward propagation continues on the NNs of nodes 2 and 3. The entire process is repeated until convergence.
2. **Inference Phase:** During inference, nodes 1, 2, and 3 observe (or measure) each new sample. Let x_1 , x_2 , and x_3 be the samples observed by nodes 1, 2, and 3, respectively. Node 1 processes x_1 using its NN and sends an encoded value \mathbf{u}_1 to node 5; similarly, nodes 2 and 3 send their encoded values towards node 4. Upon receiving \mathbf{u}_2 and \mathbf{u}_3 from nodes 2 and 3, node 4 concatenates them vertically and processes the obtained vector using its NN. The output \mathbf{u}_4 is then sent to node 5. The latter performs similar operations on the activation values \mathbf{u}_1 and \mathbf{u}_4 . This node then outputs an estimate of the label y in the form of a soft output $Q_{\phi_5}(y|\mathbf{u}_1, \mathbf{u}_4)$.

² For simplicity, η and σ are assumed here to be identical for all NNs.

3 Performance Gains

In this section, we compare the algorithms HFL and VFL, SL, and our INL for data classification problems, in terms of achieved accuracy and bandwidth requirements.

3.1 INL vs. HFL and HSL

Experiment 1: We create five sets of noisy versions of images obtained from the CIFAR-10 dataset [12]. The images are first normalized, and they are then corrupted by additive Gaussian noise with standard deviation set respectively to 0.4, 1, 2, 3, 4. For INL, each of the five input NNs is trained on a different noisy version of the same image. Each NN uses a variation of the VGG network of [13], with the categorical cross-entropy as the loss function. The architecture is shown in Figure 6. In the experiments, all five noisy versions of every CIFAR-10 image are processed simultaneously, each by a different NN at a distinct node.

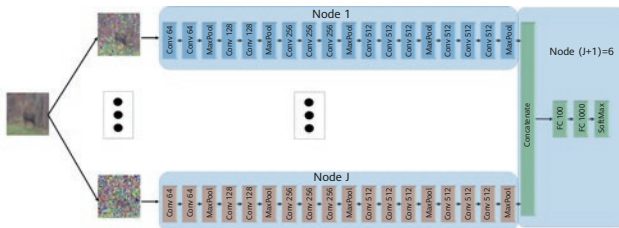


Figure 6 Network architecture. Conv stands for a convolutional layer, and FC stands for a fully connected layer.

Subsequently, the outputs are concatenated and then passed through a series of fully connected (FC) layers at node $(J + 1)$. For HFL, each of the five client nodes is equipped with the *entire* network of Figure 6. The dataset is split into five sets of equal sizes, with the split being performed such that all five noisy versions of a given CIFAR-10 image are presented to the same client NN (note, however, that distinct clients observe different images).

For HSL, each input node is equipped with an NN formed by *all* five branches with convolution networks (i.e., the entire

network shown in Figure 6, except the part at Node $(J + 1)$). Furthermore, node $(J + 1)$ is equipped with fully connected layers at Node $(J + 1)$ in Figure 6. Here, the processing during training is such that each input NN vertically concatenates the outputs of all convolution layers and then passes that to node $(J + 1)$, which then propagates back the error vector. After one epoch at one NN, the learned weights are passed to the next client, which performs the same operations on its part of the dataset.

Figure 7 shows the amount of data needed to be exchanged among the nodes (i.e., bandwidth resources) in order to get a prescribed value of classification accuracy. Observe that our INL requires significantly less data transmission than HFL and HSL for the same desired accuracy level.

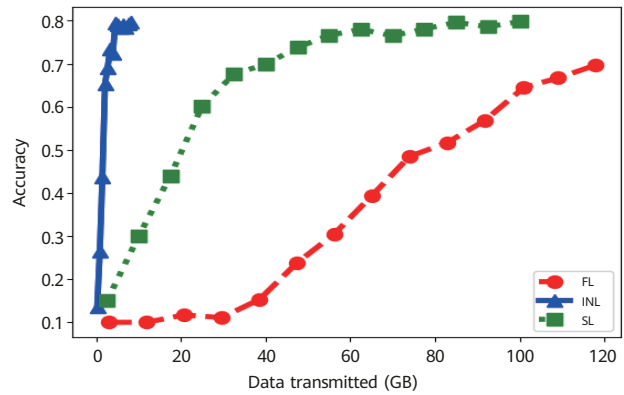


Figure 7 Accuracy vs. bandwidth cost for Experiment 1

Experiment 2: In the previous experiment, the entire training dataset was partitioned differently for INL and HFL in order to account for their unique characteristics. In this second experiment, they are all trained on the same data. Specifically, each client NN sees all CIFAR-10 images during training, and its local dataset differs from those seen by other NNs only by the amount of added Gaussian noise (standard deviation chosen respectively as 0.4, 1, 2, 3, 4). Also, to ensure a fair comparison of the three schemes, INL, HFL, and HSL, we set the nodes to utilize the same NNs fairly for each of them. See Figure 8.

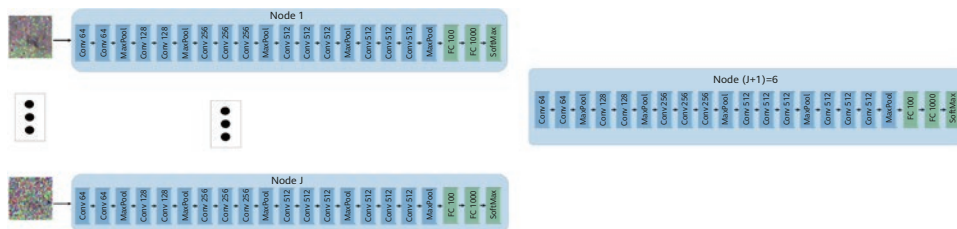


Figure 8 Used NN architecture for Experiment 2

Figure 9 shows the performance of the three schemes during the inference phase in this experiment. For HFL, the inference is performed on the image whose quality is the average among the five noisy input images used for INL. Again, observe the benefits of INL over HFL and HSL in terms of both achieved accuracy and bandwidth requirements.

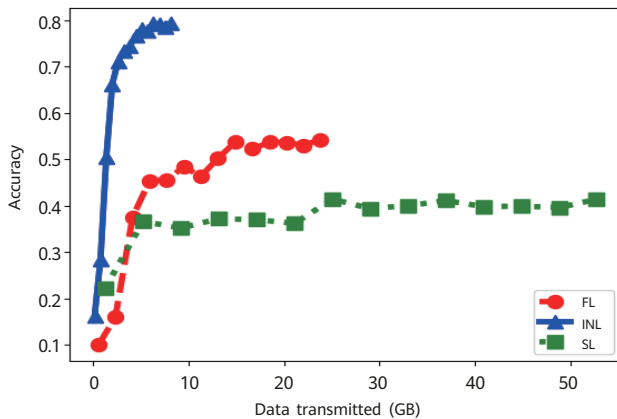


Figure 9 Accuracy vs. bandwidth cost for Experiment 2

3.2 INL vs. VFL

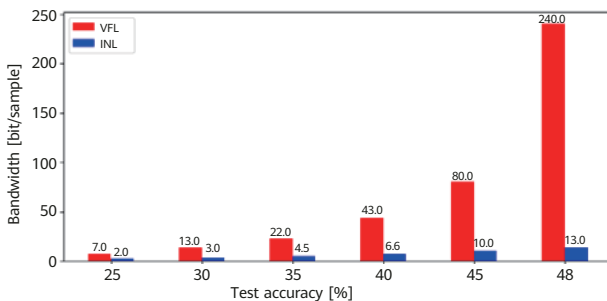
Experiment 3: In our third experiment, we compare our INL with VFL (as mentioned earlier, VSL can be viewed as

a special case of VFL). We use variations of the CIFAR-10 dataset. In particular, we consider two encoders and one decoder, which communicate over a wireless network with a channel capacity of R_1 bits per channel used from Encoder 1 to the decoder and R_2 bits per channel used from Encoder 2 to the decoder (see Figure 3b). The dimension of the latent vectors U_1 and U_2 is set to 64. At every time point during the inference phase, the input of the first encoder is a half-occluded copy of a given CIFAR-10 image, and that of the second encoder is a noisy version of the same image. The encoders use CNNs with a few convolutional layers or ResNet-18 [14]. The decoders consist of FC layers.

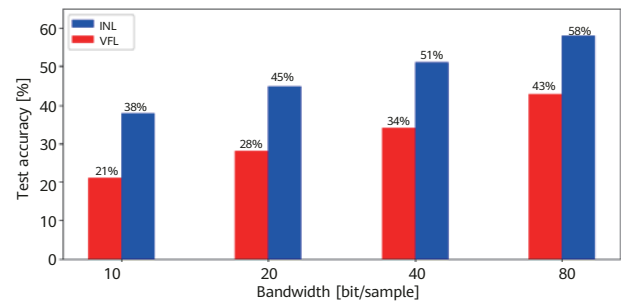
As shown in Figure 10 for CNN encoders and in Figure 11 for ResNet-18 encoders, VFL requires around three times more bandwidth than our INL in order to reach the same AI service accuracy. For some values of accuracy, the advantage of INL in terms of communication cost saving can be significantly higher. Moreover, for a bandwidth of $R_1 = R_2 = 20$ bits/sample, INL already achieves 45% accuracy, whereas VFL achieves only 28%.

4 LLM

In addition to having remarkable capabilities, large language models (LLMs) are revolutionizing AI development

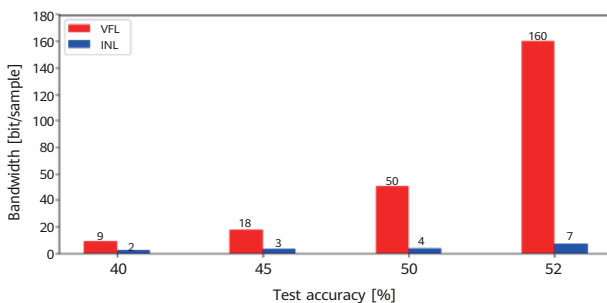


(a) Bandwidth requirements for various desired accuracy levels

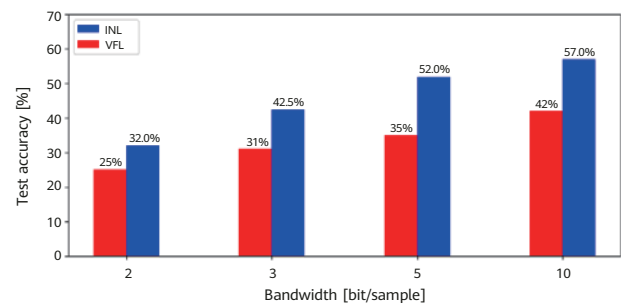


(b) Test accuracy vs. available bandwidth $R_1 = R_2 = R$ bits/sample

Figure 10 Performance gains of INL vs. VFL for Experiment 3 using CNN encoders



(a) Bandwidth requirements for various desired accuracy levels

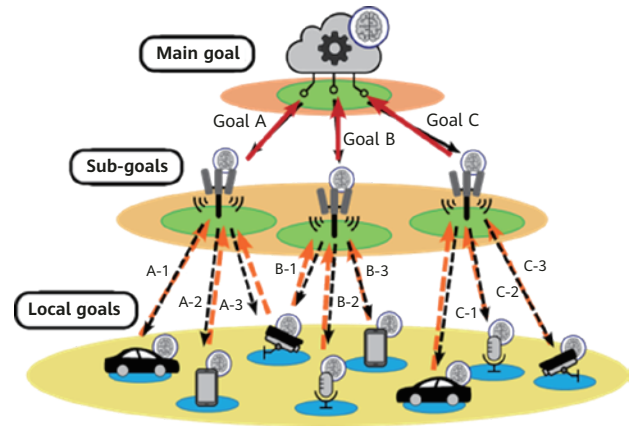


(b) Test accuracy vs. available bandwidth $R_1 = R_2 = R$ bits/sample

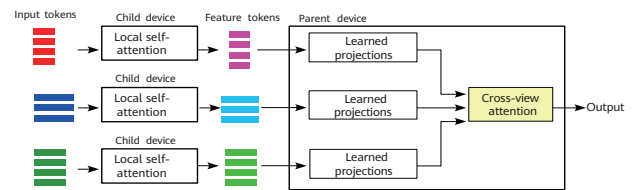
Figure 11 Performance gains of INL over VFL for Experiment 3 using ResNet-18 encoders

and potentially shaping our future. However, their multimodality, in part, causes some critical challenges for cloud-based deployment: (i) long response time, (ii) high communication bandwidth cost, and (iii) infringement of data privacy. As a result, there is an urgent need to leverage Mobile Edge Computing (MEC) in order to finetune and deploy LLMs on or in closer proximity to data sources while also preserving data ownership for end users. In accordance with the "NET4AI" (network for AI) vision for the 6G era [15], we envision a 6G MEC architecture that can support the deployment of LLMs at the network edge. Our proposed architecture includes the following critical modules.

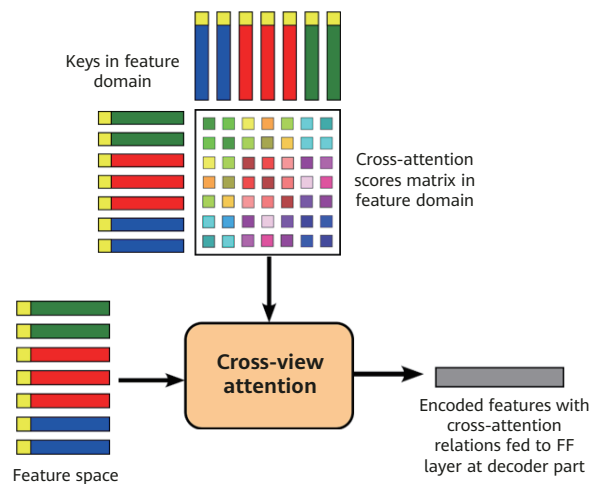
- Goal Decomposition:** The global inference task is performed collaboratively between different layers of the mobile network system. The fusion center decomposes the global goal into smaller sub-goals and assigns them to the next-layer BSs based on their respective strengths. The BSs then further decompose the subgoals into smaller ones. This process continues until it reaches the edge devices. See Figure 12a.
- Cross-View Attention:** The self-attention of transformers can only be computed for locally available sensory data. If multiple sensors acquire multi-view data that is relevant for a given inference task, it is necessary to compute how a token from a given piece of data collected at one sensor attends to another token from another piece of data collected or measured at another sensor. We call this *cross-view attention*, which is computed at a fusion center in the feature space after feature projection on a hyperplane. See Figures 12b, 12c, and 12d.
- Latent Structure-based Knowledge Distillation:** It is expected that 6G will evolve into a mobile network supporting in-network and distributed AI at the edge [15]. However, considering the excessive memory and compute requirements of LLMs, is it feasible to run such large models at the 6G edge? Also, would the network bandwidth support various agents/devices equipped with LLMs exchanging the entirety of their models for model aggregation and collaboration? A step in this direction has been accomplished in [16] recently, where devices use INL to only exchange the structure of their extracted features, not the features themselves. This structure is then utilized onsite at the device to fine-tune the locally extracted features.



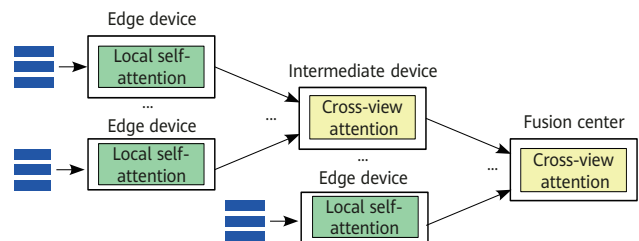
(a) Hierarchical goal decomposition for decision making over a RAN



(b) Feature projection for cross-view attention computation



(c) Cross-view attention computation in the feature space domain

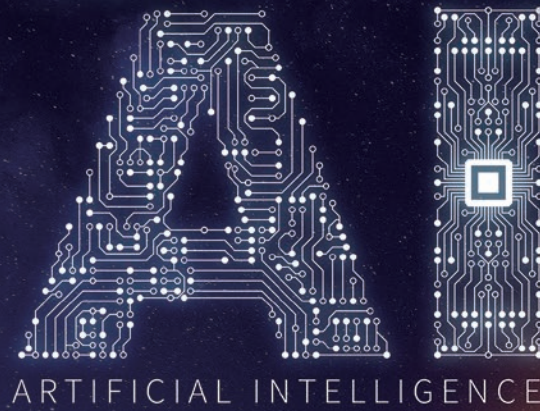


(d) Hierarchical cross-view attention computation

Figure 12 Main components of LLM-based INL

References

- [1] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 763–776, 2017.
- [2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.
- [3] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen, and T. Ranbaduge, "Vertical federated learning: Challenges, methodologies and experiments," *arXiv preprint arXiv: 2202.04309*, 2022.
- [4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [6] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *Journal of Network and Computer Applications*, vol. 116, pp. 1–8, 2018.
- [7] I. E. Aguerri and A. Zaidi, "Distributed Variational Representation Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 120–138, 2021.
- [8] I. Estella Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in *International Zurich Seminar on Information and Communication (IZS 2018). Proceedings*. ETH Zurich, 2018, pp. 35–39.
- [9] M. Moldoveanu and A. Zaidi, "In-network Learning for Distributed Training and Inference in Networks," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.
- [10] —, "On in-network learning. A comparative study with federated and split learning," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2021, pp. 221–225.
- [11] —, "In-network learning: Distributed training and inference in networks," *Entropy*, vol. 25, no. 6, p. 920, 2023.
- [12] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv: 1409.1556*, 2014.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [15] L. Huawei *et al.*, "6G: The next horizon from connected people and things to connected intelligence," *Huawei, White Paper*, 2022.
- [16] M. Sefidgaran, A. Zaidi, and P. Krasnowski, "Minimum description length and generalization guarantees for representation learning," *Advances in Neural Information Processing Systems*, vol. 36, 2023.



A Novel Federated Learning Method for Distributed Generation of 6G Air Interface Data

Mingfeng Xu, Yang Li, Wei Zhou, Hui Liu*, Jiamo Jiang
Mobile Communications Innovation Center, CAICT

Abstract

Integrating artificial intelligence (AI) with communications technologies will enhance the intelligence of communications systems, driving their evolution towards inclusive intelligence and achieving the 6G vision of "the Intelligent Network of Everything." However, creating high-quality AI models requires massive volumes of data. To reduce the significant overhead of collecting vast and diverse air interface data, this paper explores the feasibility of utilizing generative adversarial network (GAN) models to synthesize artificial air interface channel data. It also investigates the method of training distributed GAN models through federated learning (FL) architecture to fully leverage edge node computing power and reduce network transmission load. Simulation results demonstrate that these models generate channel data comparable in quality to centrally trained models, offering a novel approach for future air interface data collection.

Keywords

6G, GAN, FL

* Corresponding author

1 Introduction

Since the 2020s, a number of groundbreaking artificial intelligence (AI) applications like AlphaFold, ChatGPT, DALL-E3, and Sora have emerged. These innovations have the potential to transform lifestyles and boost productivity [1]. Integrating AI with communications systems will help build intelligent 6G wireless networks, advance society toward inclusive intelligence, and achieve the 6G vision of "the Intelligent Network of Everything" [2]. However, training and optimizing AI models require vast volumes of high-quality data, which is challenging to collect, especially massive and diverse wireless air interface data [3]. Due to the sensitivity of air interface channels to propagation environments, even slight environmental changes can cause significant fluctuations. To ensure an AI model has sufficient generalization capability, it is essential to collect air interface data across diverse physical environments, which incurs substantial human resources and material costs. To address this challenge, a novel approach has emerged that utilizes generative models to expand datasets [4]. By incorporating high-quality synthetic data into the original datasets, this approach enhances model convergence and generalization.

Training a powerful generative model requires massive volumes of data [5]. However, retrieving this data, which is typically collected and stored locally, from numerous edge nodes and sending it back to a centralized node for processing results in significant network load and resource overheads [6]. To reduce system overheads, a typical approach is to utilize a federated learning (FL) framework for distributed training of generative models [7]. This method leverages idle edge node computing power to reduce communication overheads and shorten the training time required for model convergence [8]. Additionally, during FL-based training, data remains stored locally, ensuring privacy [9].

This paper explores the method of training distributed generative adversarial network (GAN) models [10] using an FL architecture based on air interface channel data. It also compares the data generation effectiveness and performance differences between federally trained GAN models and traditional centrally trained GAN models.

2 FL-based GAN Model Training

2.1 Concepts

Figure 1 illustrates the basic architecture of a GAN, which comprises two independent neural networks: the generator G and discriminator D . The GAN framework does not impose any strict requirements on the exact structure of these two models, and they can be configured as multilayer perceptrons (MLPs), convolutional neural networks (CNNs), or transformers, etc. The main concept of GAN involves introducing a supervisory model (known as the discriminator) along with the traditional standalone generator to determine whether generated samples conform to the distribution of training samples. Through adversarial training, GANs gradually reach a Nash equilibrium between the two models.

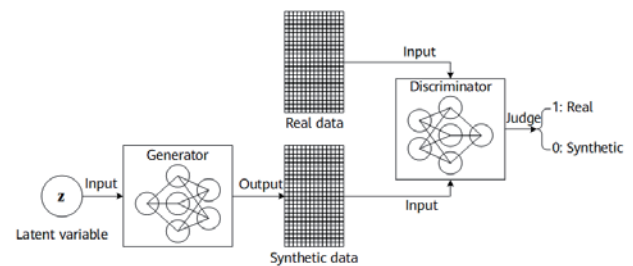


Figure 1 Basic GAN framework [11]

The generator's objective is to learn the mappings from latent variables z in a low-dimensional space to real channel data in a high-dimensional space. By inputting randomly sampled z into the generator, we can produce new synthetic data, denoted as $G(z; \theta_g)$, where θ_g represents the parameters of the generator. The discriminator's objective is to determine whether the input data is real or synthetic. The generator's output layer typically connects to a Sigmoid function, limiting its output value $D(x; \theta_d)$ to $[0, 1]$, where θ_d represents the parameters of the discriminator, and x denotes the input. This value represents the probability of the discriminator identifying the input as real data. The training objective of the generator is to obscure its output from the discriminator, whereas the discriminator's objective is to distinguish real and synthetic data, creating a minimax game. The overall optimization objective of the GAN model can be formulated as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{\text{data}}(x)} \{\log D(x; \theta_d)\} + \mathbb{E}_{z \sim P_g(z)} \{\log(1 - D(G(z)))\} \quad (1)$$

When both models reach their optimal state, the distribution P_g of the generated data approaches the distribution P_{data} of the training data, and the discriminator's probability of identifying real data is approximately 0.5.

2.2 Framework

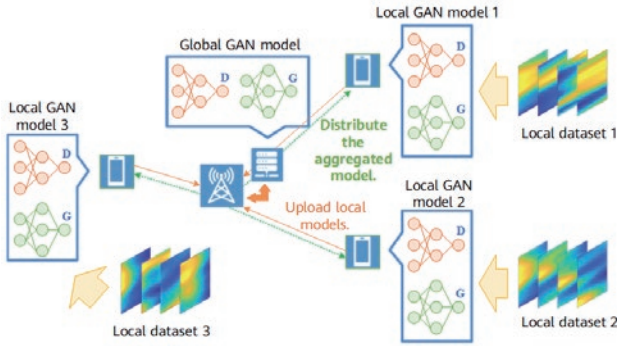


Figure 2 FL-based GAN training framework

As illustrated in Figure 2, consider a scenario where N users are involved in the FL of a global GAN model, with each user storing a different channel dataset denoted as D_n , where $n = 1, \dots, N$. Each dataset contains $|D_n|$ channel samples, following the distribution $P_n(x)$. Note that this paper exclusively addresses cases where all users' datasets conform to an independent and identically distributed (i.i.d.) pattern. To complete FL, multiple rounds of model aggregation are required. Consider the t th round of model aggregation: each user performs M rounds of local training on the generator and discriminator using their local datasets. The parameters of the two models after tM iterations of training are denoted as $\theta_n^g(tM)$ and $\theta_n^d(tM)$. These parameters are then uploaded to a centralized node, which aggregates the parameters from multiple users and returns the aggregated result. The aggregation rule calculates a weighted average of the model parameters, prioritizing users with larger data volumes [12]. The aggregated result is denoted as:

$$\begin{aligned} \theta_{FL}^d(t) &= \frac{\sum_{n=1}^N |D_n| \theta_n^d(tM)}{\sum_{n=1}^N |D_n|}, \\ \theta_{FL}^g(t) &= \frac{\sum_{n=1}^N |D_n| \theta_n^g(tM)}{\sum_{n=1}^N |D_n|}, \end{aligned} \quad (2)$$

where $\theta_{FL}^d(t)$ and $\theta_{FL}^g(t)$ represent the aggregated results of the generator and discriminator, respectively, in the t th round of training. After aggregation, the centralized node

distributes the updated global model parameters $\theta_{FL}^d(t)$ and $\theta_{FL}^g(t)$ to all participating users, who then update their local model parameters $\theta_n^d(t)$ and $\theta_n^g(t)$ accordingly. This completes a round of federated training. The process is repeated multiple times until the model converges.

2.3 Training Method

After T rounds of aggregation, the model is close to convergence. For details on the complete training process, see Algorithm 1.

Algorithm 1: FL-based GAN training process

Initialization: This step involves the following parameters: number of rounds of federated training (T), number of local training iterations (M) of each user in each round, generator parameter $\theta_n^g(0)$ and discriminator parameter $\theta_n^d(0)$ initialized by each user, data volume $|D_n|$ of each user, and convergence threshold δ .

Loop: For the t th round of federated training ($1 \leq t \leq T$),

1. Users complete M rounds of local training using Algorithm 1 from [11] to obtain the trained generator and discriminator parameters $\theta_n^d(tM)$ and $\theta_n^g(tM)$.
2. The users upload $\theta_n^d(tM)$ and $\theta_n^g(tM)$ to the centralized node, which then updates the global discriminator and generator parameters to $\theta_{FL}^d(t)$ and $\theta_{FL}^g(t)$, respectively, based on Equation 2, and distributes them back to each user.
3. The users update their local discriminator and generator parameters to the global model parameters.

Termination: This step is triggered when $t > T$ or $\|\theta_{FL}^d(t) - \theta_{FL}^d(t-1)\| \leq \delta$ and $\|\theta_{FL}^g(t) - \theta_{FL}^g(t-1)\| \leq \delta$.

Output: The trained global generator $\theta_{FL}^{g,*}$ and discriminator $\theta_{FL}^{d,*}$ are generated.

3 Simulation and Analysis

3.1 Air Interface Channel Setup

This experiment is conducted based on the air interface channel dataset following the TDL-C channel model [13] from 3GPP TS 38.901, which simulates a Rayleigh channel for macro base stations in urban areas under non-line-of-sight (NLOS) conditions. This experiment also simulates and evaluates channel data in a high-speed environment

at 300 km/h. Other simulation parameter settings include 48 subcarriers \times 14 symbols per frame, a carrier frequency of 3.5 GHz, and subcarrier spacing of 30 kHz. Four users participate in federated training, with each user's local dataset containing 2,500 channel data samples and following the i.i.d. pattern. For comparison, centralized training is performed using a dataset at the centralized node, which includes local data from all users and has a total of 10,000 channel data samples.

3.2 GAN Model Setup

The GAN model used in this experiment follows the structure outlined in [11]. Both the generator and discriminator are fully connected networks with five linear layers, as shown in Figure 3.

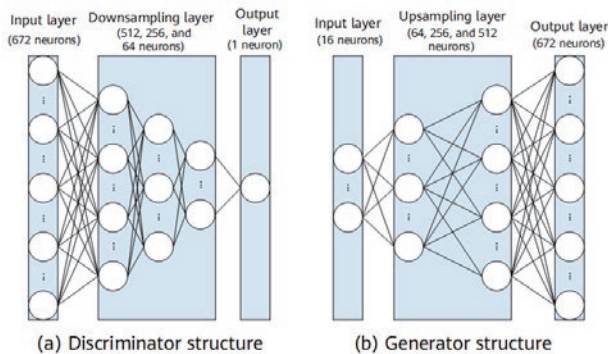


Figure 3 GAN model structure [11]

For the discriminator, we first flatten the input channel data into vectors. Because the channel values are complex, their real and imaginary parts are treated as separate data samples for further processing. Therefore, the designed input layer has 672 neurons. The hidden layer consists of three downsampling linear layers, with 512, 256, and 64 neurons, respectively. The outputs of the first four layers are processed with the LeakyReLU and LayerNorm functions. The output layer has one neuron and leverages a Sigmoid function to determine the probability that the samples input to the discriminator are real data.

For the generator, we design the input latent variable z as a 16-dimensional random vector following a standard Gaussian distribution. The hidden layer consists of three upsampling linear layers, mirroring the neuron structure of the discriminator's downsampling layers. The output layer has 672 neurons to reconstruct the complete channel matrix. It uses the Tanh function for activation.

3.3 Simulation and Evaluation

Figure 4 shows the simulation results of air interface channel strength, with 20 rounds of federated training and 10 local training sessions per round. The horizontal axis represents the number of symbols, while the vertical axis represents the number of subcarriers. Figure 4a shows the effect of using only local training instead of federated training for channel generation, whereas Figure 4b illustrates the effect of a federally trained global model. These figures illustrate that locally trained models perform poorly and produce highly blurry channel images for a given number of training sessions. In contrast, federally trained models show significant improvement in output performance and reduction in image noise. This suggests that the federated training architecture effectively leverages data from multiple nodes, leading to better model training outcomes.

Figure 5 shows the simulation results after 50 federated training rounds and 20 local training sessions per round. As can be seen from Figure 5a, despite a significant

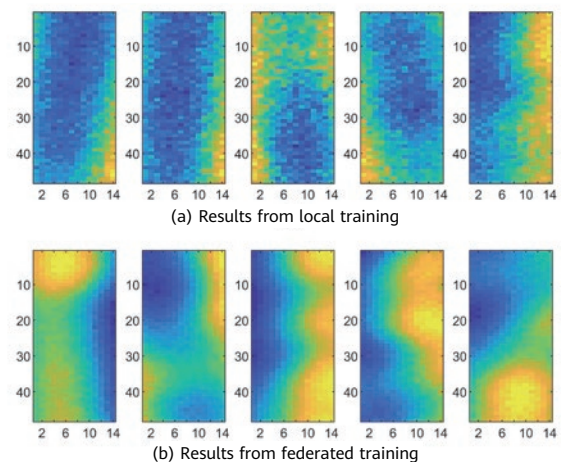


Figure 4 GAN model outputs after 200 rounds of local and federated training

improvement compared to Figure 4a, the locally trained model still produces low-quality channel samples with considerable noise due to insufficient training data. In contrast, as depicted in Figure 5b, the federally trained model generates near-real channel samples with minimal noise, demonstrating performance that significantly surpasses the locally trained model and is on par with the centrally trained model using the same dataset.

To further evaluate the performance of the FL-based GAN model, we tested the model in a channel estimation task using the normalized mean square error (MSE) between

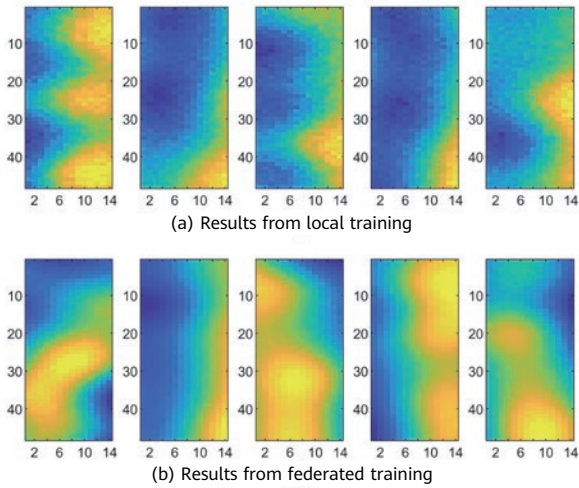


Figure 5 GAN model outputs after 1,000 rounds of local and federated training

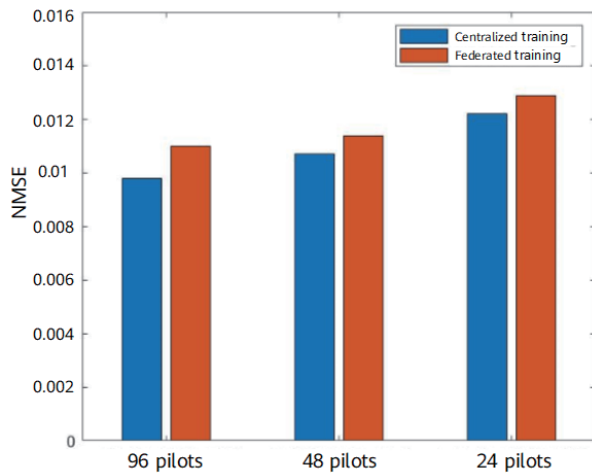


Figure 6 Channel estimation performance comparison between a centrally trained GAN model and a federally trained GAN model at an SNR of 15 dB

estimated and true values. Figure 6 compares the channel estimation results for federally and centrally trained GAN models, following the method outlined in Algorithm 2 in [11]. The results demonstrate that for any given pilot quantity, the channel estimation performance of the federally trained model is comparable to that of the centrally trained model. Centralized training uses serial computing, whereas federated training employs parallel computing. Additionally, the latency from uploading and distributing model parameters is negligible since fewer model parameters are involved in the experiment, and the transmission latency is much lower than model training latency. Therefore, when computing power across nodes is comparable, the latency in federated training decreases as the number of participants increases, resulting in a shorter training duration.

4 Conclusion

This paper explores an FL-based distributed GAN training method for air interface channel data. Simulation results demonstrate that the data generation quality of the FL-based GAN model is comparable to that of the traditional centrally trained model, offering new insights for future air interface data collection.

This paper focuses solely on generating air interface channel data in sub-6 GHz bands. The effectiveness of generating such data in complex scenarios, such as mmWave and higher frequency bands, is yet to be verified. Additionally, the impact of differences between actual and simulated channel data on data generation needs to be further explored. For distributed model training architectures like FL, challenges, such as non-i.i.d user data, heterogeneous models, and asynchronous model aggregation, still require thorough investigation. The effective volume of data that the GAN model can provide also requires further testing and verification.

References

- [1] D. Zhang, Z. P. Bhat, K.-H. Lai, *et al.*, "Data-centric artificial intelligence: A survey," arXiv preprint, arXiv: 2303.10158, 2023.
- [2] IMT-2030(6G) Promotion Group, "White paper on 6G vision and candidate technologies," June 2021.
- [3] M. Xu, Y. Li, M. Li, *et al.*, "A denoising diffusion probabilistic model based data augmentation method for wireless channel," in Proc. International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, Nov. 2023.
- [4] X. Li, K. Wang, X. Gu, *et al.*, "Paralleleye pipeline: An effective method to synthesize images for improving the visual intelligence of intelligent vehicles," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 53, no. 9, pp. 5545–5556, May. 2023.
- [5] B. Shaker, G. P. R. Papini, M. Saveriano, and K.-Y. Liang, "Generating synthetic vehicle data using decentralized generative adversarial networks," IEEE Access, Jul. 2024.
- [6] Z. Zhao, S. Bu, T. Zhao, *et al.*, "On the design of computation offloading in fog radio access networks," IEEE Transactions on Vehicular Technology, vol. 68, no. 7, pp. 7136–7149, Jun. 2019.
- [7] C. Hardy, E. L. Merrer, and B. Sericola, "MD-GAN: Multi-discriminator generative adversarial networks for distributed datasets," in Proc. IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil, May. 2019.
- [8] Zhiqin Wang, Jiamo Jiang, Peixi Liu, *et al.*, "New design paradigm for federated edge learning towards 6G: Task-oriented resource management strategies[J]," *Journal on Communications*, 2022.
- [9] C. Feng, Z. Zhao, Y. Wang, *et al.*, "On the design of federated learning in the mobile edge computing systems," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5902–5916, Jun. 2021.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.
- [11] Y. Du, Y. Li, M. Xu, *et al.*, "A joint channel estimation and compression method based on GAN in 6G communication systems," *Applied Sciences*, vol. 13, no.4, 2023.
- [12] H. B. McMahan, E. Moore, D. Ramage, *et al.*, "Communication-efficient learning of deep networks from decentralized data," in Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, USA, Apr. 2017.
- [13] 3GPP, "3GPP TR 38.901 v16.1.0 3rd generation partnership study on channel model for frequencies from 0.5 to 100 GHz (Rel 14)," 2020.



Service Quality Assurance for 6G Networks with Native AI

Guangyi Liu¹, Kaiyue Wang¹, Juan Deng¹, Jiajun Wu¹, Huanran Hu², Guanchen Lin²

¹ China Mobile Research Institute (Beijing)

² Beijing University of Posts and Telecommunications

Abstract

In its Recommendation, the International Telecommunication Union (ITU) defines "AI and Communication" as one of the six typical usage scenarios of 6G, with native intelligence being crucial in 6G. The 6G native AI network is designed to improve the performance of 6G air interfaces, achieve high-level network autonomy, and deliver high-performance AI services. To meet various requirements of diversified AI services and guarantee quality of artificial intelligence service (QoAIS) in differentiated scenarios, industry research is focusing on designing a unified network QoAIS mechanism. In this paper, we focus on QoAIS assurance for 6G networks from the following perspectives: architecture design, indicator systems, protocols and processes, and assurance technologies. We first propose the 6G network QoAIS architecture featuring hierarchical management and control. Then, we discuss the design of indicator systems for QoAIS and AI foundation models, and related protocols for convergence of communications and computing. We also introduce QoAIS assurance technologies based on the MADDPG-Adv algorithm. These findings help achieve AI full lifecycle management and on-demand scheduling of four AI elements within networks. At the end of this paper, we conclude by demonstrating the QoS assurance effect in AI inference scenarios through simulation, using two main indicators, namely, accuracy and delay. Our reinforcement learning algorithm MADDPG-Adv can improve the performance by 6% and achieve a 23% increase in the number of UE tasks that satisfy the QoAIS requirements when compared to the MADDPG algorithm.

Keywords

6G, native AI, QoAIS assurance

1 Introduction

The huge application scope of artificial intelligence (AI) in all walks of life is posing new requirements for future 6G networks. To meet some of these requirements, native AI or network AI needs to be developed to use AI as a service (AlaaS) on 6G networks [1]. This is necessary in order to deliver pervasive intelligent services for both networks and third-party users. 6G networks with native AI will support two main types of scenarios, "AI for Net" (empowering networks with AI capabilities) and "Net for AI" (enabling AI with network capabilities). In terms of AI for Net scenarios, a key technology trend lies in convergence of AI and air interfaces. Native AI technologies are necessary to improve KPIs like bandwidth, throughput, and capacity of 6G systems. And in terms of Net for AI, mobile communications networks serve as fundamental platforms for making AI ubiquitous and inclusive. The edges of 6G networks can be utilized to enable high-value AI scenarios, use cases, and services. To advance AI for Net, standardization organizations such as the ITU and 3GPP have been engaged in research and standardization of typical use cases and key technologies of air interface AI. Today, AI technologies hold immense potential in supporting applications such as modeling and estimation of complex unknown environments, intelligent signal modulation and coding, intelligent network scheduling, and intelligent network optimization and deployment. This creates significant value for related research and evolution of 6G technologies [2]. To boost Net for AI, China Mobile works with Huawei and other partners to develop a task-oriented native-AI RAN architecture, providing insights for the 6G native AI architecture [3]. From the perspective of basic theoretical models for distributed AI, key algorithms are designed and simulation experiments are conducted in [4]. This effectively demonstrates the studies on native AI algorithms and drives research on basic theoretical algorithms for network foundation models. [5] puts forward a task scheduling and resource allocation scheme for AI training services on native AI wireless networks, allowing for flexible computing server selection, data quality adjustment, and resource allocation to meet the quality of service (QoS) requirements of AI tasks. In [6], the authors design an AI-native network slicing architecture for 6G networks to enable the synergy of AI and network slicing, thereby facilitating intelligent network management and supporting emerging AI services. And aiming to develop 6G native AI applications, [7] presents a novel federated transfer learning framework, in which models are locally trained on distributed edge agents for global aggregation on edge base stations. This framework

can significantly improve accuracy of models and reduce energy consumption. Undoubtedly, the integration of 6G and AI opens up a variety of new opportunities. AI and Communication, one of the typical usage scenarios of 6G, is expected to help achieve the overall vision of 6G: Pervasive Intelligence and Digital Twin.

6G network AI provides intelligent services for high-level network autonomy and various types of users. The AI application scenarios of network autonomy can be typically classified into three types: O&M intelligence, network intelligence, and NE intelligence. By leveraging advantages in terms of real-time performance, privacy, mobility, and edge-device synergy, 6G network AI can contribute to external high-value scenarios including mobile robots, Internet of Vehicles (IoV), and extended reality (XR). Typical use cases include mobile household robots, factory transportation robots, AI-assisted autonomous driving, and AR navigation. (1) According to GII data, the mobile robot market is expected to hit CNY180 billion globally by 2027. Different types of robots meet the different requirements of diverse industries and tasks [8]. In terms of KPIs, some remotely controlled robots impose stricter requirements. For example, helicopters and humanoid robots require a delay shorter than 5 ms, necessitating device-edge synergy design. 6G network AI can empower lower delay and prominent edge-device synergy for these robots [9]. (2) IDC forecasts that global autonomous driving vehicles will see significant growth, reaching up to 89.3 million by 2026, and IoV is expected to play a pivotal role in future economic development [10]. IoV requires a one-way delay of 10 ms. Furthermore, image recognition requires a delay shorter than 5 ms, while that of fault detection, security warning, and video identification should range from 500 ms to 1s. All these usage scenarios demand that 6G networks deliver high real-time performance. (3) XR techniques can be applied at scale to industrial applications, healthcare sector, education, military drills, e-commerce, and many other scenarios. Network AI, which is closer to UEs than cloud AI, accelerates the response to XR intelligent service requests by reducing signaling interaction on non-access networks. Additionally, 6G networks can provide stronger computing power when compared to on-device AI, enabling AI inference and training services used by XR while also reducing the delay of AI computing. Another benefit of 6G network AI is that it has a smaller data transmission scope than cloud AI. This helps to protect data privacy, particularly for user information on the device side and sensitive data in industrial scenarios that are vulnerable to eavesdropping and pollution. When UEs send requests during movement,

6G network AI can make more flexible response to AI services, like model inference and real-time download.

Diverse scenarios and use cases pose different requirements for AI service types and QoAIS. Consequently, 6G network AI needs to enable AI service types on demand and ensure QoAIS. Typically, there are four types of AI services: AI inference, AI training, model optimization, and data processing and preprocessing. Among these types, AI inference can be applied to network AI scenarios that are highly sensitive to delay. For example, in scenarios such as voice recognition and real-time translation, AI models can be used to receive input data and immediately generate output data. This type of AI service can also be adopted for non-real-time scenarios, such as large-scale image classification and video analysis. AI training covers supervised, unsupervised, semi-supervised, and reinforcement learning training. Model optimization involves model pruning, quantization, knowledge distillation, and structure optimization. And data processing and preprocessing refers to data cleaning and data augmentation. Specifically, data cleaning services aim to remove or correct noise and errors in data to ensure data quality, whereas data augmentation services generate new training samples to extend datasets for images and texts [11].

In order to deliver QoAIS assurance, the design of 6G networks must consider the following four aspects: architecture, indicator systems, protocols and processes, and assurance technologies. (1) Architecture: The traditional communication architecture provides session-centric QoS assurance, resulting in a lack of closed-loop feedback. When handling AI services, the architecture needs to support in-depth integration of multiple resource elements including connections, computing, data, and algorithms. This makes it necessary to design a novel QoS architecture. (2) Indicator systems: Traditional communication QoS only covers indicators such as delay, reliability, throughput, and priority. However, AI services need to orchestrate multi-dimensional elements including connections, computing, data, and models. In terms of accurately reflecting the performance, the 5G QoS indicator systems fall short [12]. (3) Protocols and processes: AI service application, AI task control, and AI resource control require E2E protocol and process design. Additionally, corresponding protocols and processes need to be designed for new AI bearers together with transmission and mapping of QoS indicators. Because there are some problems in the existing QoS mechanism, such as coarse service differentiation granularity and long optimization period, adapting it to task-oriented AI service mode is difficult. (4) Assurance technologies: 5G QoS

assurance is performed by using sessions and connections. Core network NEs obtain the QoS flows and QoS indicators corresponding to data packets through 5-tuple mapping, and then send them to the RAN. The RAN is responsible for mapping QoS flows and data radio bearers (DRBs) and for scheduling air interface resources. This mechanism ensures communication QoS of different services and data packets [13]. To ensure QoAIS, QoS requirements for all connection, computing, data, and model resources need to be fulfilled. Additionally, three layers of QoS need to be mapped in an appropriate manner, driving the need for more advanced, complex assurance technologies. However, existing solutions fall short in ensuring QoAIS on 6G networks. 6G faces challenges like differentiated QoS requirements for various AI service scenarios and allocation of multiple resources (communication, computing, data, and model elements). To address these challenges, a novel QoAIS assurance solution needs to be developed for 6G with consideration given to the network architecture, indicator systems, protocols and processes, and assurance technologies.

2 QoAIS Architecture Design

A 6G network QoAIS architecture features a hierarchical management and control design. It performs dynamic task-oriented orchestration to achieve AI full lifecycle management and on-demand scheduling of communication, computing, data, and model resources (referred to as four element resources for short) within networks.

Because AI integrates computing, data, and model elements, a 6G network QoAIS architecture needs to be equipped with computing, data, and model functions beyond existing communication network functions. Figure 1 shows a logical function architecture of native AI in 6G. The top layer provides the management and orchestration functions, which are responsible for parsing and mapping AI service requirements and for orchestrating the AI service function chain to meet the QoS requirements of the service layer. However, because this function cannot directly manage UEs, the four elements of AI cannot be coordinated to manage, control, and guarantee task QoS. Additionally, high signaling delay at the top layer means that task management cannot be performed in a prompt manner, making it exceedingly difficult to meet strict task QoS assurance requirements. In order to manage the lifecycle of AI tasks and determine task controllers, the AI task control function is introduced to the control plane. (1) The communication control function on this plane incorporates data bearer setup, model transmission bearing, and other control functions

in addition to existing communication functions. (2) The computing control function enables control over computing resource sensing, compute node selection, and computing resource allocation. (3) The data control function provides data management functions, such as data collection node selection and data collection configuration. (4) The AI model control function controls and manages models, enabling model registration and selection. This function can be used to deploy, start, delete, modify, and monitor tasks based on the QoS requirements of AI tasks, including scheduling communication, computing, data, and model element resources, making it possible to ensure QoS in the task deployment phase. On the user plane, the task execution function is designed to execute AI tasks like AI training, inference, and verification.

To ensure the quality of AI services in 6G, there is a need to develop a QoS assurance mechanism with closed-

loop feedback. With this mechanism, the task execution layer monitors and optimizes four element resources, meeting resource QoS requirements. At the same time, the task control layer ensures the fulfillment of task QoS requirements. This layer can flexibly adjust the resource configuration and mapping between task QoS and resource QoS in cases where the resource QoS requirements cannot be met. The orchestration management layer, which is responsible for guaranteeing QoS, optimizes the mapping between AI services and AI tasks if the task control layer cannot deliver task QoS assurance. This type of three-layer QoAIS mechanism with closed-loop feedback implements on-demand collaborative mapping from AI service requirements to underlying multi-dimensional network resources, yielding differentiated QoAIS assurance.

Figure 2 shows the logical potential deployment locations of native AI functions in 6G. The AI service orchestration

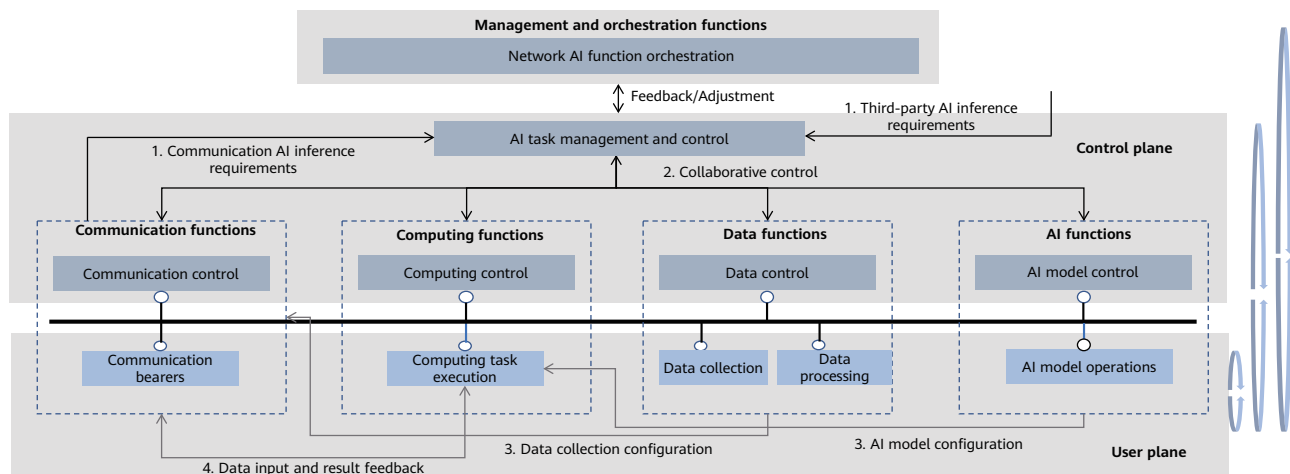


Figure 1 Logical function architecture of native AI in 6G

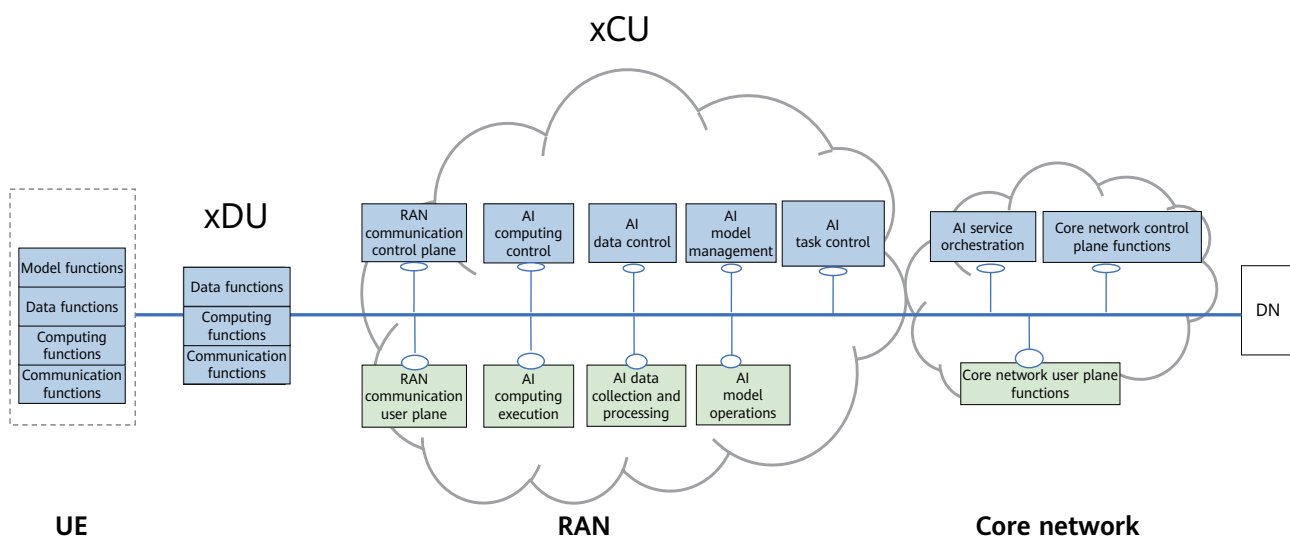


Figure 2 Logical deployment of functions of native AI in 6G

function can be deployed on the core network to receive and parse AI service requests initiated by UEs, and decompose AI services into AI tasks. This function can also orchestrate AI tasks on demand and deploy such tasks to the AI task management function, which deploys functions related to four elements and configures resources based on QoS requirements of AI tasks. This process maps task QoS to resource QoS. Furthermore, to meet real-time performance requirements of certain AI services, functions related to four elements can be deployed on the RAN. And in order to complete AI tasks, the control plane and user plane functions collaborate in terms of four elements.

3 QoAIS Indicator Systems

A QoAIS indicator system is used by the network to ensure the quality and effect of AI services. It can reflect the AI advantages of 6G networks and help strike a balance between users' intelligent requirements, fairness of network use, and limited resources. QoAIS indicator systems aim to guarantee AI service QoS, AI task QoS, and AI resource QoS, with three layers of indicators being mapped. (1) AI services provide AI technologies, services, or three AI elements (computing, data, and model elements) to the served parties on demand. AI service QoS involves the quality indicators of AI services in terms of performance, availability, reliability, etc. An AI service can be decomposed into one or more AI workflows, which can be further decomposed into one or more tasks. (2) An AI task involves the collaboration of computing, algorithms, connections, and data to achieve a specific goal of an AI service. AI task QoS focuses on priorities of task orchestration in the deployment phase and four-element assurance in the execution phase. (3) AI

resources mainly refer to four element resources involved in AI task implementation, including computing resources (e.g., CPUs and GPUs) and physical resources (e.g., air interface time-frequency resources). In the case of AI resource QoS, a range of indicators are used to measure the quality of four element resources on networks. For instance, computing QoS centers on computing delay, reliability, and concurrency. Algorithm QoS involves training convergence speed, generalization, and explainability. Data QoS covers indicators measuring collinearity as well as security and privacy levels. Connection QoS indicators measure the priorities, delay, and packet loss rate. Table 1 shows an example of the layer-based mapping of QoAIS indicators for an AI inference task.

At present, third-party foundation models are developing rapidly. To empower foundation models through networks, QoAIS indicators need to be designed for AI services based on the characteristics of foundation models. Typically, foundation model services empowered by networks consist of four phases: (1) In the model pre-training phase, all parameters need to be updated, and a significant amount of training data is required, resulting in the need for high computing power. (2) In the model fine-tuning phase, partial parameters can be updated, and a small amount of training data is required, meaning there is lower demand for computing power than during model pre-training. (3) In the model deployment phase, only forward propagation is performed, eliminating the need for parameter update or data training. Therefore, this phase poses the lowest computing power requirement among the four phases. (4) In the model optimization phase, the computing power demand is also lower than that during model pre-training. Centering on the model inference, fine-tuning, and optimization services, QoAIS indicators are simplified from different evaluation dimensions. Table

Table 1 Layer-based mapping of QoAIS indicators

Category		Indicator
AI service		Delay, power consumption, transmission rate, jitter, etc.
AI inference task		Inference speed and accuracy
AI resource	Data resource	Feature redundancy, integrity, data accuracy, and data preparation duration
	Model resource	Performance indicator limit, training duration, convergence, and optimization target matching degree
	Computing resource	Computing accuracy, duration, and efficiency
	Communication resource	Bandwidth and jitter, delay and jitter, bit error rate and jitter, reliability, etc.

2 provides the indicator systems for network-empowered foundation models.

4 QoAIS Protocols and Processes

To guarantee QoAIS of 6G networks, end-to-end protocols and processes need to be specifically designed. QoAIS is an important input for the management and orchestration as well as task control functions. Based on QoAIS, the management and orchestration function converts user requirements into AI service capability requirements that can be understood by networks. These converted requirements are called AI service QoS, which are further decomposed into task QoS as input of the control plane. Then, the

control plane maps task QoS to resource QoS requirements for connections, computing, data, and algorithms. By monitoring and optimizing allocation of these four element resources in real time, the converged control and bearer protocols for heterogeneous, multi-dimensional AI resources are accordingly designed. This satisfies the high QoS requirements for AI task transmission and execution. Figure 3 depicts the end-to-end QoAIS transmission processes.

Systematic protocols can be further developed for the function architecture of native AI in 6G. Specifically, in protocols for convergence of communications and computing, illustrated in Figure 3 in the blue dashed box, we introduce computing execution layers for computing processing of AI tasks, and computing and connection

Table 2 QoAIS indicator systems for network-empowered foundation models

Service Type	Evaluation Dimension	QoAIS Indicator
Fine-tuning or optimization	Performance	Performance indicator limit, fine-tuning duration, explainability, consistency between the loss function and the optimization target, and fairness
	Overhead	Storage overhead, computing overhead, transmission overhead, and energy consumption
	Security	Storage security, computing security, and transmission security
	Privacy	Data privacy level and algorithm privacy level
	Autonomy	Full autonomy, partial manual control, or full manual control
Inference	Performance	Inference speed, concurrency, inference error, inference accuracy, etc.
	Autonomy	Model controllability, such as automatic scheduling of inference tasks and automatic allocation of inference resources

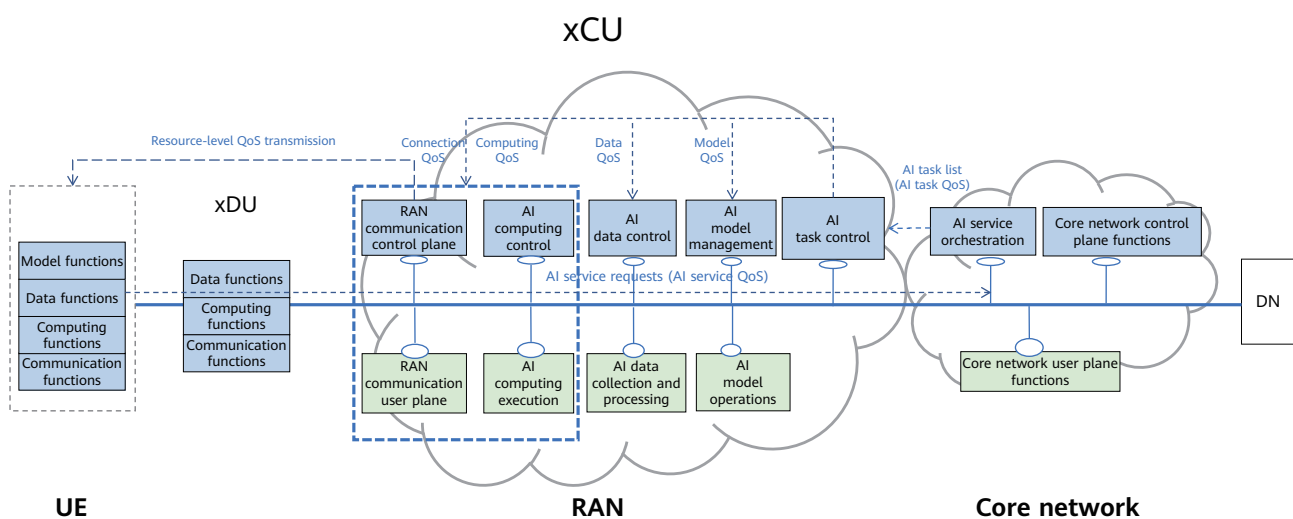


Figure 3 End-to-end QoAIS transmission processes

bearers for transmitting computing data. As shown in Figure 4, CTAP layers are leveraged on the basis of existing user-plane protocol stacks of 5G air interfaces, implementing QoS mapping and update for AI computing tasks. A network-side CTAP entity adds a header index to a data packet of a computing task according to the received communication QoS and computing QoS requirements of this task. Data packets of computing tasks with the same communication QoS requirements are filtered and mapped to one QoS flow, which is then sent to a lower-layer SDAP entity. At the SDAP layer, the QoS flow is mapped to a corresponding radio bearer based on the QoS indication information, with an index added to the SDAP packet header. Then, through processing and transmission (using protocols such as PDCP, RLC, MAC, and PHY), various types of computing data are mapped to corresponding computing and connection bearers. A newly added computing execution layer executes computing tasks on the network or UE side. It also receives computing data packets and computing QoS requirements from the upper-layer and maps computing tasks with similar computing QoS requirements to the same computing execution function. Then, this layer transmits the computing result to the next layer. Mapping from AI computing execution QoS and AI connection QoS to computing resources and connection bearers is achieved by introducing CTAP. The entire processes fulfill the QoS requirements of computing transmission and computing execution.

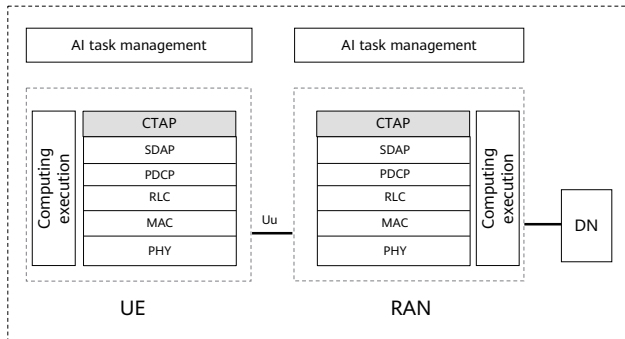


Figure 4 QoAIS protocol design

5 QoAIS Assurance Technologies

QoAIS assurance technologies are expected to meet user requirements for AI services and improve the overall system performance by scheduling AI services (like AI inference and training) and allocating communication, computing, and AI model resources. Under the QoAIS architecture, converged

control over communication, computing, and models is enabled at the task control layer to ensure the quality of AI services mainly in terms of the accuracy and delay indicators.

In this paper, we propose QoAIS assurance technologies based on cloud-edge-device synergy for AI inference services. (1) The cloud can provide AI functions and computing resources, and store full AI models. Each subsequent version of an AI model results in a more complex model structure and improved accuracy. Additionally, powerful cloud computing resources can minimize the computing delay. (2) Base stations enable AI task control, communication and computing, and AI model management. However, because these base stations have only a limited cache, they can cache only partial AI service models. (3) On the device side with certain computing resources, UEs can process certain layers of neural networks, or achieve joint inference through collaboration with base stations or the cloud.

Figure 5 illustrates the process of AI inference powered by cloud-edge-device synergy. Base stations forming agents through collaboration with each other cache multiple AI services from the cloud and allocate bandwidth and computing resources to UEs. In addition, UEs apply for AI services and report the QoAIS requirements (in terms of model accuracy and delay) to the network. Base stations leverage the AI task control function and generate the QoAIS assurance policies for allocating air interface bandwidth and computing resources and caching AI models.

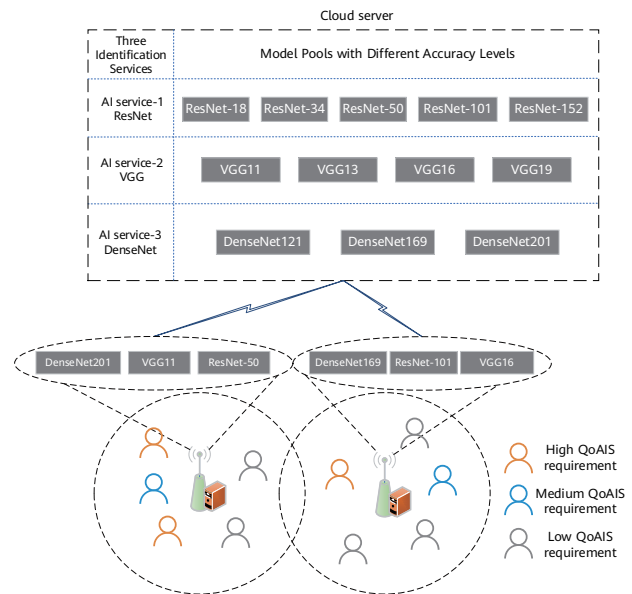


Figure 5 AI inference powered by cloud-edge-device synergy

To evaluate the quality of AI inference services, indicators covering aspects such as accuracy and delay can be used. The AI service accuracy score $Q_p^{u_i}$ can be calculated using the following formula:

$$Q_p^{u_i} = \begin{cases} v_b \\ v_b - \Delta \times v_c \end{cases} \quad (1)$$

When the allocated AI models meet or exceed user requirements, the accuracy score equals the benchmark score v_b . If the models fall short in meeting these requirements, the accuracy score is obtained by subtracting the penalty score from the benchmark score. The penalty score is the product of the penalty factor v_c and Δ , representing the difference between the required accuracy and allocated accuracy of AI models.

The AI service delay score $Q_d^{u_i}$ is calculated as follows:

$$Q_d^{u_i} = \begin{cases} v_d \\ 0 \end{cases} \quad (2)$$

where the delay score is denoted by v_d when the delay is within the required range, and equals 0 when it is beyond the range.

To continuously ensure QoAIS with dynamic, complex wireless network status, the assurance solution optimizes the service accuracy and delay of AI tasks by adopting a reinforcement learning algorithm. In this case, the algorithm state, action, and reward settings of the Markov process are as follows. (Agents 1 and 2 represent base stations 1 and 2 in Figure 5, respectively.)

- State space — Agent 1 state: [UE accessing the base station, UE applying for a service index] \times UE quantity. Agent 2 state: [UE accessing the base station, UE channel status, Base station that carries UE computing, Service index of UE computing] \times UE quantity.
- Action space — Agent 1 action: [Three AI services stored on the base station] \times Base station quantity. Agent 2 action: [Computing and bandwidth resources allocated to each UE] \times UE quantity.
- Reward functions are closely related to the optimization objectives and defined as follows:

$$R_1 = \left(\sum_{u_i \in U} (Q_p^{u_i} - q_1 - q_2) - \sum_{j=0}^1 q_3 \right) \times \rho_1 \quad (3)$$

$$R_2 = \sum_{u_i \in U} Q_d^{u_i} \times \rho_2 \quad (4)$$

where ρ_1 and ρ_2 represent the scale factors. Because the cache policy influences locations of UE computing to some extent, q_1 and q_2 are the penalty factors introduced for performance loss between two base stations and between base stations and the cloud, respectively. q_3 denotes the penalty factor introduced for repeated caching by base stations.

To address the issue that correlation between resource allocation policies affects the resource allocation algorithm, we optimize the existing MADDPG algorithm and propose our MADDPG-Adv algorithm. This algorithm breaks down a problem of collaboratively allocating three (communication, computing, and AI model) resources into two sub-problems: a global policy for obtaining the cache and a global policy for obtaining the communication and computing resources. It also enhances information transmission between different agents in MADDPG. As presented in Figure 6, agents can resolve various problems on corresponding base stations. In addition, the locations of cached AI models (on the cloud or base stations) directly impact the allocation policy of communication and computing resources. However, during status observation, cache information of AI models cannot be directly obtained until the actions are complete. Consequently, we have developed an additional information transmission mechanism in our solution. After the actions are complete, the observed network status information and the transmitted cache information are integrated for collaboratively generating the policies to asynchronously execute the algorithm operations.

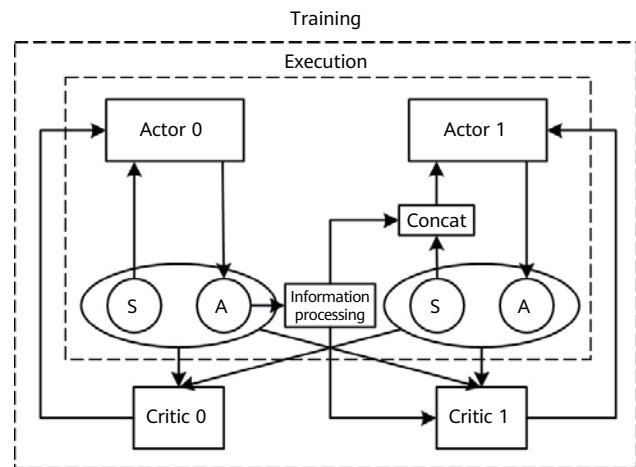


Figure 6 MADDPG-Adv architecture

6 Simulation Solution and Result

To simulate AI inference scenarios, we use a set of base stations and eight UEs. 12 types of identification AI models are stored on the cloud, with each type of AI model having three to five versions. During simulation verification, we evaluate whether the AI service meets the QoAIS requirements from two dimensions: accuracy and delay of the AI service. The optimization objectives involve maximizing the accuracy and delay scores of the AI service.

In this paper, we adopt the orthogonal frequency division multiple access (OFDMA) technology for system modeling by referring to 3GPP TR.38.901 [14]. The uplink transmission rate e_{u_i} of a UE is modeled as follows:

$$e_{u_i} = B_{u_i} \log_2 \left(1 + \frac{\beta_{u_i} h_{u_i} p_{u_i}}{\sigma B_{u_i}} \right) \quad (5)$$

where β_{u_i} is the channel loss between the UE and base station, h_{u_i} is the antenna gains, p_{u_i} is the uplink transmission power of the UE, and σ is the noise power spectral density.

The transmission delay T_{t-u_i} of a UE can be expressed as:

$$T_{t-u_i} = \begin{cases} \frac{D_{air}}{r_{u_i}} \\ \frac{D_{air}}{r_{u_i}} + T_c \\ \frac{D_{air}}{r_{u_i}} + T_b \end{cases} \quad (6)$$

where D_{air} represents the air interface transmission load,

T_c represents the round-trip time (RTT) delay between the UE and cloud (cloud computing involved), and T_b represents the wired transmission delay between two base stations (collaborative base station computing involved).

Limited computing resources of UEs and base stations are denoted by C_l and C_b (in FLOPS), respectively. Additionally, we assume that the computing resources on the cloud are unlimited, meaning that the computing delay on the cloud can be ignored. D_l is the local computing load, and D_b is the computing load on base stations. In our verification using a set of base stations, deep neural network (DNN) split computing is used, splitting computing into local computing and collaborative base station computing. In local computing and collaborative base station computing, the computing delay of the UE can be expressed as the upper part and lower part of Equation 7, respectively:

$$T_{c-u_i} = \begin{cases} \frac{D_l}{C_l} \\ \frac{D_l}{C_l} + \frac{D_b}{C_{b-u_i}} \end{cases} \quad (7)$$

To validate the performance of our algorithm MADDPG-Adv, we compare MADDPG-Adv with the MADDPG and Communication Neural Net (CommNet) algorithms. MADDPG has an architecture that enables various agents to have independent actor and critic networks, making the actions of agents more flexible and diversified. This algorithm can be extended to highly dynamic network scenarios more easily. CommNet requires a more complex communications mechanism to coordinate learning between agents, resulting in limited flexibility. Figure 7 shows the training curves and convergence results of the

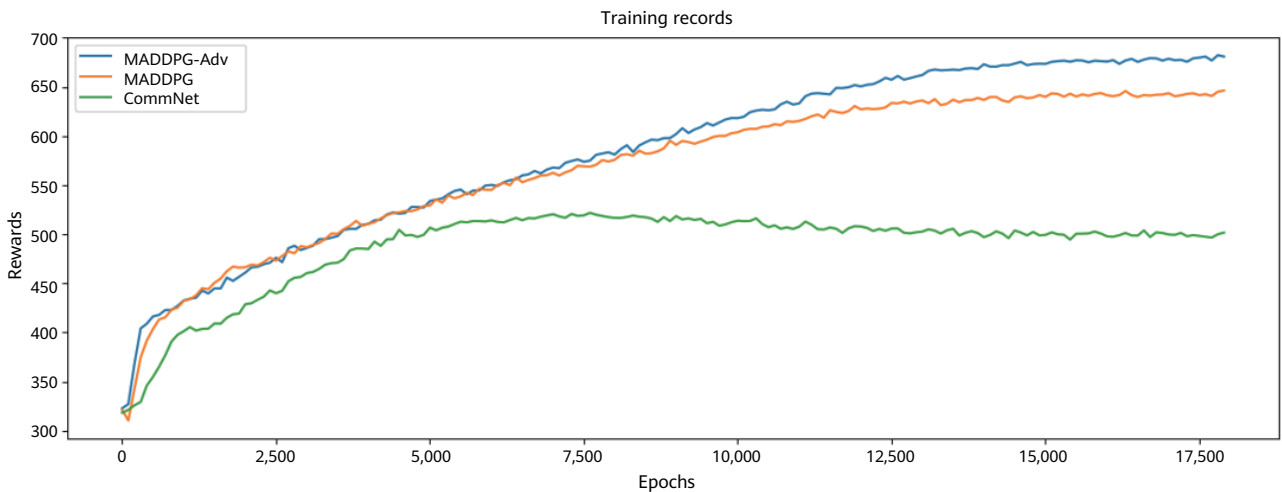


Figure 7 Comparison between performance of various algorithms

three algorithms. It demonstrates that our MADDPG-Adv algorithm has higher rewards than MADDPG and CommNet. MADDPG-Adv outperforms MADDPG by 6% and CommNet by 32%. Notably, our MADDPG-Adv algorithm is more suitable for the AI task control function.

To demonstrate that our algorithm can guarantee QoAIS, we measure the number of UE tasks for which both the accuracy and delay requirements are met. The requirements mean that the AI service version allocated to the UE must be later than or equal to the requested version, and that the UE delay must be within the required delay range. We compare the number of UE tasks meeting QoAIS requirements using the three algorithms. Figure 8 demonstrates that our MADDPG-Adv algorithm yields the largest number of UE tasks that meet the requirements, 23% higher than MADDPG.

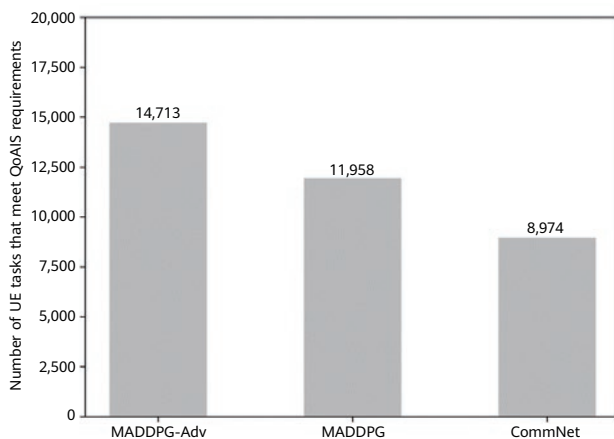


Figure 8 Comparison between the number of UE tasks meeting the QoAIS requirements

7 Conclusion and Outlook

In this paper, we focus on QoAIS assurance for native AI in 6G and discuss the architecture design, indicator systems, protocols and processes, and assurance technologies. We first propose the three-layer QoAIS mechanism with closed-loop feedback for 6G networks and develop the network function architecture. Then, we explore the design of QoAIS indicator systems, and related protocols for mapping from communication QoS and computing QoS to bearers. Additionally, we propose the QoAIS assurance technologies based on cloud-edge-device synergy and the multi-agent reinforcement learning algorithm. We simulate the AI inference scenarios and demonstrate the QoS assurance

effect by using accuracy and delay indicators. We verify that our reinforcement learning algorithm MADDPG-Adv algorithm improves the performance by 6% and achieves a 23% increase in the number of UE tasks meeting the QoAIS requirements when compared to the MADDPG algorithm.

QoAIS assurance requires comprehensive design paying full consideration to multiple aspects like architecture, protocols and processes, and technologies. In the future, we will further optimize the converged control mechanism and bearing of communication, computing, model, and other resources. By integrating our algorithm and architecture with fully designed protocols and processes, it will be possible to meet diverse requirements of various intelligent service scenarios, achieving more comprehensive assurance in terms of resource QoS, task QoS, and service QoS.

References

- [1] Guangyi Liu, Juan Deng, Qingbi Zheng, Gang Li, Xin Sun, and Yuhong Huang, "Native intelligence for 6G mobile network: Technical challenges, architecture and key features[J]," *The Journal of China Universities of Posts and Telecommunications*, 2022, 29(1): 27–40.
- [2] IMT-2030 (6G) Promotion Group, "Research report on wireless artificial intelligence technologies [R]."
- [3] Y. Yang *et al.*, "Task-oriented 6G native-AI network architecture," in *IEEE Network*, vol. 38, no. 1, pp. 219–227, Jan. 2024, doi: 10.1109/MNET.2023.3321464.
- [4] 6GANA, "Whitepaper on Distributed Learning of 6G" [R], 2023.
- [5] T. Chen, Q. Tang, and G. Liu, "Efficient task scheduling and resource allocation for AI training services in native AI wireless networks," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2023, pp. 637–642.
- [6] W. Wu, C. Zhou, M. Li, H. Wu, H. Zhou, N. Zhang, *et al.*, "AI-native network slicing for 6G networks," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 96–103, 2022.
- [7] M. Hua, T. Chen, N. Li, and H. Zhang, "Energy-efficient federated transfer learning in 6G native AI networks," *2023 IEEE Globecom Workshops (GC Wkshps)*, Kuala Lumpur, Malaysia, 2023, pp. 1746–1751.
- [8] GGII, "In-depth research report on mobile robot industry."
- [9] 3rd Generation Partnership Project (3GPP), "TR 22.874: Technical specification group services and system aspects; Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," Release 18, December 2021.
- [10] IDC (2023), "Insights into China's IoV security solution market."
- [11] 6GANA TG1, "6G Network AI Scenario Use Case Service Application Requirements [R]", 2022.
- [12] 3rd Generation Partnership Project (3GPP), "TS 23.503: Policy and charging control framework for the 5G system (5GS)," 2020.
- [13] 3rd Generation Partnership Project (3GPP), "TS 23.501: System architecture for the 5G system," 2020.
- [14] 3rd Generation Partnership Project (3GPP), "TR.38.901: Study on channel model for frequencies from 0.5 to 100 GHz," 2022.



Joint Orchestration and Management of Multidimensional Resources in 6G Intelligent Endogenous Networks

Dong Wang¹, Jianzhang Guo²

¹ China Telecom Research Institute

² China Telecom Digital Intelligence Technology

Abstract

In the 5G era, the focus has been on the cloudification and intelligentization of the core network. With the advent of 6G, the continuous cloudification and intelligentization of end-to-end (E2E) networks will make endogenous intelligence a defining feature of 6G networks, raising the requirements for joint orchestration and management of multidimensional resources, including communication, data, computing power, and AI models. This paper focuses on the technical requirements for joint orchestration and management of multidimensional resources in 6G intelligent endogenous networks, outlines the prospects of the 6G network architecture and the intelligent endogenous network solution, and proposes a solution for joint orchestration and management of connection, intelligence, sensing, and security. Additionally, this paper provides recommendations for future research, standardization, and prototype development.

Keywords

6G, intelligent endogenous network, multidimensional resource, joint orchestration and management

1 Introduction

With the implementation of new pattern industrialization in China, unprecedented development opportunities are ahead for the information and communication technology (ICT) industry [1]. The trend of ICT convergence that integrates communication technology (CT) and information technology (IT) is becoming prominent in fields with strong service requirements [2]. This trend will continue into the 6G era. The continuous ICT convergence is promoting the applications of innovative technologies, such as servitization, cloud native, and AI, to the mobile network. ICT convergence allows the network infrastructure to meet the requirements of the industrial field for deterministic quality characterized by low latency, low jitter, and high reliability, as well as requirements of HD video applications, Internet of Vehicles (IoV), and the industrial internet for high-bandwidth network infrastructure.

ICT convergence in the 5G era emphasizes the cloudification of the core network. As 5G networks see large-scale commercial development, the distributed network architecture, characterized by the separation of the user plane and control plane, has become increasingly refined. Moving forward, the industry has started systematic research on 6G technologies. The 6G network architecture is an important topic of research and crucial for realizing the 6G vision. Emerging technologies like immersive XR, metaverse, and digital twin, pose new requirements on the network in terms of high bandwidth, low latency, deterministic experience, and high computing power at the edge. Against this backdrop, E2E cloud-network convergence will become an inevitable trend in 6G [3], and the 6G network architecture will need to integrate advanced technologies.

ICT is evolving to become more diverse, converged, and intelligent. The convergence of data technology (DT) and operational technology (OT) is driving the transformation and capability upgrade of the next-generation mobile network. DT can significantly promote the intelligent transformation of networks and enhance user experience. The increasing development of AI foundation models (FMs), rising demand for intelligent computing power, and widespread adoption of intelligent computing infrastructure indicate that intelligence will play a pivotal role in driving industrial development. Furthermore, OT will facilitate the automatization and intelligent transformation and enhance network security, data security, and service security. According to the *Industrial Internet Innovation and Development Action Plan (2021–2023)*, China aims to

promote the network reconstruction of industrial equipment and upgrade internal networks of enterprises to achieve convergence of IT and OT networks.

In addition, to meet the service requirements of individual users for 6G networks, the 6G mobile communication network should be more intelligent, flexible, and secure. To enable intelligent decision-making and adaptive networking based on situational awareness, 6G networks must possess intelligence enablement and endogenous intelligence. This can be achieved through the networks' capabilities of autonomous learning and is dependent on ubiquitous computing capabilities and endogenous intelligence distributed in the resource sensing layer, function control layer, and service application layer. Endogenous intelligence will play a key role in optimizing and enhancing the network organization and service intelligence. 6G networks must be able to jointly orchestrate and manage multidimensional resources, including connections, data, computing power, models, and security, to address the challenges presented by the increasingly diversified and personalized user requirements.

2 Status Quo and Challenges

In 2021, the IMT-2030 (6G) Promotion Group launched an initiative for developing a 6G network architecture oriented to data, operation, information, and communication technology (DOICT) convergence. The promotion group identifies DOICT convergence as a direction of network development in a report of their phased progress [4]. In 2030 and beyond, 5G will fall behind with the requirements for emerging use cases such as immersive cloud XR, holographic communication, intelligent interaction, and digital twin. Therefore, the industry has realized that 6G networks must be highly intelligent and autonomous. The mobile network, as an important implementation of CT, is fully integrated with IT, with network functions virtualization (NFV), containers, software-defined networking (SDN), and API-based capability exposure giving full play to their roles in the system [5, 6]. The growth of the digital economy depends on the massive number of connections and data collection, modeling, and analysis. As operation and production requirements increase, OT and DT will infuse new energy into the evolution of networks. However, DOICT convergence is still in its infancy. Effectively orchestrating and managing multidimensional resources to achieve DOICT convergence in 6G intelligent endogenous networks presents a significant challenge.

In terms of intelligence enablement and endogenous intelligence, the existing network architecture should be improved to realize the 6G vision. Currently, 5G networks mainly use centralized, standalone AI, such as the Network Data Analytics Function (NWDAF). Network elements (NEs) like the Access and Mobility Management Function (AMF) and Session Management Function (SMF) lack AI capabilities and depend on centralized analysis provided by the standalone AI. This could pose certain challenges. Centralized AI requires massive volumes of data, leading to significant consumption of communication network resources. Additionally, in a large-scale network, centralized processing, analysis, and feedback are insufficient to meet the requirement for real-time performance [7]. Furthermore, centralized AI analysis requires highly aggregated computing power resources. Consequently, it is challenging to utilize multinode and edge computing power resources effectively. Effectively utilizing network AI capabilities, including the collaboration of AI resources (such as AI agents), is essential for realizing the 6G vision of intelligence enablement and endogenous intelligence. Extensive research is conducted regarding this topic.

The European Telecommunications Standards Institute (ETSI) has established several specification groups, including Experiential Networked Intelligence (ENI) and Zero-touch network and Service Management (ZSM). It has started standards formulation and research projects focusing on Autonomous Networks (AN) and network enablement. A number of technical reports and standards addressing the framework, application, and interfaces related to network intelligence have been released. 3GPP TS 28.535 [8] defines the closed-loop service level specification (COSLA), which involves monitoring, analysis, decision, and execution for closed-loop management and control in communication service assurance. Adjusting the resources for communication services relies on continuous iteration of steps in the management and control loop. The closed loop control of communication services is deployed in the preparation phase and takes effect in the preparation and during the lifecycle management. Future research on closed-loop control should focus more on multidimensional intelligent resources.

The academic community is also interested in the research on intelligence enablement and endogenous intelligence of 6G networks. A team led by Ping Zhang [9], an academician of the Chinese Academy of Engineering at Beijing University of Posts and Telecommunications, proposed Ubiquitous-X, a 6G information exchange hub. Based on the understanding of intelligence and human

consciousness, their innovative paradigm fully integrates humans, machines, things, and intelligence and defines new communication objects for future 6G networks. [10] proposed a simplified AI-enabled intent-driven 6G radio access network (RAN) architecture. This architecture is oriented to the networking of integrated terrestrial and non-terrestrial networks. Based on fog RANs (F-RANs), this architecture offers a flexible and reconfigurable system for communication, computing, caching, and control collaboration. [11] suggested an AI-driven 6G network architecture with endogenous intelligence. This architecture features AI models at the user layer, edge layer, central layer, and core network layer to provide the network with the capabilities of sensing and prediction, enabling the network to make decisions on offloading and scheduling in both centralized and distributed modes. [12] introduced the concept of *intelligent-concise RAN*. It proposes adopting a unified, integrated RAN architecture aiming to achieve performance objectives like wide coverage, huge capacity, massive connections, ultra-low latency, and high self-organization. Enabled by AI, this flexible and reconfigurable architecture features a variety of technologies and allows for collaboration of communication, computing, caching, and control (4C) and simplified networking.

Endogenous intelligence is recognized as a pivotal feature of 6G networks. To optimize the performance of 6G networks, we must conduct thorough research on sensing, joint orchestration, and management of multidimensional resources. This research should be based on network resource sensing and differentiation of user requirements, with a focus on leveraging endogenous intelligence as a key feature.

3 6G Network Architecture Prospect

The cross-domain convergence of new technologies will accelerate the development of the 6G network architecture and digital transformation in the industry. 6G, as the next-generation mobile communication system, deeply integrates communication technologies, computing networks, and AI technologies. The development of 6G exhibits strong interdisciplinary and cross-domain characteristics. 6G will evolve toward DOICT convergence. With 5G as its foundation, 6G will fully support global digital transformation and integrate next-generation ICT with various industries. This will accelerate development through digital transformation.

The evolution of the 6G network architecture must focus on user and customer requirements and cloud-network convergence to comprehensively improve mobile network service capabilities and user experience. Therefore, 6G networks require E2E integrated architecture design to provide higher flexibility and capability exposure, thereby implementing E2E orchestration and management of multidimensional resources.

The new 6G network architecture must meet the following requirements:

Research on the 6G network architecture must focus on Connection+ to enable the network to meet the requirements of emerging services, including immersive XR, holographic communication, human-machine interaction, and machine-machine interaction. In addition to essential connection services, these services require edge computing, high-precision environment and object sensing, and network AI services. Therefore, 6G networks should provide services that integrate connection, sensing, intelligence, and computing.

The architecture should incorporate innovative technologies while ensuring the implementation feasibility. In this sense, the 6G network architecture design should align with the evolution trends in mobile communication and DOICT convergence. At the same time, the impact of emerging technologies on 6G networks, such as AI, cloud computing, big data, and blockchain, should be considered. The architecture design must focus on an E2E systematic approach to avoid fragmentation and incompatibility. In terms of key technologies and protocols, horizontal collaboration between terminals, the wireless network, and the core network, and vertical collaboration across domains and layers are required. Specifically, joint orchestration and management of multidimensional resources should involve the operation management domain, bearer network domain, and cloud-network resource domain.

The architecture should allow for more flexibility in 6G networks to build an open industry ecosystem. 6G networks must be more adaptive to services by supporting on-demand orchestration and deployment, service loading, and traffic scheduling.

The architecture design should also consider 2B requirements. 5G networks play an important role in the growth of the digital economy and the digital transformation of industries. The 6G network holds significant promise for the 2B sector and will become increasingly vital in driving digital transformation. 2B and 2C networks have distinct service characteristics, deployment requirements, and management solutions. Therefore, the 6G network architecture design should address the requirements for both 2B and 2C networks. In light of this, we propose an architecture consisting of three layers and four planes for 6G networks, as shown in Figure 1.

This architecture is guided by cloudification and servitization. From bottom to top, the three layers are the cloud-network resource layer, network function layer, and application enablement layer.

- **Cloud-network resource layer:** As the provider of infrastructure and resources for 6G network deployment, the cloud-network resource layer supports a range of network functions. The resources include computing resources (such as CPUs, GPUs, and FPGAs), storage resources, network resources, and heterogeneous resources. IPv6/SRv6 transport networks are used to connect cloud-based network functions.
- **Network function layer:** This layer provides Connection+ services and is divided into four planes, namely, the control plane, user plane, data plane, and intelligence plane, to meet the requirements for different services and use cases.
- **Application enablement layer:** This layer aggregates network service capabilities and common application

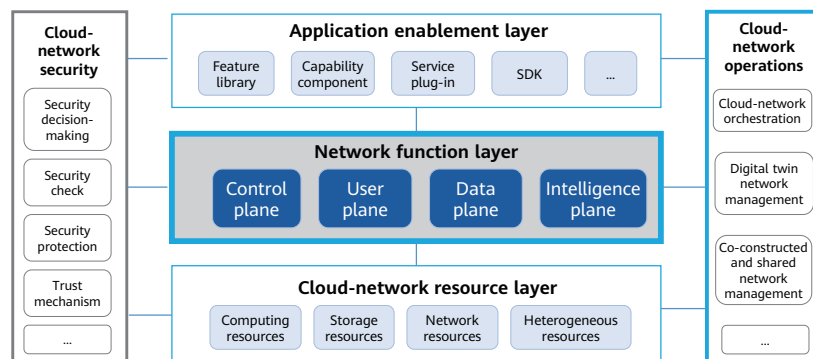


Figure 1 Layered architecture of 6G networks

service components, and provides services for applications or peripheral ecosystems through capability exposure and the application enablement framework. This design implements unified application enablement management and facilitates application development and deployment.

Cloud-network operations management involves operations management across all layers, including network management, service handling, and charging and settlement. With the evolution of cloud-network integration, new functions will be introduced, including intelligent cloud-network orchestration and scheduling, digital twin management, and blockchain-based co-constructed and shared network management. These functions will render the network more flexible and intelligent, promote efficient utilization of network resources, and optimize service deployment.

The four planes are an extension of 3GPP's mobile core network framework. To meet the requirements for Connection+ services, the logical functions of the four planes are designed as follows:

- **Control plane:** As the central component of network management, the control plane is responsible for managing network connection services, intelligent services, computing power services, and sensing services in a unified manner. It interworks with other planes to provide integrated control, identity authentication, mobility management, session management, policy control, and computing resource allocation and management for multiple access modes.
- **User plane:** The user plane supports network programmability and allows for flexible definition of data processing policies. Its main functions include tunnel management, data flow identification, service awareness, deterministic communication assurance, data encapsulation, data forwarding, and traffic steering. The user plane functions as a policy, action, and routing node between users and the data layer, processing and transferring data, including user data, environment object sensing data, and AI task data.
- **Data plane:** The data plane separates data from the service logic. The data plane manages various types of data and provides data for the control plane, user plane, and intelligence plane through standard interfaces. The data encompasses static data and dynamic real-time data, such as user registration details, network status, and connection information.
- **Intelligence plane:** The intelligence plane is the core of 6G network intelligence and enables comprehensive intelligence across the core and access networks. In addition to addressing the intelligence requirements of the network itself, the intelligence plane should also meet the requirements of user and service applications. The intelligence plane provides network AI functions, including data modeling, model training, inference and decision-making, knowledge graph, and feedback and evaluation. These functions enable intelligent network management and service optimization.

4 6G Intelligent Endogenous Networks

With the continuous development of AI technologies, the industry has reached a consensus that AI will be one of the core capabilities of 6G networks. To improve user experience of AI services and intelligent network operations, 6G intelligent networks need to process massive, multimodal, heterogeneous, and privacy-sensitive data that is distributed across network nodes, supporting model training for services and meeting requirements for automatic closed-loop management of inference, decision-making, evaluation, and optimization. In 5G networks, AI is typically deployed externally and is primarily used for data analysis and prediction. This deployment mode has several issues, including weak distributed learning capabilities, a lack of flexibility in data processing, and the absence of decision feedback mechanisms and knowledge accumulation. Therefore, 6G networks should have the following features:

- **Collaborative group intelligence:** Given the massive volume of multimodal and heterogeneous data distributed across a network, with network nodes at different levels performing different functions, and high-performance machine learning (ML) consuming significant resources, 6G intelligent endogenous networks will be characterized by ubiquitous connectivity, ultra-distribution, and multinode cross-domain collaboration.
- **Support for various AI learning methods:** 6G will offer a diverse range of services, necessitating support for a variety of learning methods to provide users with the ultimate experience. In addition to basic ML methods such as supervised learning, reinforcement learning, and deep learning, 6G should also integrate multinode collaborative learning methods of the

distributed AI architecture, including federated learning, group learning, and multiagent reinforcement learning. Additionally, 6G should leverage knowledge accumulation to improve learning efficiency, including knowledge-driven transfer learning and lifelong learning [12, 13].

- **Enhanced data sensing and processing:** Endogenous intelligence requires 6G networks to have independent data processing and analysis NEs with dedicated sensing and processing modules. In this way, data can be separated from forwarding and controlling, resulting in a distributed and collaborative data plane. By configuring flexible rules of data sensing and performing real-time processing and AI analysis on network data, data is seamlessly integrated into the network.
- **Autonomous decision-making:** AI-enabled decision-making capabilities or NEs are added to 6G networks to interact with the inference and decision-making modules of NEs, forming a closed-loop decision-making system across nodes in a complex heterogeneous network. The network can adaptively select or intelligently make decisions based on real-time network conditions and AI-driven analysis and inference results. This should be underpinned by decision evaluation and validation capabilities, ensuring real-time and effective service assurance.
- **Knowledge accumulation:** 6G networks should be capable of knowledge extraction and accumulation to construct knowledge graphs that facilitate network operations autonomously. Knowledge such as network graphs, expert experience, and prior rules are accumulated in the knowledge base. Nodes can query and obtain AI service-related knowledge in real time to improve learning efficiency and intelligence levels, thereby implementing data- and knowledge-driven network operations.

5 Orchestration and Management of Multidimensional Resources

To discuss the orchestration and management of multidimensional resources, it is necessary to clarify the dimensions of the resources involved in DOICT convergence of 6G networks. Generally, network cloudification requires unified scheduling of communication and intelligent computing resources, including data storage, models, and computing power. As wireless sensing technology continues

to develop, integrated sensing and communication (ISAC) applications are also integrated. In addition, information security has always been a concern in network communication. 6G should consider endogenous security requirements. Specifically, security assessment should be incorporated into the E2E process of resource orchestration and management.

Figure 2 shows a system analysis solution for joint orchestration and management of multidimensional resources. The multidimensional resource orchestration and management in the 6G intelligent endogenous networks necessitate both horizontal and in-depth integration. Horizontally, cloud storage capabilities and AI models should be integrated with connections based on technology maturity. Additionally, as computing power requirements increase, the optimal computing power resource pool should be selected based on the computing power latency and transmission latency for service instances with different resource requirements. This ensures that computing tasks that could not be completed locally in time are efficiently executed. Communication and cloud computing belong to different technical domains. They are interdependent yet independent of each other, and there are noticeable differences in their standards. The in-depth integration of the two is a long-term process as well as a challenge to be addressed.

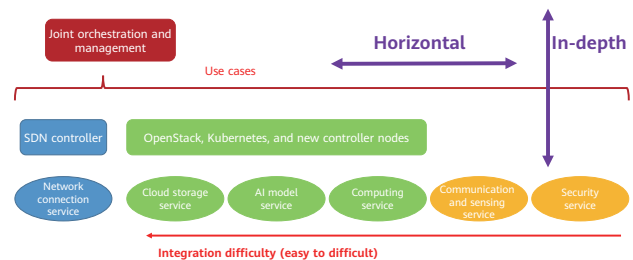


Figure 2 Requirement analysis of joint orchestration and management of multidimensional resources

Currently, sensing applications rely on electromagnetic waves in the wireless air interface, so ISAC is still limited to the access network. Sensing services require independent controllers and data processing, and independent of the communication system. As ISAC applications become more mature, service applications will be implemented based on the same core network and service platform as mobile communications, sharing the management and control capabilities of the 6G core network. In this way, signals and services will be processed based on common intelligent computing data resources.

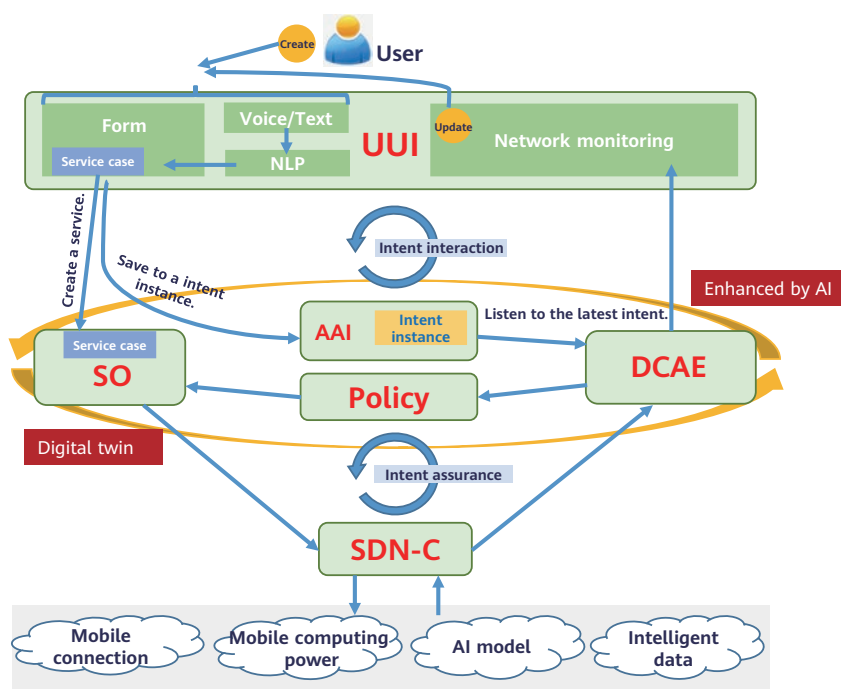


Figure 3 Architecture of joint orchestration and management of multidimensional resources

As shown in Figure 3, to provide differentiated and personalized intelligent services for users and flexibly schedule cloud-network computing resources across the network, intent-based services are provided for all intelligent use cases, implementing E2E integrated dynamic orchestration and management of the access network, bearer network, core network, air-space network, and cloud computing resources. To achieve closed-loop autonomy, the architecture has two closed-loop structures. The outer closed-loop structure comprises two data flows, which are used for intent creation and delivery and intent modification and satisfaction. The data flow of intent creation and delivery is as follows: Use-case User Interface (UUI) > Service Orchestrator (SO) > Software-Defined Networking Controller (SDN-C) > Data Collection, Analysis, and Events (DCAE) > UUI. The data flow of intent modification and satisfaction is as follows: UUI > Active and Available Inventory (AAI) > DCAE > Policy > SO > SDN-C > DCAE > UUI. The inner closed-loop structure implements user intent assurance and includes four phases: monitoring to analysis (M2A), analysis to decision (A2D), decision to execution (D2E), and execution to monitoring (E2M). Tasks in each phase are defined and implemented by modules (such as the intent instance module, policy module, and service orchestration module) in the architecture and deployed on the open network automation platform (ONAP). This architecture is highly feasible due to the lightweight nature of its modules, which can be seamlessly decoupled.

Additionally, it employs a standard network intent verification process and interfaces.

The intent input and translation module in the intent-driven network is implemented based on the UUI. This module is designed to parse users' natural language input, either in voice or text form, to determine their network intent requirements. Network intent requirements encompass the requirements for network service quality, basic network configuration, and network status. Network service quality requirements are essentially network configuration parameters, including those derived from intent-aware network quality requirements. The basic network configuration includes details about user locations. The network status involves real-time traffic, latency, and jitter, which are used to determine whether the network status meets the intent requirements and to guide network configuration policy adjustment. Entity extraction and identification are performed based on the voice or textual input of network intent requirements using natural language processing (NLP) methods, such as the BERT algorithm. After NLP, the natural language requirements are translated into corresponding parameters. SDN-C and SO are responsible for SDN control, fulfilling the intent, and feeding back to DCAE. The intent verification module verifies these policies to evaluate whether the network status meets the service level agreement (SLA). Finally, DCAE sends the evaluation result to UUI for the user's decision-making.

The intent instance management technology in the intent-driven network can implement closed-loop management of user intents and network status monitoring. When DCAE detects that the network status changes, it notifies UII to request user intervention. On UII, the user modifies the intent instance parameters stored in AAI. Then, AAI notifies DCAE of the notification to update the stored user intent and intent translation result parameters and read and update the stored network status feedback information.

The NLP function in UII classifies user intents based on matched use cases or services, extracts parameters from text, fills the parameters in the automatically generated service request form, and displays the form to the user for confirmation or modification. The service request form specifies the service type, requirements, and SLA clauses, and user intents in a formatted way.

SO translates a customer service model into a service delivery model that defines how to design services in the network. SDN-C converts the service delivery model into a corresponding network configuration model and applies the model to a physical network.

The advancement of such model-driven design is reflected in three aspects: (1) The model for intent translation is standardized through SO, which facilitates the standardization of the intent-driven network solution. (2) The model instances in each step are stored in the AAI database and can be retrieved using RESTful interfaces. (3) The data model is decoupled from the code logic for accessing (reading/writing) data, so that the code logic can be improved and developed independently of the data model, thereby minimizing the impact on other components.

The inner closed-loop of intent assurance consists of four phases: monitoring, analysis, decision-making, and execution. The monitoring phase is implemented on SDN-C. It collects monitoring and performance data from the network controller and forwards the data to DCAE. SDN-C determines the data to be collected based on the SLA parameters obtained from the user intent during intent translation.

In the analysis phase, DCAE detects network exceptions by analyzing the monitoring data received from SDN-C and provides feedback for Policy. If an exception is detected, DCAE notifies Policy to take corrective measures.

The decision-making phase is implemented on Policy. Policy makes closed-loop decisions based on the data received from DCAE and issues appropriate recommendations to SO to perform service changes.

The execution phase involves a typical SDN orchestration and control workflow. SO and SDN-C apply the new network configuration to the physical network.

All data, including monitoring data, service models, network resources, and configurations, is stored in AAI and can be retrieved and shared in every phase.

6 Intent Instance: Requirement Entity for Multidimensional Resource Orchestration

The intent instance is a type of important entity oriented to 6G intelligent endogenous networks. An intent instance can contain information such as the original intent, intent translation, and network status information. This data entity is created to enable effective storage, management, and interactive update of user intents. It is used to standardize and associate the storage and interactive update of user intents, the storage, reading, and update of intent translation parameters, and the storage of network status feedback information. Intent instance management technology implements closed-loop awareness and feedback of user intents, enabling closed-loop management of user intents and network status monitoring. It summarizes user information, including original intents, intent translations, and network status feedback, and allows this information to be exported after anonymization. The data can be used to develop specifications that guide subsequent training and application of intelligent intent-based network (IBN) algorithms and models.

Closed-loop orchestration management uses orchestrators to abstract resource status based on user intent requirements and deploy and configure different service intents in the form of network slices and virtual networks to implement E2E automatic service deployment, load balancing, and service assurance. Based on the ONAP, the E2E automatic service deployment technology implements automatic arrangement, coordination, and management of complex computer systems, middleware, and services. This way, network resources are automatically and intelligently managed and orchestrated in all scenarios. Service requirements are extensive and varied. The closed-loop management system, deployable in containers, ensures that closed-loop services remain highly reliable and scalable, preventing service interruptions that might occur due to sudden spikes in requirements. Additionally, a cloud-edge collaboration platform can be deployed at the edge. This

approach leverages cloud-edge collaboration to address issues such as heavy loads from large data volumes on the wireless communication cloud platform and incomplete data analysis on the wireless communication cloud computing platform.

In terms of the intent instance management model, we need to design an available resource library to store intent instances and related configurations and receive intent translations from UUI. Additionally, a data collection and analysis module is required to query user intent in the available resource library and process intent updates. Furthermore, a service orchestration module needs to be designed to manage intent instances, scenarios, and resources.

Intent instance management involves the addition, deletion, modification, and query of intent instances. In the process of creating an intent instance, UUI reads the use case service ID, binds the user intent to the created intent instance ID and the read user service ID, and stores them in the available resource library. Additionally, the available resource library provides a user intent monitoring interface for the data collection and analysis module to listen to possible user intent updates. During the intent modification and update process, UUI reads the new use case service ID, stores the ID in the available resource library, and associates the ID with the existing intent instance. The intent instance management allows intents to be queried and deleted.

Intent instance management associates the intent-related information based on mappings between intent instances

and data tables (for example, using intent instance IDs as key IDs). For instance, IDs of use case services in various scenarios can be associated with intent instance IDs to meet user requirements in multiple scenarios. To handle continuously updated user intents and intent translations, we can associate them with unified intent instances. Additionally, associating network changes and intent assurance with unified intent instances enables the evaluation and optimization of the network support for user intents.

Intent instance management supports the functions of data summary and anonymized data export. Information about network intents can be exported for training and application of intelligent algorithm models.

As shown in Figure 4, the intent-driven network translates a user service intent into a series of executable network tasks through steps including intent input, intent translation, conflict detection, intent execution, and intent assurance. Transformation, verification, deployment, configuration, and optimization are automatically performed based on user and operator intents to achieve the target network status and automatically resolve network exceptions, thereby ensuring network reliability. However, in a network environment with diverse requirements in various scenarios, current intent-driven networks face issues such as inaccurate intent translations, inflexible intent policies, and complex intent management.

Compared with traditional AI models, FMs have significant advantages in understanding intents, inference, and

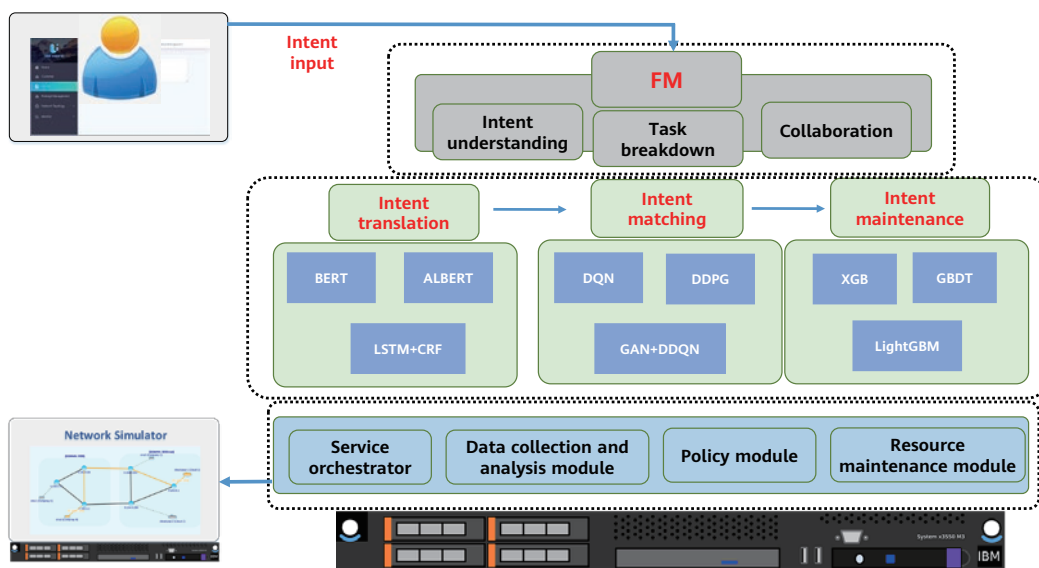


Figure 4 Application of the intelligent orchestration and management model

decision-making. Integrating FMs with intent-driven networks can effectively improve the intent translation accuracy and simplify intent management and network O&M. Specifically, applying FMs to intent identification and translations enhances the accuracy of identifying user intents and requirements; using them for policy matching and execution enables intelligent alignment of user requirements with network configuration plans, achieving automatic, user requirement-based network customization; and employing them in network monitoring and maintenance allows for the detection of network environment changes, thereby supporting autonomous network optimization, deployment, or configuration.

FMs can be deployed in the operations domain to achieve an overall architecture featuring "small model collaboration with FMs at the core." In this architecture, FMs break down and classify user intent tasks (such as intent translation, matching, and maintenance) for small models to complete. Using this collaborative approach and focusing on intent-driven methods, we aim to enhance the intelligence of E2E orchestration and management.

Furthermore, in view of the requirements for joint E2E scheduling and orchestration of multiple elements in the next-generation architecture, FMs, by enabling E2E intent-driven processes, can help comprehensively improve E2E orchestration and network performance.

7 Suggestions for Future Work

Focusing on 6G intelligent endogenous networks, this paper explores a DOICT convergence-oriented architecture with endogenous intelligence at its core. It also offers compatible solutions for the joint orchestration and management of multidimensional resources. However, the current research on 6G network architecture is still in its early stages, with much work yet to be completed.

During the evolution of 6G networks, we need to focus more on supporting multimodal, heterogeneous, and all-scenario access and explore ways to meet the personalized network requirements of users in different industries. We also need to find ways to address the challenges posed by the current network architecture design.

Additionally, future research should focus on further optimizing the intelligent endogenous network architecture to implement integrated management and scheduling of multidimensional resources. This includes enhancing the

customization capabilities of regional AI, service-specific AI, and edge AI, as well as improving the sensing processing and decision-making capabilities of AI embedded in NEs. This way, we can continuously improve the network architecture and enhance network intelligence, allowing networks to meet the increasing user requirements and address the challenges in emerging use cases.

Future research should also explore methods to meet the requirements of new services, such as immersive XR, metaverse, and digital twin, for extremely high network performance. The application scenarios of intelligent endogenous networks should be extended to include smart cities, intelligent transportation, and industrial internet. By deeply integrating intelligence with various industries, we can make intelligent services ubiquitous, thereby making people's lives more convenient and facilitating social and economic development.

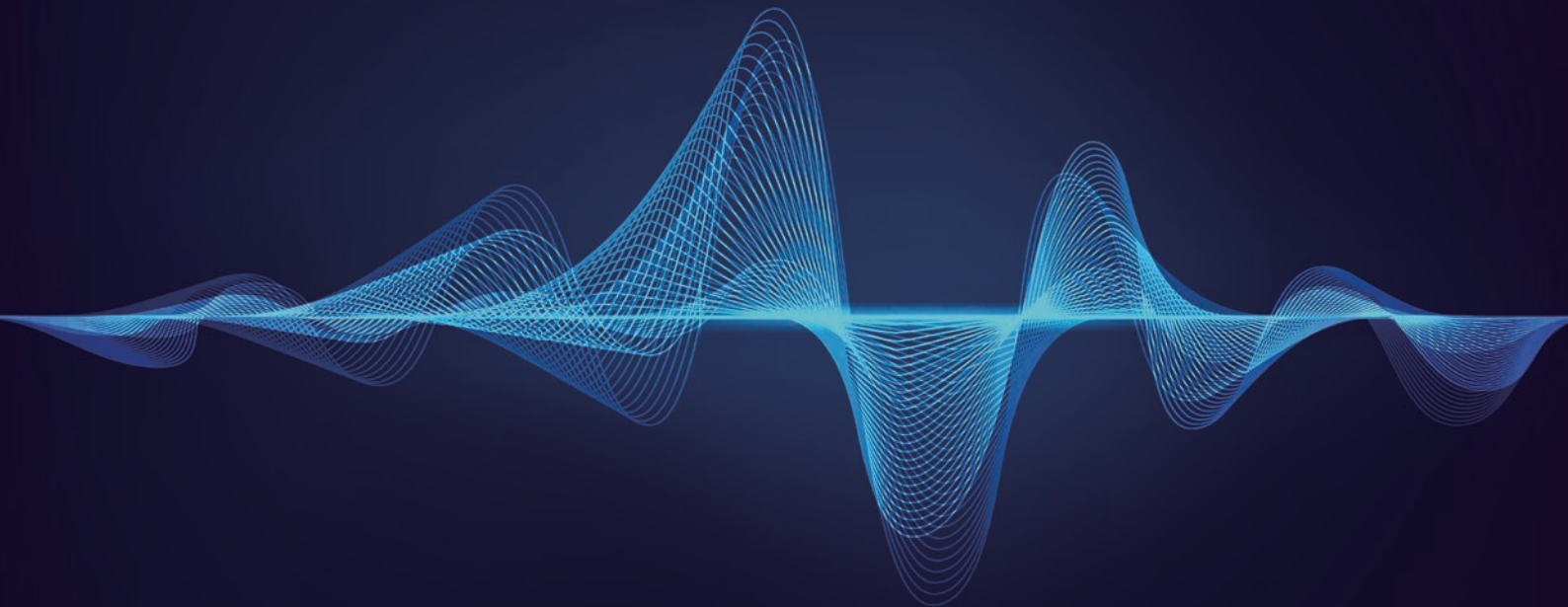
Considerable research needs to be conducted on the joint orchestration and management of multidimensional resources in 6G intelligent endogenous networks. To realize on-demand services across all scenarios in the 6G era, the industry should work together to carry out prototype research and development and standards formulation.

8 Conclusion

The consensus within the industry is that 6G networks must incorporate endogenous intelligence as a key feature. This requires improving the orchestration and management of multidimensional resources to support DOICT convergence in the 6G intelligent endogenous networks, which will enhance network service capabilities. Effective coordination of resources such as connections, data, computing power, and algorithms is necessary to achieve endogenous intelligence and improve network intelligence. Our research focuses on providing future directions for joint orchestration and management of multidimensional resources in the 6G intelligent endogenous networks.

References

- [1] "2024 nian xinxitongxin (ICT) ye shi da qushi" 2024 年信息通信业 (ICT) 十大趋势 [J]. *Hulianwang tiandi* 互联网天地, 2023(12): 12.
- [2] W Wang, J Zhou, and B Huang. ODICT integrated network 2030 [J]. *ZTE Technologies Journal*, 2022, 28(01): 47–56.
- [3] P Zhang, W Li, and K Niu. "6G xuqiu yu yuanjing" 6G 需求与愿景 [M]. Posts & Telecom Press, 2021.
- [4] "IMT-2030 (6G) tuijinzu fabu 6G baipishu he jishubaogao miaohui 6G weilai" IMT-2030(6G) 推进组发布 6G 白皮书和技术报告 描绘 6G 未来 [J]. *Information Technology & Standardization*, 2021(10): 6.
- [5] IMT-2030. "6G wangluo jiagou yuanjing yu guanjian jishu zhanwang" 6G 网络架构愿景与关键技术展望 [R]. 2021
- [6] Z Long. "SDN yu DFV jishu zai dianxin hexinwang yanjin zhong de yingyong" SDN 与 DFV 技术在电信核心网演进中的应用 [J]. *China New Communications*, 2020, 22 (21): 86–87.
- [7] B Lei and Y Chen. "Bianyuan jisuan yu suanliwangluo—5G+AI shidai de xinxing suanli pingtai yu wangluo lianjie" 边缘计算与算力网络——5G+AI 时代的新型算力平台与网络连接 [J]. *Zhongguo xinxihua* 中国信息化, 2020(12): 113.
- [8] 3GPP. TR 28.535, "Management and orchestration; Management services for communication service assurance; Requirements [R]," 2024.
- [9] P Zhang, W Xu, H Gao, *et al.* "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks [J]," *Engineering*, 8: 60-73, 2022.
- [10] N Jiang, C Zhang, C Kang, *et al.* "Low earth orbit satellite network based on communication, sensing and computing integration: Architecture and key technologies [J]," *Radio Communications Technology*, 2023, 49 (05): 842–852.
- [11] T Zhang, Y Ren, S Yan, *et al.* "Artificial intelligence driven 6G networks: Endogenous intelligence [J]," *Telecommunications Science*, 2020, 36 (09): 14–22.
- [12] M Peng, Y Sun, and W Wang. "Intelligent-concise radio access networks in 6G: Architecture, techniques and insight [J]," *Journal of Beijing University of Posts and Telecommunications*, 2020, 43 (03): 1–10. DOI:10.13190/j.jbupt.2020-079.
- [13] Yu M, Li P, Xing Y, *et al.* "A method to improve the performance of network data analytics function based on transfer learning [C]," 2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2023: 1–5.
- [14] Y Liu, Y Xing, and P Chen. "Prospects for 6G network architecture [J]," *ZTE Technologies Journal*, 2023, 29 (05): 16–20.



An Exploration on 6G-oriented Transmission and Reception Solutions for Non-Orthogonal Superimposed Pilots

Han Xiao, Wenqiang Tian, Xufei Zheng, Wendong Liu, Jia Shen
OPPO Research Institute

Abstract

Pilot design is a fundamental issue in wireless communications systems. In current systems, pilot transmission and data transmission compete for wireless transmission resources. However, the introduction of an AI receiver with powerful nonlinear processing capabilities can relax the need for pilot orthogonality, opening up possibilities to re-exploring the resource allocation between pilots and data in 6G. In this paper, we investigate the non-orthogonal superimposed pilot (SIP) and AI receiver solutions. First, we present the overall framework of the SIP+AI solutions, and demonstrate their performance advantages through simulation in basic transmission scenarios. Then, we introduce various potential implementations in complex scenarios with multi-party verification. We also propose an inference cancellation-based AI receiver, taking into account the effectiveness in multi-layer transmission and the scalability in practical deployment. The research findings in this paper provide insights into potential research and drive the promotion and standardization of topics related to the relationship between pilot and data resource allocation in future 6G systems.

Keywords

SIP, AI receiver, interference cancellation, multi-layer transmission

1 Introduction

Accurate channel estimation is crucial for ensuring reliable and effective links in wireless communications systems, with pilot design playing a central role in the estimation performance. A range of pilots [1] with predefined patterns and sequences are introduced in existing 5G systems, offering various functions such as channel estimation, resource scheduling, link adaptation, and beam management. Popular pilot solutions include DeModulation Reference Signal (DMRS), Channel State Information Reference Signal (CSI-RS), and Sounding Reference Signal (SRS). AI-based pilot solutions, such as pilots with deep learning-based pattern and sequence design [2–7], also exhibit immense potential in performance improvement. However, in all these solutions, the pilot symbols and data symbols are orthogonally allocated on time-frequency resources, meaning that pilots and data contend for limited transmission resources. This results in high pilot overheads and reduces spectral efficiency of data transmission, limiting the throughput performance of a system. Additionally, many of the candidate technologies proposed for 6G — such as larger-scale multiple-input multiple-output (MIMO), higher-precision beam management, ultra-high-speed mobility, scenario sensing, and high-precision positioning [8, 9] — have a heavy dependency on pilot transmission. These technologies pose diversified requirements on pilot designs that generate different resource overheads, making resource competition between data and pilot transmission more complex. Consequently, there is a pressing need to design a novel transmission policy for pilots and data.

In this paper, we propose a non-orthogonal solution called superimposed pilot (SIP) and an AI receiver solution. Specifically, the transmit end non-orthogonally superimposes pilots and data on time-frequency resources, and the receive end applies an AI receiver to perform effective channel estimation and symbol detection. Our solutions enable sharing of time-frequency resources between pilots and data, significantly enhancing the spectral efficiency. We first present an overview of the SIP+AI framework and conduct simulation verification in basic transmission scenarios. Additionally, we introduce various implementations in complex scenarios with multi-party verification. We also propose an interference cancellation-based AI receiver to ensure that the SIP+AI solutions work effectively in multi-layer and high-speed scenarios and that they are scalable in practical deployment. Our study and simulation verification results demonstrate the technical advantages and application potential of SIP with an AI

receiver. These solutions are expected to play a pivotal role in future 6G research and standardization.

2 Superimposed Pilots with AI Receiver

2.1 Basic Framework

In traditional solutions, pilots and data are transmitted orthogonally, resulting in severe resource competition. To address this issue, an innovative SIP solution [10] superimposes and sends pilot symbols and data symbols at the transmit end. By simultaneously transmitting pilots and data using the same time-frequency resources, this solution can achieve spectrum resource sharing, thereby significantly improving the spectral efficiency of a communications system.

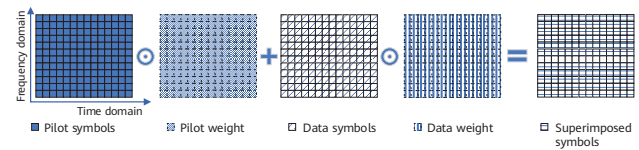


Figure 1 Non-orthogonal SIP mode

As shown in Figure 1, the pilot symbol matrix and data symbol matrix in the time-frequency resources can be weighted and superimposed at the transmit end, generating a matrix of superimposed symbols for transmission. Consider a downlink system with N_t transmit antennas, N_r receive antennas, and L layers. This system has S subcarriers in the frequency domain and T symbols in the time domain. During downlink transmission, the transmitter superimposes pilots and data non-orthogonally, obtaining the following superimposed symbols:

$$\mathbf{S} = \text{sqrt}(\mathbf{W}) \odot \mathbf{D} + \text{sqrt}(\mathbf{V}) \odot \mathbf{P}$$

where $\mathbf{D} \in \mathcal{C}^{L \times T \times S}$ represents the data symbol tensor, \mathcal{C} represents the set of complex numbers, $\mathbf{P} \in \mathcal{C}^{L \times T \times S}$ represents the pilot symbol tensor, and $\mathbf{S} \in \mathcal{C}^{L \times T \times S}$ represents the superimposed symbol tensor. $\mathbf{W} \in \mathcal{R}^{L \times T \times S}$ and $\mathbf{V} \in \mathcal{R}^{L \times T \times S}$ represent the weight tensor of data and pilots, respectively, where \mathcal{R} denotes the set of real numbers. $\text{sqrt}(\cdot)$ denotes square root calculation, and \odot denotes the Hadamard product. The superimposed symbols are transmitted to the receive end over channels. The received signal can be expressed as

$$\mathbf{Y}_r = \sum_{l=1}^L \mathbf{H}_{rl} \odot \mathbf{S}_l + \mathbf{N}_r$$

where $\mathbf{Y}_r \in C^{T \times S}$ denotes the received signal for the r th receive antenna, $1 \leq r \leq N_r$ and $1 \leq l \leq L$ are the receive antenna index and layer index, respectively. $\mathbf{H}_{r,l} \in C^{T \times S}$ denotes the equivalent channel for the r th receive antenna and l th layer. $\mathbf{N}_r \in C^{T \times S}$ is the additive white Gaussian noise. Finally, the received signals \mathbf{Y}_r obtained using the N_r receive antennas are concatenated, acquiring the final received signal $\mathbf{Y} \in C^{T \times S \times N_r}$.

In the SIP solution framework, each transmission resource unit carries both pilot and data information. This enables the AI receiver to perform channel estimation using the pilot information in superimposed symbols while reaching a high spectral efficiency. Notably, the power allocation ratio for pilot and data symbols needs to be carefully configured during superposition in order to strike a balance between an equivalent signal-to-noise ratio (SNR) of data transmission and channel estimation performance. This ensures that signals can be effectively restored at the receive end. To effectively process and receive superimposed signals, a cutting-edge AI receiver is applied at the receive end. The input of the AI receiver is the received signals obtained after modulation, non-orthogonal pilot superposition, and channel transmission, and the output is the log-likelihood ratios (LLRs) or information bits of layers. This AI receiver can serve as an integrated receiver for simultaneous implicit channel estimation and data recovery. It can also serve as a modular receiver that performs explicit channel estimation and data recovery separately.

2.2 Simulation Analysis in Typical Scenarios

This section provides the simulation results of the SIP+AI solutions (using the framework described earlier) in the basic configuration environment. The settings include 4 GHz carrier frequency, 30 kHz subcarrier spacing, 16QAM modulation, LDPC channel coding, and SVD pre-coding. The 4P+LMMSE baseline solutions use the four-symbol orthogonal pilot design and linear minimum mean square error (LMMSE) receiver. The covariance matrix of LMMSE-based channel estimation is derived from statistics of 10^5 channel samples, whereas the AI receiver is implemented based on ResNet [11].

Figure 2 presents the link-level simulation result of the channel model of Urban Macrocell (UMa) in terms of the block error rate (BLER). The SIP+AI solutions achieve BLER performance comparable to that of the baseline orthogonal

pilot solutions. This means that the SIP+AI solutions do not cause any additional performance loss in BLER. Additionally, the AI receiver at the receive end can effectively receive data in the SIP+AI solutions.

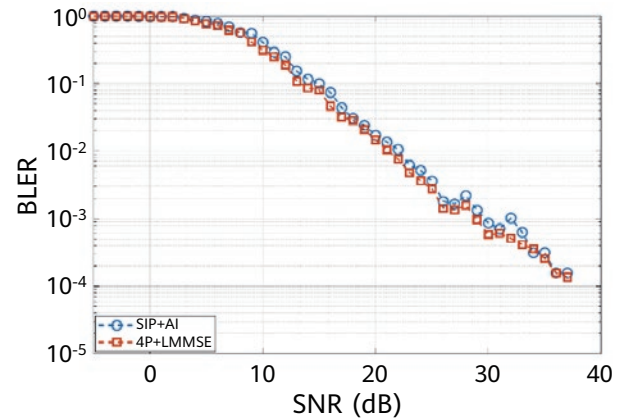


Figure 2 Comparison of BLER performance in typical scenarios (UMa, $N_t = 1$, $N_r = 1$, $L = 1$, $T = 12$, $S = 624$, 300 km/h, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$)

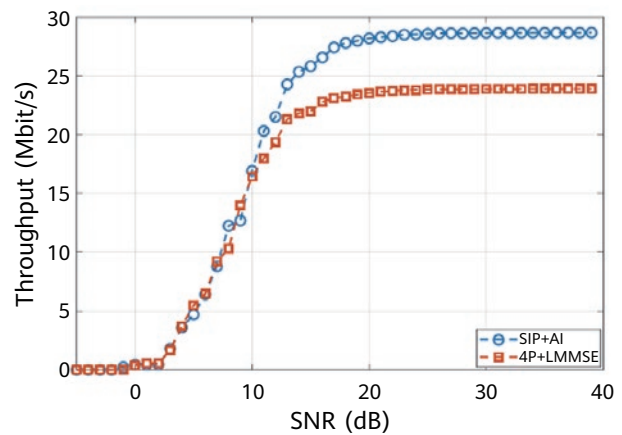


Figure 3 Comparison of throughput performance in typical scenarios (UMa, $N_t = 1$, $N_r = 1$, $L = 1$, $T = 12$, $S = 624$, 300 km/h, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$)

Figure 3 compares the solutions by the throughput performance. The SIP+AI solutions, with comparable BLER performance, can transmit more data information bits at the same bit rate. This is because these solutions avoid resource overheads caused by orthogonal pilot design of the baseline solutions. In turn, the SIP+AI solutions have higher spectral efficiency and throughput performance.

Related research has been conducted on transmission of superimposed data and pilots [11–15]. Considering complex channel conditions, the SIP+AI solutions need to be further improved in terms of performance and adaptability, particularly for multi-layer transmission and practical deployment. This requires diversified improvement to SIP+AI receiver.

3 Complex Scenarios and Multi-Party Verification

3.1 Superimposed Pilots in Complex Scenarios

Multi-layer transmission uses multiple transmit and receive antennas to simultaneously send and receive multiple independent data layers. This significantly enhances the data transmission capability and spectral efficiency of a communications system. In orthogonal pilot solutions of existing communications systems, pilots of different layers are orthogonally configured, enabling good estimation for channels at different layers. Additionally, by using the precoding function for creating equivalent channels featuring low channel correlation, interference among data of different layers can be further reduced, enabling data to be effectively received. However, in multi-layer transmission, the SIP+AI solutions cause more severe inter- and intra-layer interference than traditional orthogonal pilot solutions. This means that the AI receiver must deal with intra- and inter-layer data and pilot interference when receiving each layer. To resolve the interference challenge, we further our exploration in improving the SIP and AI receiver. Various superposition pilot patterns can be designed for handling different levels of interference.

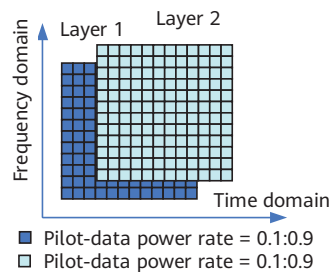


Figure 4 SIP pattern for two-layer transmission

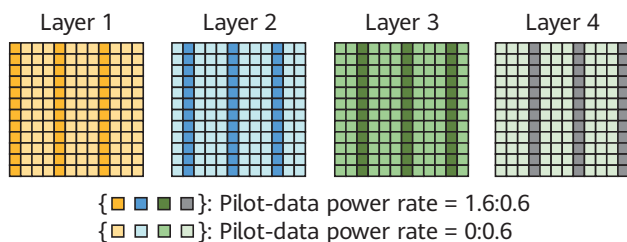


Figure 5 SIP pattern for four-layer transmission

We assume that pilots can still be superimposed on the full time-frequency resources in two-layer transmission with low inter-layer interference, as illustrated in Figure 4. As the number of layers increases to 4, the interference becomes more severe. SIP for different layers can be orthogonally configured on time-frequency resources to further reduce inter-layer pilot interference, as shown in Figure 5.

In practical deployment, the AI receiver for SIP also faces challenges with the storage and computational complexity of the model on the terminal side. Consequently, there is a critical need to design a lightweight AI receiver that is suitable for terminal deployment for SIP in multi-layer transmission. This will enable the receive end to effectively receive data while acquiring system gains from non-orthogonal pilot design.

3.2 Potential Implementations

3.2.1 Data Processing and Augmentation

Data is key to the training and construction of an AI receiver. Data and the techniques used for processing and augmenting it directly determine the ultimate performance of the AI receiver. To enhance the receiving performance and generalization capability of the model, multiple data augmentation techniques can be leveraged. These techniques can be categorized into two types, (1) common augmentation techniques in the computer field and (2) adaptive augmentation techniques derived from wireless communications knowledge.

- (1) New samples of received signals can be generated through linear sum of different received signals, thereby diversifying model inputs. Further to this, data can be flipped in the antenna, layer, subcarrier, and time-domain symbol dimensions, making the model learn complex and diverse features of received signals.
- (2) By introducing communications knowledge, some techniques adopt time-frequency aliasing. Specifically, a weighted sum of a symbol in a resource unit and symbols of the neighboring resource units of a received signal is calculated and used as the resource unit symbol after data enhancement. The aliasing process aims to simulate the potential inter-symbol and inter-carrier interference during signal transmission. Other techniques use frequency-domain clipping to enable the model to learn features of received

signals working on different bandwidths. In this way, model integration can be performed after an entire received signal is received using various sub-bands, enhancing performance and receiving flexibility in practical deployment. Additionally, random rotation and Gaussian noise can be added to received signals in order to further simulate changes of noise and signal phase in links, making the model more robust against noise and channel distortion.

3.2.2 Model Design

There are two potential directions for model designs: (1) using existing efficient model architectures in the AI field and (2) optimizing communication signals based on existing model architectures. (1) Transformer architectures, such as Swin Transformer and Vision Transformer, can be used to advance extraction for time- and frequency-domain features thanks to their self-attention mechanisms. In this type of implementation, different frequency and time domain scales need to be considered while designing the window division policy in order to adapt to communications data with high spatiotemporal correlation. (2) Existing convolutional neural networks can be optimized for communications scenarios. Specifically, a single traditional convolutional layer is decomposed into group convolution and linear layers, significantly improving the usage of computing resources. To make feature extraction more efficient, a channel split mechanism can be introduced, and various types of convolutional processing (e.g., square, horizontal, and vertical convolutions) can be applied. This type of design not only enriches the feature dimensions, but also retains the key information of the original input through direct mapping. Additionally, multiple types of foundation models can be jointly designed as potential solutions, for example, building convolutional layers in Transformer architectures or building MLP-Mixer in U-Net architectures.

3.2.3 Lightweight Model

Transmitting non-orthogonal superimposed pilots and data poses challenges in terms of the model's storage and computational complexity in practical terminal deployment. This therefore calls for a lightweight model design. To reduce the memory usage of the model, half-precision storage (i.e., parameter quantization) can be utilized, in which model parameters are converted from floating point numbers

to low-bit-width integers. Another way to make models more lightweight is to simplify network architectures. For example, replacing traditional convolutional layers with depthwise separable convolution or optimizing the model connection mode significantly reduces the model's storage and computational complexity while maintaining its ability to process high-dimensional signals. Furthermore, knowledge distillation can be considered. It is a model compression technique that can transfer knowledge from large models to small models. This technique enables a small model to achieve performance comparable to that of a complex model while consuming relatively few resources, making it possible to address challenges such as limited computing resources on the terminal side.

3.3 Multi-Solution Verification

The 5G+AI Research Group has driven research on developing and verifying solutions that can be used to verify the assumptions described in Section 3.1, Topic "Receiver Design for Non-orthogonal Superimposed Pilot and Data" in 2024 6G wireless Communication AI Competition jointly held by IMT-2030 (6G) Promotion Group and IMT-2020 (5G) Promotion Group. The competition results show that many innovative solutions with various data processing approaches, different model architectures, and unique optimization can effectively receive data transmitted through SIP. This demonstrates that SIP with AI receiver can be achieved through diversified solutions, holding immense potential in application.

Table 1 provides system configuration parameters corresponding to two SIP scenarios described in 3.1, fully considering evaluation in different complex scenarios. Scenario 1a involves inter-layer/intra-layer interference in multi-layer transmission, complex channel conditions caused by high UE speed, and generalization at different speeds. Scenario 1b focuses on complexity issues on the terminal side. In contrast, scenario 2 introduces higher inter- and intra-layer interference and the issue of low resilience against interference of higher-order modulation in multi-layer transmission.

During the competition, multiple teams formulated various solutions by combining different potential enhancements. Table 2 shows the average data recovery accuracy (1-BER) of their solutions with 5–35 dB SNR. Despite the tight competition schedule, many teams were still able to offer SIP+AI solutions with acceptable performance.

Table 1 Scenario parameters

Parameter	Scenario 1	Scenario 2
S	624	96
T	12	12
N_t	2	32
N_r	2	4
L	2	4
Modulation scheme	16QAM	64QAM
UE speed	3–120 km/h	3 km/h
Storage complexity	1a: ≤ 100 MB 1b: ≤ 20 MB	≤ 100 MB

Table 2 Accuracy for data recovery

Solution	Scenario 1	Scenario 2
1	0.9393	0.9771
2	0.9390	0.9774
3	0.9391	0.9773
4	0.9386	0.9771
5	0.9386	0.9770
6	0.9386	0.9769
7	0.9384	0.9769
8	0.9384	0.9768
9	0.9382	0.9768
10	0.9390	0.9757

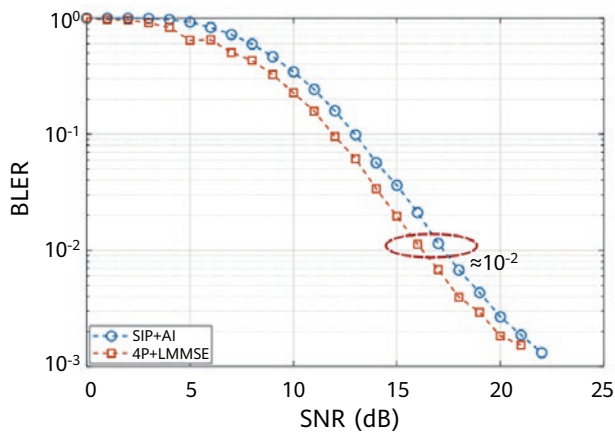


Figure 6 Comparison of lightweight models on BLER performance in scenario 1

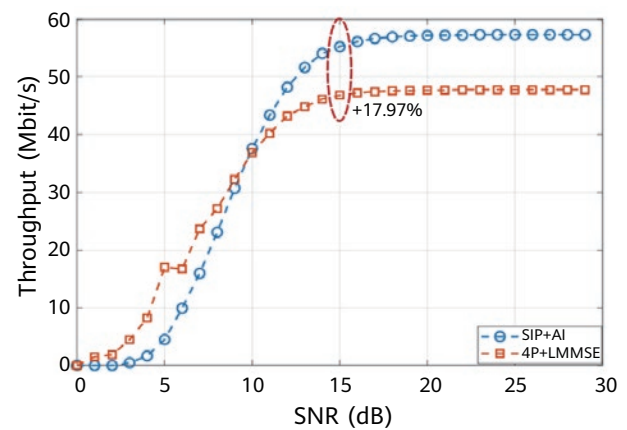


Figure 7 Comparison of lightweight models on throughput performance in scenario 1

Figure 6 compares the SIP+AI solutions with the traditional 4P+LMMSE solutions on BLER performance, showing that the SIP+AI solutions can yield performance comparable to that of the traditional orthogonal solutions in a complex scenario. Figure 7 further compares their throughput performance. In terms of saving resource overheads caused by orthogonal pilot design, the SIP+AI solutions improved throughput performance by 17.97% with respect to an effective SNR range (about 15 dB for 10^{-2} BLER) when compared to traditional solutions. These experiments are conducted using lightweight models with less than 20 MB storage complexity. In multi-party verification, the minimum storage complexity can be reduced to 2 MB with the data recovery precision reaching more than 90%. This means that the complexity requirements can be met in practical terminal deployment.

4 Interference Cancellation-based AI Receiver for SIP

4.1 Design Approaches

As described in scenario 2 in Section 3.3, to address inter-layer interference in multi-layer transmission of SIP, patterns shown in Figure 5 are introduced. Although such special pattern configuration at the transmit end can mitigate the interference to some extent, it will make processes such as model lifecycle management and signaling configuration more complex in practical applications [16]. At the same time, a simple AI receiver at the receive end usually lacks the generalization capability to adapt to different configurations (e.g., modulation schemes and

numbers of layers). It is crucial to design an AI receiver that can further resolve inter-layer interference and have scalability for different modulation schemes and layers, with unified pattern configuration at the transmit end. In [10], an interference cancellation-based AI receiver for SIP is proposed. The proposed solution uses orthogonal superposition code to configure SIP for different layers orthogonally at the transmit end, making it possible to use unified pattern design with low inter-layer pilot interference. This simplifies the transmit end design and achieves full superposition for SIP in multi-layer transmission, suitable for high-speed scenarios. In the AI receiver, a specific interference cancellation structure is introduced to eliminate intra- and inter-layer interference before channel estimation and symbol detection at each layer. Furthermore, a mechanism is designed for this AI receiver to use the same models for concurrently processing different layers. This mechanism enables scalability and low storage complexity of the same AI receiver for different numbers of layers. In addition, an output cropping mechanism is applied to achieve generalization of the proposed AI receiver with different modulation and coding schemes (MCSs).

4.2 Simulation Performance

In this section, adhering to the design approaches outlined in Section 4.1, we provide the simulation results of the SIP+AI solutions with the settings of 4 GHz carrier frequency, 30 kHz subcarrier spacing, and LDPC channel coding. The 4P+LMMSE baseline solutions are used for comparison.

Figure 8 compares performance in high-speed scenarios: 300 km/h and 900 km/h. Our proposed solutions featuring simplified pilot configuration deliver higher throughput gains in multi-layer scenarios. Figure 9 and Figure 10 compare the solutions by their MCS generalization and scalability for different numbers of layers. The mixed solutions using the same models are on par with the specific solutions using specifically trained models for different numbers of layers L and MCSs m . $m = \{3, 7, 14\}$ corresponding to modulation scheme {QPSK, 16QAM, 64QAM} and target bit rate {449/1024, 490/1024, 719/1024}. These experiments validate good generalization and scalability capabilities of the solutions, showing that they can handle challenges in practical deployment.

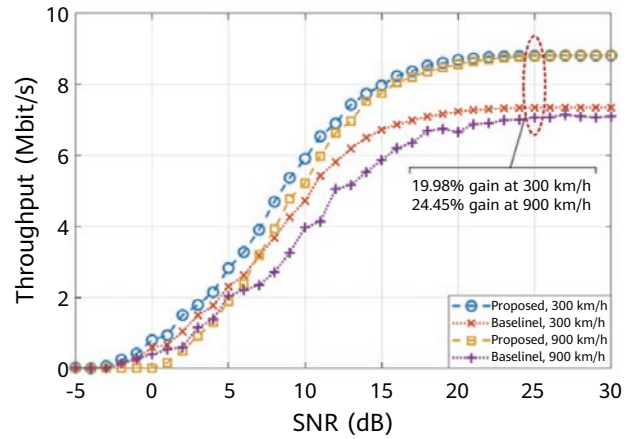


Figure 8 Throughput comparison (CDL-D extension, $N_t = 4$, $N_r = 4$, $L = 2$, $T = 12$, $S = 96$, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$)

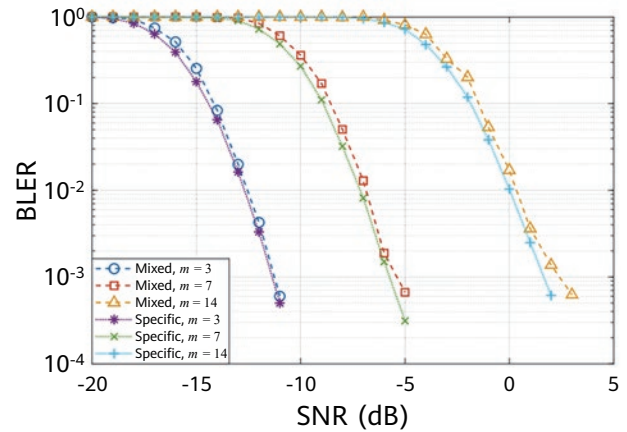


Figure 9 MCS generalization comparison (3 km/h, CDL-C, $N_t = 32$, $N_r = 4$, $L = 2$, $T = 12$, $S = 96$, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$)

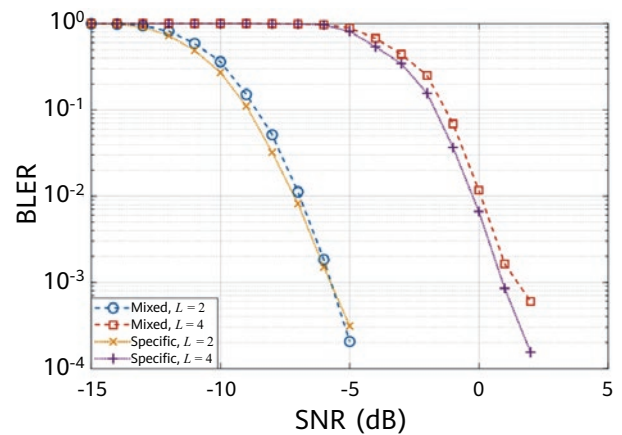


Figure 10 Layer scalability comparison (3 km/h, CDL-C, $N_t = 32$, $N_r = 4$, $L = 2$, $T = 12$, $S = 96$, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$)

5 Conclusion

In this paper, we focus on our studies into non-orthogonal SIP and AI receiver. We first present the overall framework of the SIP+AI solutions and demonstrate their performance advantages through simulation in basic scenarios. Then, we introduce various potential implementations in complex scenarios with multi-party verification. We also propose an interference cancellation-based AI receiver to make our solutions suitable for practical deployment. Finally, we comprehensively show that the SIP+AI solutions can drive potential research, promotion, and standardization of topics related to the relationship between pilot and data resource allocation in future 6G systems. The solutions are expected to contribute to breakthroughs in 6G system design.

References

- [1] Tang H, Yang N, Zhang Z, Du Z, and Shen J, "5G NR and enhancements: From R15 to R16[M]," Elsevier, 2021.
- [2] Ma X and Gao Z, "Data-driven deep learning to design pilot and channel estimator for massive MIMO[J]," IEEE Transactions on Vehicular Technology, 2020, 69(5): 5677–5682.
- [3] Sohrabi F, Attiah K M, and Yu W, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO[J]," IEEE Transactions on Wireless Communications, 2021, 20(7): 4044–4057.
- [4] Chun C J, Kang J M, and Kim I M, "Deep learning-based joint pilot design and channel estimation for multiuser MIMO channels[J]," IEEE Communications Letters, 2019, 23(11): 1999–2003.
- [5] Xu J, Zhu P, Li J, *et al.*, "Deep learning-based pilot design for multi-user distributed massive MIMO systems[J]," IEEE Wireless Communications Letters, 2019, 8(4): 1016–1019.
- [6] Soltani M, Pourahmadi V, and Sheikhzadeh H, "Pilot pattern design for deep learning-based channel estimation in OFDM systems[J]," IEEE Wireless Communications Letters, 2020, 9(12): 2173–2176.
- [7] Mashhadi M B and Gündüz D, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems[J]," IEEE Transactions on Wireless Communications, 2021, 20(10): 6315–6328.
- [8] Wang C X, You X, Gao X, *et al.*, "On the road to 6G: Visions, requirements, key technologies, and testbeds[J]," IEEE Communications Surveys & Tutorials, 2023, 25(2): 905–974.
- [9] Quy V K, Chehri A, Quy N M, *et al.*, "Innovative trends in the 6G era: A comprehensive survey of architecture, applications, technologies, and challenges[J]," IEEE Access, 2023, 11: 39824–39844.
- [10] Xiao H, Tian W, Jin S, *et al.*, "Interference cancellation based neural receiver for superimposed pilot in multi-layer transmission[J]," arXiv preprint arXiv:2406.18993, 2024.
- [11] Aoudia F A and Hoydis J, "End-to-end learning for OFDM: From neural receivers to pilotless communication[J]," IEEE Transactions on Wireless Communications, 2021, 21(2): 1049–1063.
- [12] Jing X, Li M, Liu H, *et al.*, "Superimposed pilot optimization design and channel estimation for multiuser massive MIMO systems[J]," IEEE Transactions on Vehicular Technology, 2018, 67(12): 11818–11832.
- [13] Ma J, Liang C, Xu C, *et al.*, "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems[J]," IEEE Journal on Selected Areas in Communications, 2017, 35(12): 2696–2707.
- [14] Hoehner P and Tufvesson F, "Channel estimation with superimposed pilot sequence[C]," Seamless Interconnection for Universal Services. Global Telecommunications Conference. GLOBECOM'99. (Cat. No. 99CH37042). IEEE, 1999, 4: 2162–2166.
- [15] Ye H, Li G Y, and Juang B H, "Deep learning based end-to-end wireless communication systems without pilots[J]," IEEE Trans. Cogn. Commun. Netw., 2021, 7(3): 702–714.
- [16] Chen W, Lin X, Lee J, *et al.*, "5G-advanced toward 6G: Past, present, and future[J]," IEEE Journal on Selected Areas in Communications, 2023, 41(6): 1592–1619.



Semantic Digital Twins: Enhancing Performance in Wireless Communication and LLM Inference

Peiyao Chen, Yiqun Ge, Qifan Zhang, Wuxian Shi, Zheyuan Wei
Ottawa Wireless Advanced System Competency Centre

Abstract

This paper presents a novel technique that combines integrated sensing and communication (ISAC) with large language models (LLMs) to extract features from sensor data, with the goal of creating a semantic digital twin (SDT). The proposed SDT is a dynamic collection of semantic tokens that are updated using a variant feature clustering method, where the same cluster represents identical events or objects. Unlike traditional digital twins, this approach incorporates spatiotemporal aspects by assimilating historical data, which enhances the performance of wireless communication systems and LLM inference. By combining ISAC principles with advanced language processing capabilities, the methodology goes beyond real-time replication of the physical environment, providing a comprehensive understanding of system behavior. In wireless communication, SDT enables precise beamforming and personalized user localization, enhancing system flexibility and user experience. In AI inference, SDT enhances LLM inference results by enabling precise visual cropping, context-aware prediction, and effective prompt engineering, thereby supporting applications in intelligent medical diagnosis and transportation. Overall, this combined approach offers significant opportunities for performance improvement in both wireless communication and language understanding domains, facilitating the development of more intelligent systems.

Keywords

ISAC, LLM, SDT

1 Introduction

The development of artificial intelligence (AI), represented by large language models (LLMs), has opened doors to exciting possibilities across various industries such as manufacturing, healthcare, and transportation. From intelligent management of production lines to automated control of transportation systems and precise prediction in medical diagnostics, the applications of LLMs are gradually reshaping our lives and work, offering a more efficient, safer, and healthier future for society. LLMs possess the capability to understand and process contextual information, which enables them to generate coherent and meaningful responses, mimicking human-like communication.

With the advancement of wireless communication technology, 6G wireless systems are evolving toward higher frequencies (including millimeter-wave and even terahertz), wider bandwidths, and larger-scale antenna arrays [1]. On the one hand, communication systems are acquiring capabilities similar to sensing systems, utilizing widely-covered mobile communication signals to extract distance, angle, material, and other information from radio waves through analysis of direct, reflected, and scattered signals [2], thereby achieving a perception of target objects or environmental attributes and states. On the other hand, sensing technologies, through high-precision positioning, environment reconstruction, etc., provide real-time replication of the physical world, i.e., constructing a parallel digital world, known as a "digital twin". The digital twin helps enhance communication performance, for example, through precise beamforming and efficient channel state information (CSI) detection. Integrated sensing and communication (ISAC) is expected to become a trend [3, 4]. Additionally, the combination of LLMs with sensing technologies endows devices such as sensors and cameras with intelligent perception capabilities. They not only identify, detect, and collect vast as well as diverse data but also possess the ability to analyze and optimize the data, enabling them to perceive and understand the external environment. The future will witness the integration of ISAC with LLMs, jointly driving the era of 6G "Artificial Intelligence of Things (AIoT)" [5].

In this vision, there will be a significant portion of communication between sensors, robots, and other intelligent devices in future communication systems, particularly in the context of 6G [6]. The emergence of LLMs enables intuitive and efficient communication between

humans and machines, as well as among machines, advancing the concepts of semantic communication — which enhances efficiency by reducing the volume of data transmission and focusing on conveying meaning rather than just raw data — in the realm of 6G research. Through LLMs, information from various modalities (such as images, audio, and point clouds) can be extracted and transformed into a common tokenized representation. These discrete tokens extracted from the LLM vocabulary encapsulate the semantics of underlying data, regardless of their original modalities. This offers exciting possibilities for seamless communication and information exchange among different devices and systems. Additionally, this token-based semantic communication method makes it easier to integrate information into knowledge graphs and other semantic representation frameworks, facilitating decision-making based on a comprehensive understanding of the environment. With context-aware communication, devices can dynamically adjust their behavior based on the surrounding environment and the overall system goals.

To realize the vision of efficient communication and AIoT, the demand for LLMs will extend beyond human users to encompass a vast network of IoT devices. However, performing LLM inference directly on these devices is often impractical due to their limited computational capabilities. Traditional cloud-based solutions, because of the round-trip communication between users and distant data centers, introduce significant delays, hindering real-time applications that demand instant responses. This is particularly problematic for time-sensitive tasks such as autonomous driving or industrial control, where milliseconds are crucial. This challenge necessitates the need for online LLM inference facilitated by advanced wireless systems, particularly at the base station (BS) level in the context of the upcoming 6G era.

Thus, in future wireless communication systems, in addition to managing and controlling wireless communication and providing connectivity and communication services to user devices, BSs also serve as central hubs for various AI models, each of which is designed for specific functionalities and applications. These pre-trained and validated models are strategically allocated to BSs within the core network, bringing AI capabilities closer to end users. Figure 1 shows such a future system.

This paper contributes to advancing wireless communication and AI inference efficiency by integrating ISAC with LLMs to establish a semantic digital twin (SDT). The rest of the

paper is organized as follows. Section 2 introduces the detailed framework of SDT in wireless communication systems, focusing on the integration of ISAC and LLMs. Section 3 explores the applications of SDT in wireless communication and enhancing AI inference capabilities. Section 4 concludes this paper.

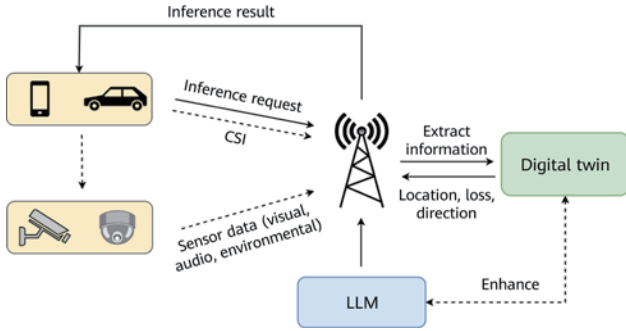


Figure 1 Integrating ISAC and LLM in a 6G system

2 Semantic Digital Twin

The concept of digital twin is revolutionizing our understanding and management of complex systems. Digital twins enable network operators to optimize performance by identifying coverage gaps, mitigating interference, and efficiently allocating resources. As we move toward 6G and beyond, digital twins are becoming essential for creating virtual replicas of physical wireless environments, encompassing everything from BSs and user devices to the surrounding terrains. These digital representations are constantly updated with real-time data, facilitating continuous monitoring, analysis, and prediction of system behavior.

In the vision of efficient communication, the integration of semantic communication and digital twin technology presents a compelling prospect for the future of 6G intelligent systems, namely, the tokenized representation of the real-time physical cellular world. Sensors and wireless-related characteristics with tokenized representation, such as CSI and channel quality indicator (CQI), play a crucial role in building an accurate and effective SDT in this context.

2.1 Semantic Sensor Data

The presence of LLMs enables sensors to comprehend specific tasks or objectives assigned to them when processing raw data. LLMs enhance efficiency because

they allow sensors to focus attention and processing capabilities on relevant aspects and extract more meaningful information for specific scenes and tasks. This information will contain both semantic concepts related to tasks and additional attributes of objectives, and will be represented by the semantic token T^s . For instance, a camera does much more than display basic image pixels. It detects individuals performing specific actions within a particular scene and encodes additional attributes such as their location and movement. Similarly, environmental sensors can convey semantic concepts such as "comfortable", "humid", or "polluted" based on predefined thresholds and environmental models, in addition to reporting temperature values, and provide alerts indicating whether the environment is abnormal.

2.2 Tokenized Radio Channel Measurement

Given that the entire communication network functions as a vast sensor, wireless-related characteristics can significantly enhance the perception and comprehension of the physical world. This is achieved by extracting distance, velocity, and angle information from wireless signals. For example, analysis of CSI can reveal the presence of obstacles, identify different types of interference (e.g., co-channel interference or external sources), and detect the movement of objects within the coverage area. These insights can be encoded as tokens T^c such as "obstacle", "interference", or "movement" along with relevant parameters like location, loss, or direction. Furthermore, if BSs possess a comprehensive RF map of the environment and sufficient computational power, they can potentially reconstruct the coarse but complete CSI from these tokenized representations. This reconstruction ability facilitates a more efficient and compact representation of CSI, reducing the amount of data that needs to be transmitted while preserving essential information about the wireless environment.

2.3 Semantic Digital Twin Representation

In this paradigm (illustrated in Figure 2), the semantic digital twin becomes a dynamic collection of semantic tokens, continuously updated with information from various sensors and radio channel measurements. Each sequence

of tokens, representing a specific aspect or event within the environment, carries not only its inherent semantic meaning but also temporal and spatial context. Every piece of information within the digital twin is tagged with a timestamp and location stamp, creating a three-dimensional representation of the environment that encompasses time, space, and semantics (for event descriptions). This enriched digital twin transcends the role of a passive data collector and becomes an active participant in understanding and interpreting the environment.

Such a semantic digital twin is established by BSs. During the establishment, the main challenge lies in token fusion at each timestamp. Assume that the lengths of tokens T^s and T^c are equal, or an additional neural network projection will be employed to align their lengths. Inspired by [7], we divide tokens $T = \{T^s, T^c\}$ into a certain number of clusters using token features, and then fuse the tokens in the same feature cluster, as shown in Figure 3. Note that tokens within the same feature cluster correspond to identical

events or objects, and the number of fused tokens varies across different feature clusters. The clustering method used in the paper is based on a hybrid feature clustering method using semantic tokens. It comprises two main parts: token KNN, which focuses on clustering based on spatial similarity of features, and token fusion, which utilizes large-scale models to consider semantic similarity. During model training, multiple semantic token attributes belonging to the same target or event are aggregated into a cluster using semantic graphs.

- **Feature cluster:** A variant of the "density peaks clustering based on k -nearest neighbors (DPC-KNN)" algorithm [8] is used to create a feature cluster. Since the cluster centers are distinguished by their higher density compared to neighboring tokens as well as their relatively large distance from tokens with higher densities, both density ρ and relative distance δ should be considered. Given a set of tokens T , let $NN_k(t_i)$ be the k -th nearest token to t_i according to semantic similarity. The k -nearest neighbors $KNN(t_i)$ of t_i is defined as:

$$KNN(t_i) = \left\{ j \in T \mid \frac{t_i \cdot t_j}{\|t_i\| \|t_j\|} \leq \frac{t_i \cdot NN_k(t_i)}{\|t_i\| \|NN_k(t_i)\|} \right\}. \quad (1)$$

Then, the local density ρ_i of token t_i is obtained by calculating the mean distance to k nearest neighbors:

$$\rho_i = \exp \left(-\frac{1}{k} \sum_{t_j \in KNN(t_i)} \frac{t_i \cdot t_j}{\|t_i\| \|t_j\|} \right). \quad (2)$$

The relative distance is calculated by:

$$\delta_i = \begin{cases} \min_{j: \rho_j > \rho_i} \frac{t_i \cdot t_j}{\|t_i\| \|t_j\|}, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i \\ \max_j \frac{t_i \cdot t_j}{\|t_i\| \|t_j\|}, & \text{otherwise} \end{cases}, \quad (3)$$

where ρ_i is the local density of token t_i .

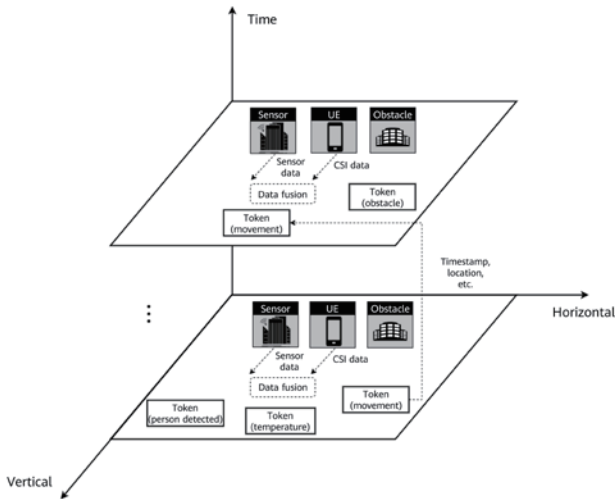


Figure 2 Semantic digital twin

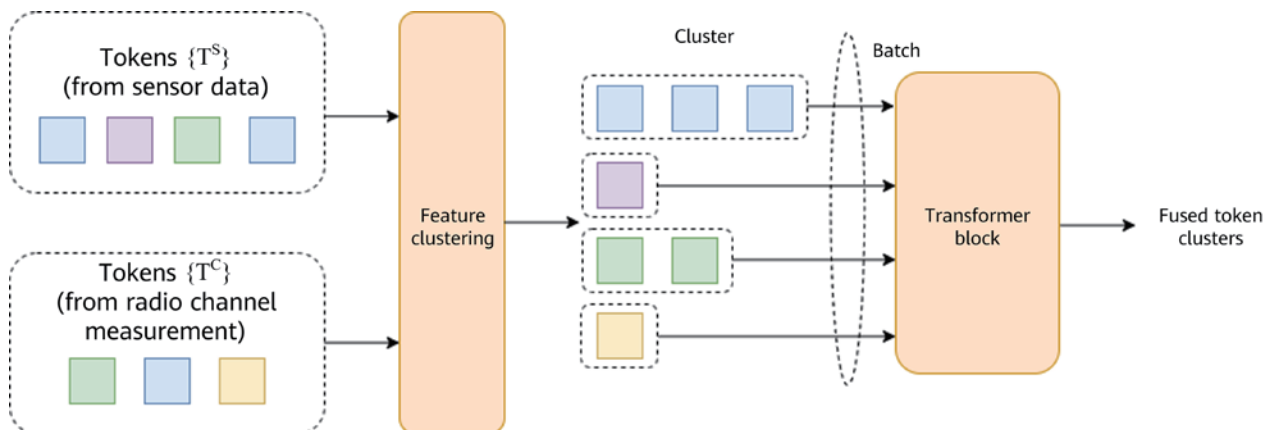


Figure 3 Token fusion process

Let $s_i = \rho_i \times \delta_i, i \in \{1, \dots, |T|\}$ denote the token score for each token t_i . Then, the cluster centers are determined by selecting the tokens with the highest scores s_i , and other tokens are then assigned to the nearest cluster center based on the semantic distances.

- **Token fusion:** A transformer block is applied to each feature cluster to capture the semantic relationships and information interaction between different tokens in the same feature cluster, resulting in fused token clusters \tilde{T}_n .

For feature clusters at different timestamps, pairing is based on similarity distances, meaning that clusters will only match if the similarity distance between their centers is less than the given threshold d_c . When making decisions or similar tasks, all corresponding feature clusters spanning across time and space are considered, enhancing accuracy and opening up new possibilities for more harmonious interaction between the physical and digital worlds.

3 Applications of Semantic Digital Twins

The spatiotemporal SDT plays a crucial role in both wireless communication and LLM inference.

3.1 In Wireless Communication

By analyzing and integrating historical and real-time data, SDT can help optimize resource allocation and signal processing. Specifically, in technologies like beamforming, it precisely locates signal transmission directions to maximize signal reception efficiency.

In traditional beamforming techniques, directional transmission typically relies on the geographical position of

devices or specific signal sources. However, through SDT, the system can recognize and understand specific user activities or states, such as identifying a user's posture or behavior while reading a book. This personalized localization transcends rigid geographical boundaries and signal sources, focusing instead on user behavior and needs. Based on this information, the system adjusts the beamforming direction of the antenna array to precisely target specific user devices. Furthermore, the system can quickly respond to changes in user posture or environmental conditions, dynamically adjusting the beam direction to maintain communication continuity and efficiency. These capabilities enhance the flexibility and adaptability of communication systems and significantly improve user experience and service quality.

Figure 4 presents a real-time demonstration of the SDT detecting a person holding a book. In our demonstration, multiple types of sensing devices are used, such as cameras and lasers. To align the data collected by these diverse sensing devices, the token fusion method proposed in Section 2.3 is employed, enabling feature extraction and matching of targets and objects across multiple devices. The detection is divided into environmental detection and semantic detection. The former involves detecting static objects that current LLMs can handle effortlessly. The latter refers to understanding and detecting human actions, which requires analyzing and integrating the relative positions and states of the target individual and surrounding objects. In the demonstration, we maintain two queues: one for semantic states $S(p)$, and the other for the relative positions $L(p, o)$ of individuals and objects corresponding to these states, where p refers to the index of detected persons and o denotes the index of detected objects. Subsequently, contrastive learning between semantic states and relative positions is employed to enhance the precision of detecting and understanding human postures. The entire process is depicted in Figure 5.

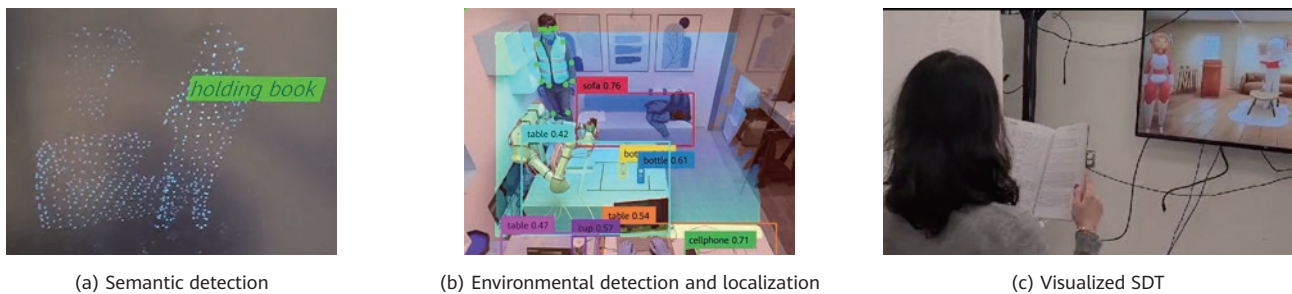


Figure 4 Demonstration of SDT with a person holding a book

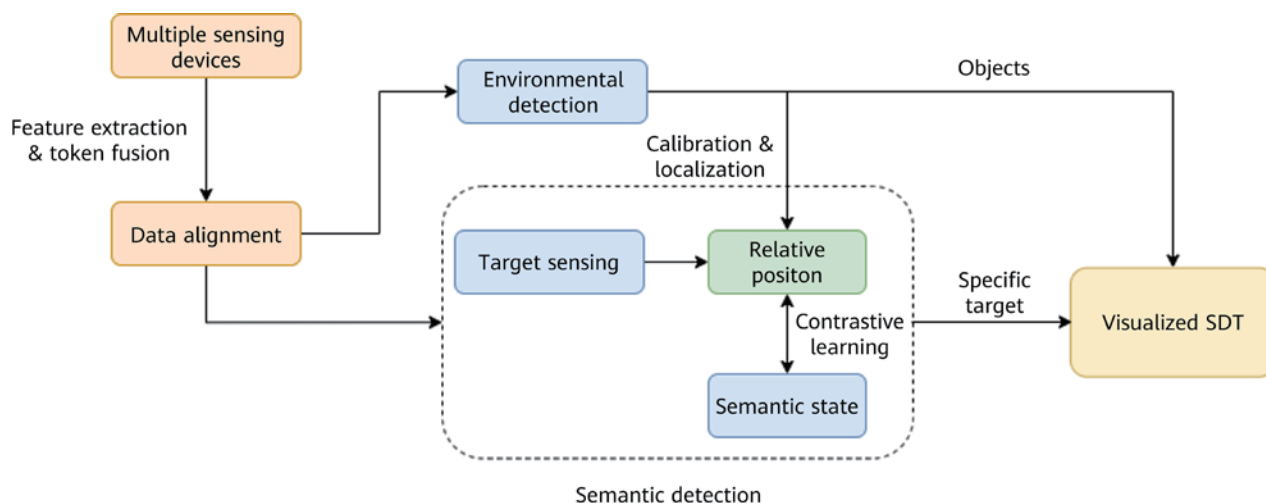


Figure 5 Process of SDT construction

3.2 In Enhancing AI Inference

SDT can significantly enhance AI inference in several key aspects:

- Precise visual cropping:** Effective performance in visual question answering (VQA) tasks using multimodal LLMs is crucial for applications in medical diagnosis and intelligent transportation. As introduced in [9], the size of the visual subject in the question significantly affects model sensitivity. Larger visual subjects tend to improve accuracy in related question answering. Conversely, smaller or blurry details often challenge models, impairing their ability to process subtle visual cues effectively. Therefore, precise image cropping, which enables models to focus on critical visual regions, notably enhances accuracy and efficiency in VQA tasks. Unlike conventional methods (e.g., [9]) that focus on single-image cropping, SDT provides a global view of the environment through token representation, enabling more accurate cropping.
- Context-aware prediction:** Current visual LLMs are optimized for single-image tasks and lack temporal memory. Direct training of video LLMs is resource-intensive due to the voluminous nature of video data. Tasks involving actions like "pick up" and "put down" require contextual information for accurate interpretation that single-frame analysis alone may not provide. Enhancing inference tasks, especially those predicting regular actions or scenes, can benefit from SDT's spatiotemporal knowledge
- Effective prompt engineering:** SDT can assist in improving and optimizing prompt engineering for LLMs by analyzing and understanding past language data. This refinement enables the inference engine to make more informed and contextually relevant decisions. Consider a scenario where a robot is tasked with retrieving food. If the robot solely relies on its onboard sensors, its capabilities are inherently limited to its immediate surroundings, lacking access to historical environmental information. In such cases, if no visible food items are nearby, the robot might fail to complete its task. However, integrating the SDT's spatiotemporal awareness into the inference process expands the robot's perception beyond its immediate environment. The SDT's collective memory function provides insights into past events and the environment's history, filling knowledge gaps for the robot. For instance, its prompt may contain information about the location of food items in a particular drawer, even if the robot cannot directly observe them. Armed with this background knowledge, the inference engine can guide the robot effectively, enabling it to successfully retrieve the desired food. This example underscores the transformative impact of integrating SDT technology with robotic inference, enhancing robots' intelligence and adaptability in complex environments.

4 Conclusion

This paper introduces a novel approach integrating ISAC with LLMs to establish an SDT, where semantic tokens represent sensor data from sensing devices and radio channel measurements. These tokens are fused according to their feature clusters. By assimilating historical data, SDT enhances wireless communication performance, particularly in providing precise beamforming and personalized user localization. Additionally, it enhances the precision and efficiency of AI inference tasks through accurate visual cropping, context-aware prediction, and effective prompt engineering. This integrated method holds significant promise for advancing intelligent systems in both domains. Future research can further explore SDT's potential across diverse applications, including autonomous driving, smart manufacturing, and environmental monitoring, achieving comprehensive deployment and advancement of IoT technologies.

References

- [1] Wen Tong and Peiyong Zhu, "6G: The next horizon – From connected people and things to connected intelligence," Cambridge University Press, May 2021.
- [2] Zhi Zhou, Xianjin Li, Jia He, *et al.*, "6G integrated sensing and communication - sensing assisted environmental reconstruction and communication," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023: 1–5.
- [3] Danny Kai Pin Tan, Jia He, Yanchun Li, *et al.*, "Integrated sensing and communication in 6G: Motivations, use cases, requirements, challenges and future directions," in Proceedings of 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S). 2021: 1–6.
- [4] Alireza Bayesteh, Jia He, Yan Chen, *et al.*, "Integrated sensing and communication (ISAC) — From concept to practice," Communications of Huawei Research, 2022: 4–25.
- [5] Jing Zhang and Dacheng Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," IEEE Internet of Things Journal, 2020, 8: 7789–7817.
- [6] one6G, whitepaper, "6G & robotics, use cases and potential service requirements," June 2023. Available online: <https://one6g.org/resources/publications/>
- [7] Wang Zeng, Sheng Jin, Wentao Liu, *et al.*, "Not all tokens are equal: Human-centric visual analysis via token clustering transformer," in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11101–11111.
- [8] Mingjing Du, Shifei Ding, and Hongjie Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," Knowledge-Based Systems, 2016, 99: 135–145.
- [9] Jiarui Zhang, Mahyar Khayatkhoe, Prateek Chhikara, and Filip Ilievski, "Visual cropping improves zero-shot question answering of multimodal large language models," arXiv preprint arXiv:2310.16033, 2023.



Digital Twin Online Channel Model: Vision, Progress, and Challenges

Junling Li ^{1,2}, Weitian Zhang ¹, Chengxiang Wang ^{1,2}, Chen Huang ^{2,1}

¹ National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University

² Pervasive Communication Research Center, Purple Mountain Laboratories

Abstract

This paper introduces the digital twin online channel model (DТОСМ), a novel approach that synchronizes the performance of digital and physical networks, providing continuous visualization and accurate prediction of dynamic channel changes. Unlike traditional offline channel models, DТОСМ offers real-time sensing and accurate representation of dynamic wireless channels, enabling efficient optimization of 6G networks. We start this paper by tracing the development path of DТОСМ, highlighting its vision and key challenges. Next, we delve into the underlying principles, mechanics, and practical applications of DТОСМ in typical 6G environments. Then, we demonstrate the DТОСМ platform's ability to provide and visualize real-time channel information. Finally, we outline future research directions and the challenges that need to be addressed in promoting the application of DТОСМ.

Keywords

digital twin, 6G online channel modeling, channel map, environment sensing, machine learning

1 Introduction

The 6G vision has been summarized as "global coverage, all spectra, full applications, all senses, all digital, and strong security" [1]. Conventional network optimization methods are often costly, time-consuming, and prone to errors, relying on techniques such as channel measurement, trial and error, and engineering expertise. Furthermore, real-world wireless networks are characterized by numerous adjustable parameters, which can render nonlinear predictions of network performance and user traffic extremely complex. As a result, offline optimization methods fall short of expectations, with most networks typically operating at around 60% of their potential, leaving significant room for improvement [2]. To successfully deploy and optimize 6G networks, it is essential to have a precise online channel model that can accurately reflect real-world physical environments and enable real-time network optimization. However, testing communication systems on live networks poses a significant challenge because the performance often differs significantly from lab simulations. The main reasons for this discrepancy are: 1) Live network tests often lack channel measurements, and the channel models used in lab simulations fail to accurately mirror the actual test environment. 2) Even if a suitable channel model is used, its parameters are often static, making it difficult to adapt to the dynamic and time-varying conditions encountered on live networks. Consequently, channel simulation models struggle to keep pace with real-time test scenarios [3, 4]. In light of these challenges, there is a pressing need for a digital twin online channel model (DTOCM) that can accurately capture the characteristics of live networks [16].

DTOCM enables real-time monitoring and analysis of channel conditions, offering visualization and prediction of dynamic channel changes [5]. As 6G networks begin to support hybrid services, immersive communication, and dynamic environments, DTOCM realizes low-cost, low-complexity, and high-accuracy channel information acquisition. This, in turn, can help mitigate the issue of large-scale pilot overheads, reduce estimation errors, and ultimately improve overall network performance [6]. The viability of DTOCM hinges on two crucial factors. First, the close connection between the physical environment and channel characteristics can be effectively captured through physical positioning and environment characteristics. Such an inherent relationship suggests that similar physical locations and environment characteristics often lead to similar channel characteristics, allowing a significant

amount of data generated by the base station to be reused. This reusability plays a significant role in constructing an accurate channel digital twin model. Second, as communications networks expand and frequency bands become more diverse, advancements in positioning and sensing precision have improved the capabilities of precise positioning and environment cognition. These enhanced capabilities enable digital representations of wireless channels to be more detailed and accurate, aligning perfectly with the objectives of DTOCM.

In this paper, we introduce the DTOCM, a novel approach that leverages physical environment sensors and multiple sensing methods to dynamically monitor the environment in real-time. Such an approach enables a mapping to be established between actual communication environment parameters, channel parameters, and channel characteristics. By integrating environment sensing with channel modeling, DTOCM enhances the prediction accuracy of changes in the wireless signal propagation environment. This, in turn, enables more accurate simulation of real-world wireless propagation scenarios, providing robust support for communication network optimization and decision-making.

This paper is organized into six sections. Section 2 will delve into the benefits and vision of DTOCM. Section 3 will provide a detailed introduction to the proposed DTOCM framework, covering its overall working principles, a three-step construction mechanism, and typical applications in 6G communication scenarios. Section 4 will describe the real-time provision and visualization of channel information enabled by the proposed DTOCM framework. Section 5 will discuss some of the open research issues related to DTOCM. And finally, Section 6 will summarize the main points of this paper.

2. DTOCM's Advantages and Vision

2.1 Advantages

The DTOCM proposed in this paper has four key advantages:

- Reduced pilot overhead:** By preloading basic channel state information (CSI), devices can quickly obtain channel parameters as they move through a continuous 3D space. This can be achieved through simple database lookups or calculations based on location, orientation, and antenna parameters. By leveraging DTOCM, communication network designers can overcome channel research challenges, significantly reducing pilot overheads and ultimately enhancing 6G network performance.

- Real-time CSI provision and prediction:** DTOCM boasts robust real-time channel information provision and prediction capabilities, enabling it to adapt seamlessly to dynamic network conditions. By extracting and predicting channel information in real-time, DTOCM ensures that the 6G communication network remains responsive to changing environmental conditions. This capability also helps to calibrate the performance between channel model simulations and actual environment conditions, thereby enhancing the reliability and robustness of network performance.
- Channel information visualization:** DTOCM offers powerful visualization of channel information, enabling the prediction and highlighting of changes in the communication environment. Through this visualization, DTOCM displays node locations and motion status in real-time, and provides real-time monitoring of channel characteristics and other related information. In this way, DTOCM significantly enhances environment sensing and decision-making capabilities in future 6G networks.
- Network performance optimization:** DTOCM allows real-world scenarios to be simulated in a virtual environment, providing critical sensing and feedback that can be used to optimize the deployment of communication networks in the real world. This approach enables comprehensive testing in a controlled virtual environment, facilitating rapid iteration and improvement. And by leveraging the insights gained through simulation, DTOCM ensures a balance can be struck between performance and efficiency during the optimization of future 6G networks.

2.2 Vision

DTOCM captures real-time CSI changes in dynamic environments, displays the location and motion status of entities and nodes, and outputs channel large-/small-scale fading information, channel characteristics, and other relevant data in real-time. These capabilities, as depicted in Figure 1(a), enable real-time accurate predictions and support informed decision-making to optimize communication network performance. Sensors in the physical space monitor the environment in real-time, mapping actual conditions to the digital twin's channel parameters and characteristics. When combined with AI algorithms, DTOCM can predict changes in the wireless propagation environment across space, time, and frequency domains.

This vision is further illustrated in Figure 1(b), which depicts three typical application scenarios of the 6G network. In the first scenario, the transmitter and receiver are stationary, but the scattering environment changes dynamically. DTOCM displays real-time CSI, including amplitude, phase, delay, Doppler, and angle. In the second scenario, the transmitter and receiver are in motion within a dynamic environment. DTOCM provides real-time CSI and tracks the location and motion of each node in real-time. And in the third scenario, when a mobile entity such as an unmanned aerial vehicle (UAV) is in motion, DTOCM uses channel information changes to deduce the location of the UAV.

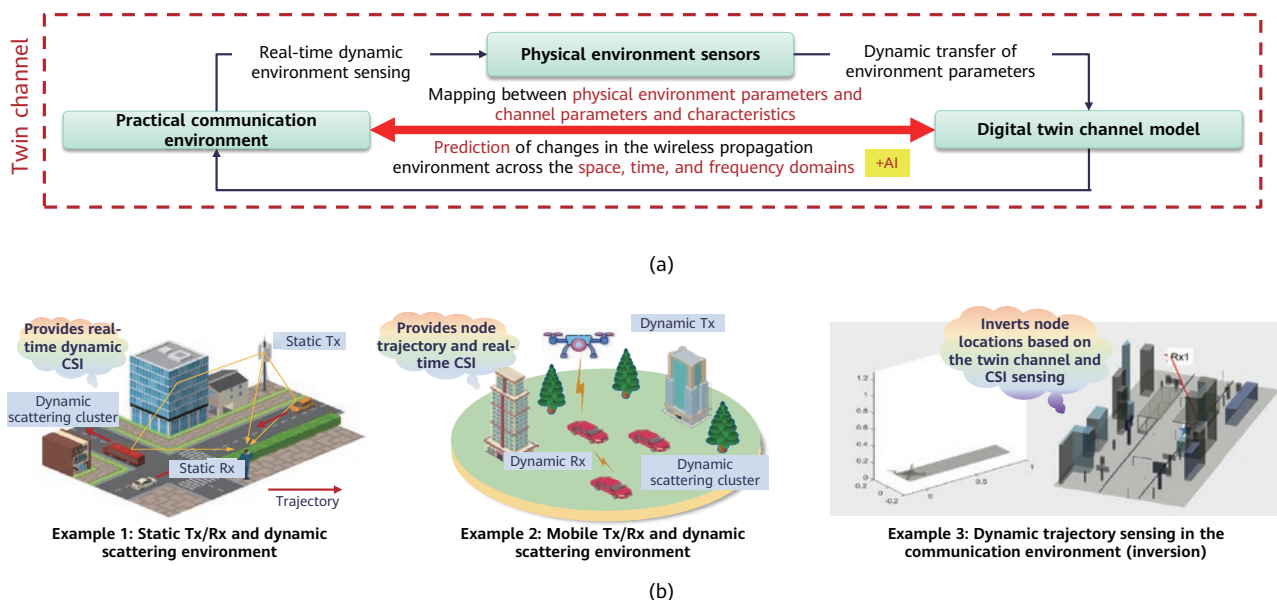


Figure 1 DTOCM's (a) vision and (b) typical application scenarios

3 DTOCM Framework

3.1 Overall Framework

DTOCM builds on a comprehensive and exclusive classification of 6G scenarios and uses real-time sensing of environmental information to identify specific communication scenarios. It then pairs these scenarios with the model parameters of the 6G pervasive channel model (6GPCM) [7, 8] to implement a data-driven approach for identifying communication scenarios and a model-driven approach for all-scenario channel simulation.

The process begins with the collection of environment and channel measurement data at the physical network layer. This data is then used to reconstruct the environment and create a digital twin layer. Within this digital twin layer, ray tracing (RT) and geometric-based stochastic modeling (GBSM) are used to generate a wireless channel map, effectively creating a virtual representation of the physical environment and its corresponding channel characteristics. The wireless channel map serves as a coarse model for AI prediction. Using a known temporal channel database, the AI model then makes accurate and effective predictions across the space, time, and frequency domains [9]. These predictions fill any gaps in the wireless channel map, enabling the model to handle unknown scenarios, future time periods, and unfamiliar frequency bands. This process is performed to create a relatively complete digital twin channel map, which can be utilized for network parameter optimization and environment information inversion at the physical network layer. Furthermore, the physical network layer can provide the digital twin channel map with the necessary parameters or data required for network performance calibration.

3.2 Machine Learning–based Scenario Identification

The classification of communication scenarios in the current 5G standard channel model is limited and coarse, meaning it cannot accurately describe the diverse and complex communication scenarios of the future 6G network. In particular, the vision of global coverage, all spectra, and full applications for 6G networks requires a more detailed and comprehensive classification of wireless communication scenarios.

To build a real-time and accurate DTOCM (Figure 2), it is essential to classify 6G wireless communication scenarios in detail and pair them with the corresponding channel model parameters. This involves identifying and modeling communication scenarios by using a data-driven machine learning method based on the environment sensing data. The process encompasses the following steps:

1. To enable comprehensive communication scenario identification, real-time environment sensing is achieved by leveraging various data sources, including electronic maps, remote sensing images, and point cloud data [10, 11]. These data sources provide valuable information that is used to extract environment parameters, which are essential for accurate channel modeling.
2. Next, the extracted environment parameters are processed using a machine learning model (e.g., a neural network) to classify communication scenarios into different categories, such as aeronautical, aerospace, terrestrial, and marine communication. The communication scenarios classified in this step are further subdivided into more specific environments, including satellites, space stations, aircraft, UAVs, indoor environments, cities, seas, and islands. A neural network–based classification network plays a crucial role in this step, enabling a more refined understanding of various environments.
3. The final step is to identify different communication scenarios and match them with the corresponding channel model parameters of the 6GPCM [7, 8]. This involves mapping the appropriate channel model parameters to each scenario, ensuring that accurate multi-scenario channel measurement can be performed while also maintaining compatibility with the 6G channel model.

3.3 Offline Channel Map Initialization and Environment Reconstruction

After the communication scenario is identified, the next step is to reconstruct the digital twin environment and initialize the offline channel graph. To accomplish this, multimodal environment sensing techniques (e.g., images, videos, 3D electronic maps, and point clouds) are used for 3D scene reconstruction. The reconstructed virtual scene is then imported into RT software for the simulation of channel characteristics. By comparing the differences between the RT simulation results and actual measurement data, the

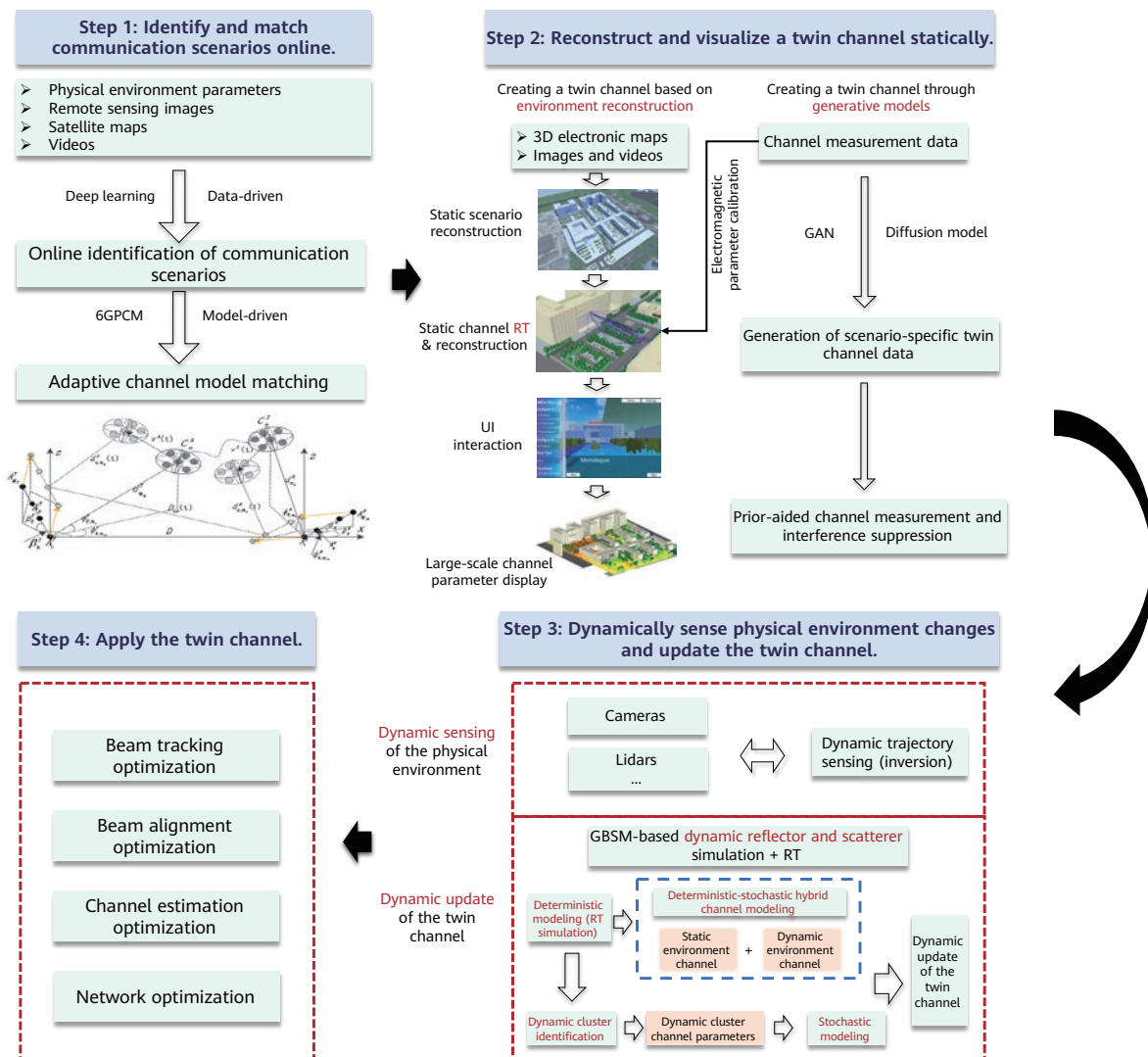


Figure 2 Process of building a DTOCM

electromagnetic coefficients used in the simulation can be calibrated to improve the accuracy of the RT channel reconstruction. In addition to the virtual environment, the digital twin layer also contains virtual data, or "twin data." Generative models can be used to create virtual environments for different scenarios from measurement data, providing prior support for channel measurement and interference suppression.

3.4 Environment Sensing and Dynamic Update of the Twin Channel

The final step is to implement real-time updates of dynamic environment channel characteristics. To achieve this, physical environment sensors such as cameras and lidars are used to monitor dynamic changes in the environment in real-time, for example, monitoring vehicle and pedestrian movements.

These changes are then reflected in the twin environment, and the corresponding channel characteristics are re-simulated. DTOCM employs a novel static-dynamic hybrid channel modeling algorithm to update dynamic channel information [12]. For static objects in the environment, RT modeling is used for simulation, as described in Section 3.3. However, for dynamic objects, a different approach is taken. Specifically, dynamic cluster parameters are extracted from measurement data, and GBSM is used to simulate the dynamic scattering channel.

3.5 Applications of DTOCM

Twin channels, constructed through the steps outlined earlier, can be used in various application scenarios to enhance the capabilities of a 6G communication network. Using the digital twin channel offers a number of key

benefits, including beam alignment, beam tracking, channel estimation, and network optimization [13].

- **Beam alignment optimization:** By leveraging the channel information obtained through DTOCM, the network can select beams that offer the best performance. This reduces the effort required to align the beams while also improving accuracy, ultimately delivering the best possible signal quality. Additionally, DTOCM pre-selects the best beams based on relevant metrics such as power spectral density, thereby reducing pilot overheads and channel estimation complexity.
- **Beam tracking optimization:** DTOCM improves the speed and accuracy of beam tracking by generating channel statistics within an optimized detection beam range. This is achieved through the Sparse Bayesian Learning approach, which leverages channel sparsity to select beams with the maximum expected signal-to-noise ratio. This facilitates fast beam alignment and efficient tracking, which is especially critical in responding to dynamic environmental changes.
- **Channel estimation optimization:** DTOCM enables optimized channel estimation by guiding pilot design and pre-acquiring channel information in advance, which significantly reduces channel estimation overheads. The use of preliminary data of the channel map simplifies the channel estimation process and improves the overall efficiency of the 6G communication network.
- **Network optimization:** Unlike traditional channel modeling, DTOCM seeks to predict channel characteristics in unknown environments, at future times, and across unknown frequency bands — it goes beyond simply representing known channel characteristics in specific scenarios, historical times, and known frequency bands. By analyzing sensing data from the wireless propagation channel, DTOCM can predict future channel conditions, which are then used to fine-tune network parameters and optimize communication performance based on the latest CSI.

4 DTOCM's Visualization Platform

The DTOCM framework we propose in this paper enables the extraction and visualization of real-time channel information on 6G networks, making it possible to implement network performance analysis and intelligent network resource scheduling. This is achieved by reconstructing the scene environment, where static

scenes are modeled using Blender to capture the physical characteristics of the "wireless valley," while dynamic scenes involving moving objects (e.g., UAVs and pedestrians) are developed using Unity. The reconstruction platform enables interactive simulation, enhancing the authenticity and applicability of the model by allowing users to select and modify the locations of the transmitter (Tx) and receiver (Rx) using mouse clicks. The system outputs detailed channel characteristics (including CSI, path loss, and other related data), which are dynamically updated based on user interaction and scene changes. Additionally, RT and GBSM technologies are employed to accurately predict and simulate wireless channel changes, enabling real-time updates [14].

Figure 3 shows the system displaying a real-time update of the environment and its corresponding digital twin environment. In the figure, (a) shows the environment change sensed in real-time, while (b) shows the corresponding digital twin environment, which is updated in real-time to reflect the movement of pedestrians. In the lower half of the figure, (c) shows a depth analysis in a depth map, where different colors indicate distance and depth, and (d) shows the corresponding channel characteristics, including the power distribution of the azimuth and elevation. When a pedestrian moves and either blocks or changes the signal reflection surface (e.g., passing through a large metal structure or a building), the power distribution of the azimuth shows a new peak value or a decrease of an existing peak value. This can be visually represented in the power distribution diagram, which allows users to view the position and height changes of new peaks or original peaks. As pedestrians move between floors or buildings, changes in their position relative to the surroundings alter the path of the vertical propagation signal, leading to shifts in the power distribution of elevation angles. This can result in an increase in power at certain elevation angles, where reflections enhance the signal. A digital twin channel precisely captures the dynamic behavior of changing environments, offering an accurate simulation of real-world channel conditions.

Figure 4 illustrates the simulation results of a space-time prediction channel model based on a generative adversarial network-gated recurrent unit (GAN-GRU) architecture. The GAN network can perform feature learning on measurement data, effectively doubling the amount of data, while the GRU network captures the mapping between a physical environment change and a space-time channel characteristic change. As shown in Figure 4, the GAN-GRU

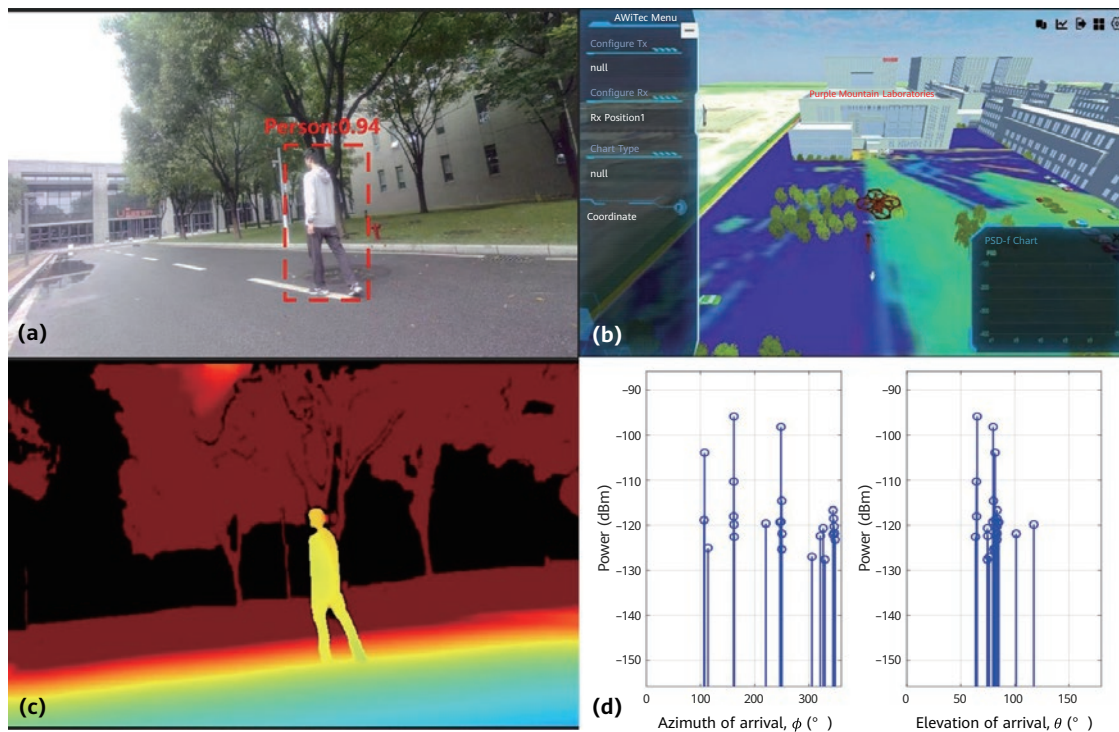


Figure 3 Real-time channel information visualization of DTOCM
 (a) Sensing information; (b) Digital twin environment; (c) Depth map; (d) Channel characteristics [14]

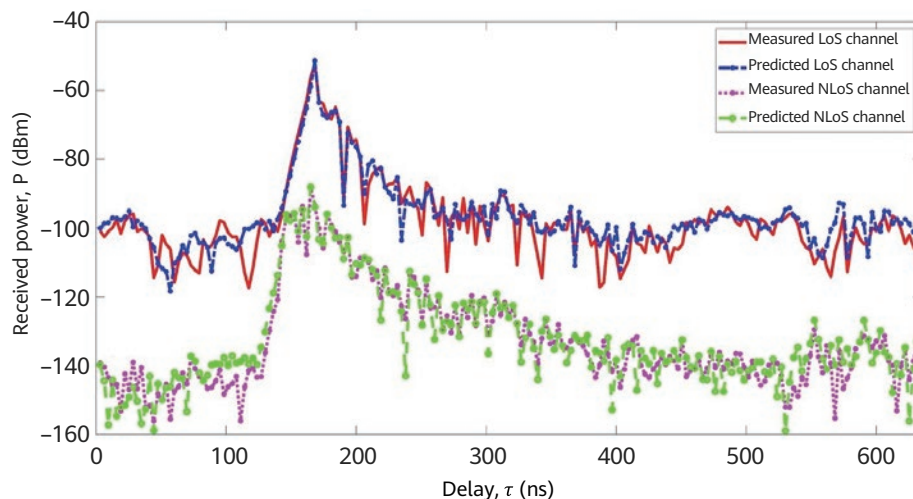


Figure 4 CIR of a measured channel and a predicted channel in LoS and NLoS scenarios within a 6 GHz frequency band [15]

space-time prediction channel model delivers accurate channel predictions in both line-of-sight (LoS) and non-line-of-sight (NLoS) environments within an indoor corridor. The receive power of the predicted channel accurately matches the receive power of the measured channel, and most paths of the channel are successfully predicted. This is clearly demonstrated in the channel impulse response (CIR) comparison diagram, which provides conclusive evidence of the feasibility of digital twin channel modeling [15].

5 Future Outlook

While DTOCM has the potential to revolutionize 6G network management, there are still several key challenges that need to be addressed. In this section, we will discuss some of the open research issues related to DTOCM in future 6G networks.

5.1 High-Accuracy RT

RT simulation is a critical component of DTOCM but its accuracy is influenced by three main factors. First, the complexity and details of the 3D model used in the RT simulation have a significant impact on its accuracy. However, it is a challenging task to create a highly detailed 3D model because a large amount of computing resources and powerful data input are required. This can be particularly difficult in a frequently changing dynamic environment. Second, the electromagnetic characteristics of the material represented in the model also play a crucial role in determining the accuracy of RT simulation. It is essential to properly distribute material properties such as reflectivity, absorption, and dielectric constants, because these properties determine how waves interact with the surface. Third, the basic algorithms used in the RT simulation also have a significant impact on its accuracy. Improving the capability of the algorithms to process complex interactions such as diffraction, scattering, and multiple reflections is a prerequisite for improving simulation accuracy. Additionally, the computational efficiency of these algorithms also affects the feasibility of real-time simulation, which is necessary for DTOCM.

5.2 Multimodal Data Convergence

Effectively perceiving and converging different types of multimodal data from various sources is a significant challenge in promoting DTOCM. Multimodal data includes not only static data collected from sensors and maps, but also real-time data dynamically generated from mobile entities such as UAVs and vehicles. To overcome this challenge, AI algorithms that are more efficient and lighter weight are required to enable the effective integration and utilization of multimodal data. This will be crucial in promoting the development and maturity of DTOCM.

5.3 Real-Time Processing and Reduced Latency

Real-time processing of sensing data and reduced latency play a vital role in ensuring the accuracy and responsiveness of DTOCM. To achieve this, dynamic data from various sources, such as sensors, UAVs, and user equipment (UE), must be effectively integrated in real-time. This involves calculating the corresponding changes in space-time-frequency channel characteristics based on the location

changes of the transmitter and receiver in the physical world and calculating the changes of the reflector and scatterer in the communication scenario. Furthermore, it is essential to minimize latency in environment sensing and data processing in order to ensure the channel map is promptly updated. This involves developing low-latency algorithms and potentially leveraging edge computing to process data closer to its source. Addressing these challenges will enhance the model's ability to provide accurate, real-time CSI.

5.4 Assisted Continuous 3D Space Radio Channel Modeling

Future networks are facing a significant challenge as channels transition from traditional, localized radio environments to complex, 3D spaces where multiple users and base stations continuously interact and move. This continuous 3D propagation environment is inherently intricate, with limited geographical data, and the varying requirements of different transmission technologies only add to the complexity to apply DTOCM in these environments. As a result, it is essential to develop a digital twin model that combines static mapping and real-time sensing to provide real-time and accurate channel information.

6 Conclusion

This paper presents a DTOCM framework that enables accurate modeling of dynamic channels, ultimately driving the optimization of 6G networks. We have outlined the vision, challenges, construction process, and working principles of DTOCM, and highlighted its ability to provide and visualize real-time channel information. Our research seeks to inspire further development and investigation of DTOCM, with the goal of enhancing the efficiency, adaptability, and performance of future networks.

References

- [1] C.-X. Wang *et al.*, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 2, pp. 905–974, 2023.
- [2] Z.-Q. Luo *et al.*, "SRCON: A data-driven network performance simulator for real-world wireless networks," *IEEE Commun. Mag.*, vol. 61, no. 6, pp. 96–102, 2023.
- [3] C.-X. Wang, J. Huang, H. Wang, X. Gao, X. You, and Y. Hao, "6G wireless channel measurements and models: Trends and challenges," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 22–32, 2020.
- [4] R. He *et al.*, "A kernel-power-density-based algorithm for channel multipath components clustering," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 11, pp. 7138–7151, 2017.
- [5] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 1, pp. 1–30, 2022.
- [6] X. Lin, L. Kundu, C. Dick, E. Obiodu, T. Mostak, and M. Flaxman, "6G digital twin networks: From theory to practice," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 72–78, 2023.
- [7] C.-X. Wang, Z. Lv, Y. Chen, and H. Haas, "A complete study of space-time-frequency statistical properties of the 6G pervasive channel model," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7273–7287, 2023.
- [8] C.-X. Wang, Z. Lv, X. Gao, X. You, Y. Hao, and H. Haas, "Pervasive wireless channel modeling theory and applications to 6G GBSSMs for all frequency bands and all scenarios," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9159–9173, 2022.
- [9] C. Huang, C.-X. Wang, Z. Li, Z. Qian, J. Li, and Y. Miao, "A frequency domain predictive channel model for 6G wireless MIMO communications based on deep learning," *IEEE Trans. Commun.*, vol. 72, no. 8, pp. 4887–4902, 2024.
- [10] F. Zhang *et al.*, "A radio wave propagation modeling method based on high-precision 3-D mapping in urban scenarios," *IEEE Trans. Antennas Propag.*, vol. 72, no. 3, pp. 2712–2722, 2024.
- [11] P. Koivumaki, G. Steinbock, and K. Haneda, "Impacts of point cloud modeling on the accuracy of ray-based multipath propagation simulations," *IEEE Trans. Antennas Propag.*, vol. 69, no. 8, pp. 4737–4747, 2021.
- [12] T. Qi, C. Huang, J. Shi, J. Li, S. Chen, and C.-X. Wang, "A Novel Dynamic Channel Map for 6G MIMO Communications," in *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, 2024, pp. 809–814.
- [13] R. Levie, Ç. Yapar, G. Kutyniok, and G. Caire, "RadioUNet: Fast radio map estimation with convolutional neural networks," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 6, pp. 4001–4015, 2021.
- [14] S. Xiao, H. Zhang, M. Yao, C. Cui, J. Li, C. Huang, and C.-X. Wang, "Demo: A novel 3D environment-aware digital twin online channel modeling platform," in *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, 2024, pp. 297–298.
- [15] Z. Li *et al.*, "A GAN-GRU based space-time predictive channel model for 6G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 9370–9386, 2024.
- [16] J. Li, C.-X. Wang, C. Huang, T. Qi, and T. Wu, "Digital Twin Online Channel Modeling: Challenges, Principles, and Applications," *IEEE Veh. Technol. Mag.*



AI-Driven Innovations in RF and Antenna Design

Guangjian Wang, Lingjun Yang, Jimmy Jian, Chandan Roy, Li Pan, Guolong Huang, Hua Cai, Wen Tong

Abstract

The rapid development of artificial intelligence (AI) technologies has led to extensive application of these technologies in radio frequency (RF) and antenna design. In this paper, we widely explore the application of AI technologies in crucial phases of RF and antenna design, including RF circuit design, antenna shape optimization, array syntheses, and electromagnetic efficient simulation, where AI technologies are employed to significantly improve design efficiency through automated design processes, fast simulation and optimization, and intelligent design tools. These technologies offer unique and advanced capabilities for handling complex issues, including multi-objective optimization, design refinement, high-dimensional problem solving, non-linear problem analysis, and multi-physical field coupling. We also discuss the challenges faced by the application of AI, potential solutions, and the future development of RF and antenna technologies. Our work offers a valuable reference for advancing RF and antenna design and the evolution of communication technologies.

Keywords

AI, RF design, antenna design, electromagnetic simulation, machine learning, filter

1 Introduction

In the information age, radio frequency (RF) and antenna technologies are the cornerstone of modern wireless communications systems, connecting people and the world. Without these technologies, we would be unable to communicate through mobile devices or keep in touch with our family and friends. Satellite communications would be paralyzed, making global information transmission impossible. The Internet of Things (IoT) would become an unrealizable fantasy, and smart devices would never be able to communicate with each other. RF and antenna design is a crucial component of a wireless communication system because the performance of RF devices and antennas has a significant impact on the quality and efficiency of the entire system [1].

However, the rapid development of technology and the growing demand for communications have created several challenges for RF and antenna technologies. These challenges include improving the transmission efficiency of RF signals, designing more compact and efficient antennas, and optimizing the use of increasingly scarce spectrum resources. Conventional design approaches rely on experience and trial-and-error tests to address these challenges. These approaches are laborious and computationally inefficient, especially when there are a large number of co-dependent component parameters. Rapidly evolving artificial intelligence (AI) technologies have been widely applied in RF and antenna design [2] due to their advanced data processing and learning capabilities, introducing innovative ideas and approaches to RF and antenna design [3].

In this paper, we explore innovative techniques to improve the performance of RF and antenna technologies. We conducted in-depth research on key fields such as RF circuit design, antenna design, and electromagnetic simulation to unlock the potential of AI in these fields. Through experiments and analysis, we investigated the use of AI technologies in simulating complex electromagnetic phenomena and achieved automatic adjustment of simulation parameters through reinforcement learning, significantly improving simulation efficiency. Figure 1 illustrates the composition of this paper. Section 2 outlines different AI technologies and AI models that can be used in RF and antenna design. Section 3 exemplifies the use of AI in RF circuit design, antenna design, and electromagnetic simulation. Section 4 discusses the unique advantages of AI in RF circuit design, antenna design, and electromagnetic simulation. Section 5 provides implementation examples of AI-based circuit design. Section 6 summarizes the challenges and prospects of AI in RF circuit design. Section 7 offers a conclusion to our work.

2 AI Models for RF and Antenna Design

AI is redefining RF and antenna design. AI technologies, especially machine learning (ML) and deep learning (DL), enable machines to learn and solve complex problems by imitating human cognitive processes and identifying data patterns. These technologies have been applied in RF and antenna design to improve design efficiency and enhance the design results, especially in solving high-dimensional, nonlinear, and multi-physical field coupling problems.

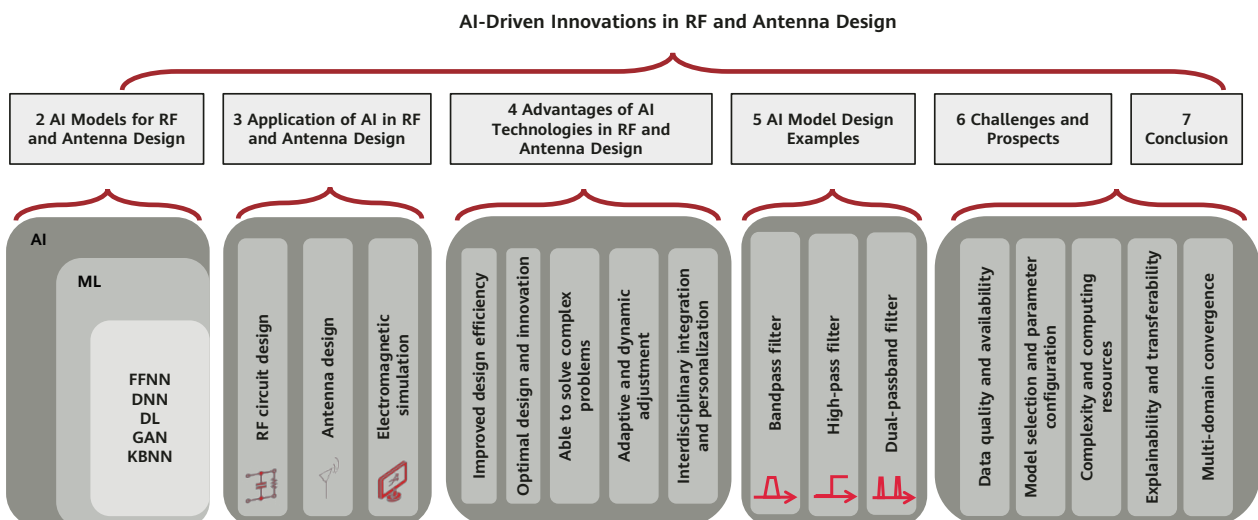


Figure 1 Composition of the paper

In RF design, feedforward neural networks (FFNNs), such as multilayer perceptrons (MLPs) and radial basis function (RBF) networks, are excellent solutions to non-dynamic modeling problems [4]. Wavelet neural networks (WNNs) can solve problems with high nonlinearity or sharp changes due to their local property of hidden neurons, facilitating model training and improving model precision. Extreme learning machines (ELMs) are FFNNs with only one hidden layer that can learn quickly and yield good performance in complex electromagnetic parameter modeling scenarios (especially when only small-sized training datasets are available). An ELM can accurately simulate the behavior of RF circuits to optimize the design of filters and amplifiers. Dynamic neural networks [5], recurrent neural networks (RNNs), and time-delay neural networks (TDNNs) play a significant role in characterizing the dynamic behavior of a non-linear device or circuit in the time domain [6]. Deep neural networks (DNNs) provide an advanced solution to complex modeling problems due to their multi-layer structure [7]. Generative adversarial networks (GANs) demonstrate significant advantages in the generation of novel designs [8, 9]. Knowledge-based neural networks (KBNNs) [10] use the current equivalent circuit and empirical model in computer-aided design (CAD) for microwave components, eliminating the need to prepare a large amount of training data and improving the extrapolation capability of the model. These technologies can be integrated to accelerate the design process and advance the overall design performance and reliability.

3 Application of AI in RF and Antenna Design

3.1 RF Circuit Design

AI is transforming RF circuit and antenna design by accelerating the design process and significantly improving circuit performance. For instance, in filter and coupler design, AI can be used to accurately calculate and optimize frequency responses, coupling attenuation characteristics, and group delays, meeting the diverse range of frequency and power selection requirements [13, 14]. In amplifier design, AI technologies introduce intelligent algorithms to optimize key parameters such as the gain and noise coefficient, achieving excellent amplifier performance under changing working conditions [15]. In oscillator design, AI can be used to design highly stable oscillators with low phase noise, delivering a high spectral purity even in

complex electromagnetic environments. AI can also be employed to accurately calculate the impedance matching network, significantly improving the transmission efficiency and overall performance of RF circuits.

In filter design, numerous techniques can meet the performance requirements of filters. These techniques can be selected based on the application scenarios. For instance, microstrip and stripline technologies can be optimally integrated with planar circuits. Conventional waveguide technology is suitable for high-frequency applications with a low transmission loss. The innovative substrate-integrated waveguide (SIW) can be used in high-frequency applications due to its low transmission loss and the ability to be optimally integrated with planar system structures. The design parameters involved in these techniques vary depending on the filter structure. For instance, the length and width of a microstrip are major geometric variables of a microstrip filter. The diameter of a via and the distance between a via and the next via are major variables of an SIW filter. The aperture length and window length are key geometric variables of a waveguide filter. The impact of geometric variables on filter performance is often characterized or evaluated through ML technology. A well-trained ML model is expected to predict the impact of geometric variables on the filter response. Such a model can be used to quickly improve the filter design without a computationally expensive and inefficient electromagnetic model [16].

3.2 Antenna Design

AI technologies enhance the efficiency and accuracy of antenna design. On the one hand, AI can be used to quickly generate antenna simulation results, eliminating the use of any model. On the other hand, AI can be used to fine-tune the size, shape, and structure of antennas or add specific structural elements, improving antenna performance in terms of gain, bandwidth, directivity, and other indicators. Additionally, AI can analyze the impacts of different polarization modes on signal transmission and reception, enabling engineers to select an optimal polarization mode, such as linear polarization, circular polarization, or elliptical polarization, based on scenario-specific requirements. This further improves signal transmission efficiency and reception quality. In antenna array design, AI algorithms can be used to design an optimal antenna layout, determine the number of elements, and optimize beam forming and control performance, significantly enhancing antenna

directivity and gains. In multi-band antenna design, AI technologies take antenna parameters and performance requirements for different frequency bands into account, achieving advanced compatibility and efficient operation of antennas on multiple frequency bands — the diversity of spectrum resources is essential in modern communications systems. In conclusion, the integration and application of AI technologies have resulted in many innovative antenna design solutions. These solutions significantly improve design efficiency and system performance, laying a robust foundation for the evolution of wireless communications.

Figure 2 illustrates the process of AI-based antenna design. First, engineers select a basic geometric shape based on their experience to meet the performance requirements and then find the optimal design parameters. During parameter optimization, updating model parameters through full-wave electromagnetic simulation is the most time-consuming step. This step can be replaced by an ML-based surrogate model to save computing resources and accelerate the antenna design process. A surrogate model can be trained on the dataset obtained through full-wave simulation, a process that generates output parameters such as S and gain parameters on the operating frequency

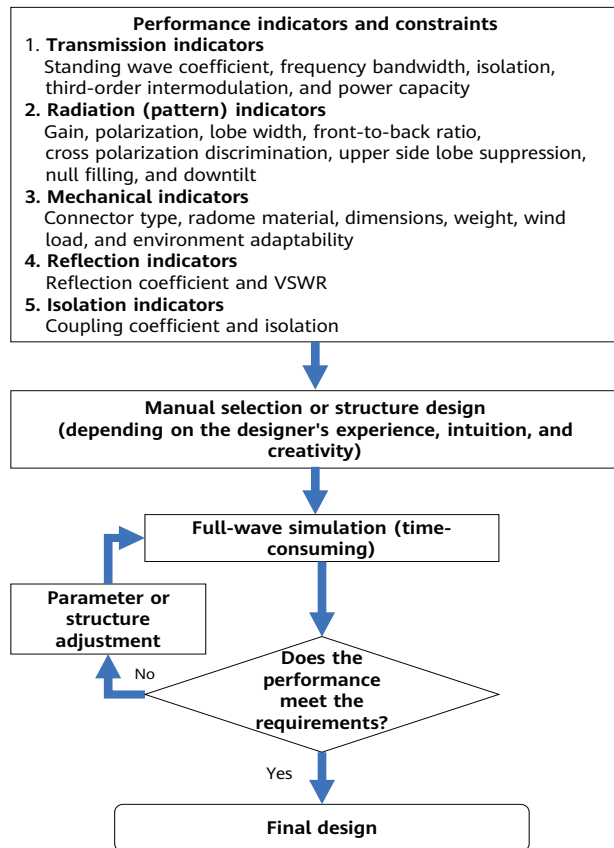


Figure 2 Antenna design process

bands by changing the geometric shapes of antennas. In the training process, the result generated by the surrogate model gradually approaches the full-wave simulation result, eliminating the need for full-wave simulation. Most surrogate model-assisted antenna optimization approaches [17] use the Gaussian Process (GP) surrogate model due to its easy implementation, processable analytical results, and ability to quantify uncertainties. Training of a surrogate model requires approximately 80% of datasets, while the remaining 20% is used to test model accuracy. If the test error does not meet the requirements, more data can be generated for training, or the surrogate model can be improved to repeat the GP process. After optimal device parameters are obtained through the surrogate model, the device needs to be verified and fine-tuned based on full-wave simulation to achieve optimal model accuracy and design performance. In antenna analysis, ML can be performed to predict various characteristics, such as radiation patterns and resonance frequencies, from simulation or measurement data. A complex surrogate model with multiple outputs can be integrated with multi-objective evolution algorithms to find the optimal antenna design that meets the performance requirements of multiple indicators [2]. Integrating different neural networks in antenna design is also a topic with significant research value. In Figure 2, the first step is selecting an antenna type, which can be completed by engineers with the help of a Support Vector Machine (SVM)-based recommendation system [18]. After the basic geometric shape of the antenna is determined, AI approaches such as artificial neural networks (ANNs) and stacked ensemble learning models can be used to calculate the values of model parameters.

3.3 Electromagnetic Simulation

In electromagnetic simulation, AI can be used to comprehensively evaluate system performance through multi-physical field coupling simulation, including electromagnetic, thermal, and structure simulation. Interdisciplinary simulation offers in-depth insights into RF and antenna design, guaranteeing advanced design robustness and adaptability. AI algorithms can be used to quickly calculate electromagnetic radiation and scattering, and extract key information from measurement data to optimize design parameters, demonstrating significant potential in the inversion analysis of electromagnetic simulation. This approach not only improves design accuracy but also accelerates the implementation of design concepts in products. Additionally, AI technologies can also be applied

in real-time electromagnetic simulation to quickly obtain design results and significantly accelerate design iteration. Fast iteration is critical to expediting prototype design and testing, significantly reducing the time required for product development. Last but not least, AI plays a critical role in quantifying uncertainties that may affect design reliability in electromagnetic simulation. Such uncertainties can be evaluated using AI technologies to provide a more reliable basis for decision-making in the design process, thereby improving the design success rate.

In forward inversion calculation between the source and the scattering field, matrix inversion needs to be performed during conventional full-wave simulation of electromagnetic scattering and radiation to calculate the induced current. This time-consuming process can be expressed as:

$$\vec{J} = (\vec{I} - \vec{\chi} \cdot \vec{G}_D)^{-1} \cdot (\vec{\chi} \cdot \vec{E}_{inc}) \quad (1)$$

To accelerate this process, researchers have proposed many DL-based non-iterative approaches and found that GAN-based AI approaches yield better performance than other neural networks (such as U-Net), especially when complex scatterers are involved. A prime example is the forward-induced current learning method (FICLM) [19]. It calculates induced currents through neural network mapping first and then calculates the scattering field by multiplying the Green function by the induced current. The accuracy of the calculated scattering field increases with the number of input schemes. As illustrated in Figure 3, the input scheme includes many combinations of the incident field and the background corresponding to dielectric constants. This meets the requirement of solving the electromagnetic scattering problem in wave physics.

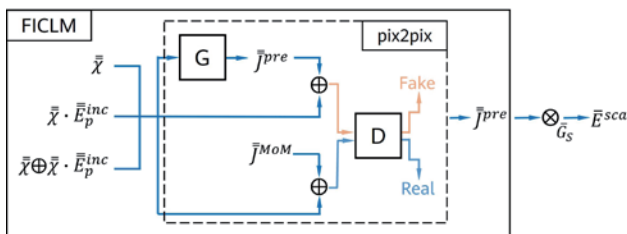


Figure 3 FICLM flowchart

In inversion and optimization, AI models and algorithms can be used to perform inversion analysis on electromagnetic phenomena to extract background information from electromagnetic measurement data and optimize device design parameters. Three types of neural network solvers were developed to solve electromagnetic inversion



Figure 4 Flowchart of the first type

problems. Figure 4 shows the design flowchart of the first type, which directly inverts the physical parameters of the scatterer from the electromagnetic measurement result. The following model expresses the learning process:

$$R_l = \min_{R_\theta, \theta} \sum_{m=1}^M f(R_\theta(\vec{E}_m^s), \vec{\chi}_m) + g(\theta) \quad (2)$$

We fit each pair of physical parameters $\vec{\chi}_m$ and scattering field \vec{E}_m^s in the training data to obtain the neural network R_l , which directly maps the scattering field to the physical parameters. We introduced a regularization term $g(\theta)$ to avoid overfitting. However, this model has limited inversion capabilities because it learns a large amount of unnecessary information (known information about wave physics) [20].

The second type still uses the conventional objective function approach, where a neural network is trained as a component to learn iterative solvers [21]. The third type integrates an approximation solver (such as a back propagation) with a DNN. It converts the input \vec{E}_m^s of the DNN from a measured electromagnetic field to an approximation solution $\vec{\chi}_m^a$ of a physical parameter (see the model flowchart in Figure 5), thereby reducing the learning workload of the neural network and simplifying the learning process [22, 23]. Inspired by the similarity between the iterative approach of inverse scattering problems (ISPs) and the DNN architecture in the third type of neural network solvers, researchers improved the DNN topology by incorporating three CNN modules that are cascaded and trained separately. Initially, researchers focused on the contrast information of the third type. The approximate and actual results of contrast information are used as the input and output of the CNN modules. Later on, inspired by wave

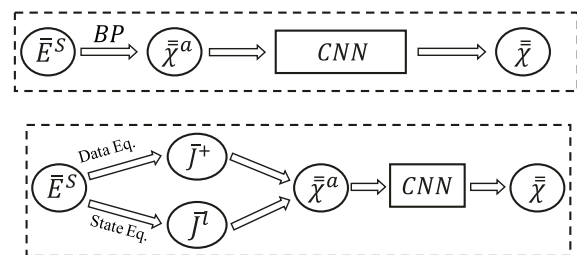


Figure 5 Flowchart of the third type

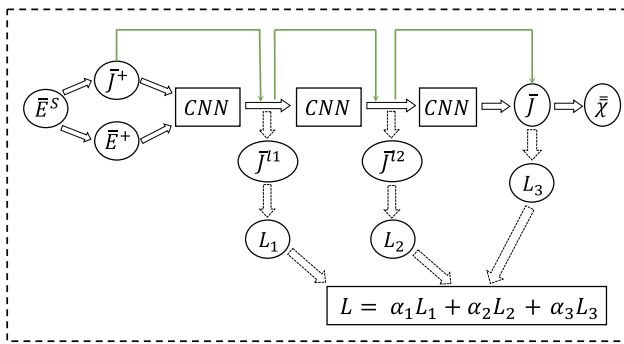


Figure 6 Electromagnetic inversion flowchart for neural networks

physics, they developed an enhanced neural network [24]. This network introduces the induced current and electric field in inputs and outputs to significantly advance inversion performance, as shown in Figure 6. The following model expresses the learning process of the neural network that integrates electromagnetic physics:

$$R_l = \min_{R_\theta, \theta} \sum_{m=1}^M f(R_\theta(\bar{J}^+, \bar{E}^+), \bar{J}^{l1}, \bar{J}^{l2}, \dots, \bar{J}) + g(\theta) \quad (3)$$

where $\bar{J}^{l1}, \bar{J}^{l2}, \dots, \bar{J}$ indicates the density of the output induced current of each CNN module.

4 Advantages of AI Technologies in RF and Antenna Design

4.1 Improved Design Efficiency

AI technologies have significantly accelerated and innovated RF and antenna design by automating the design process, including repetitive tasks such as parameter calculation and model generation. The fast simulation and optimization capabilities of AI algorithms enable designers to formulate high-quality design solutions efficiently. The emergence of intelligent design tools further enhances design efficiency and visualize the design process.

AI surrogate models significantly improve design efficiency by using ML algorithms and models to replace complex RF/microwave components. This approach achieves the same accuracy as electromagnetic models do and makes circuit models more computationally efficient by replacing the time-consuming full-wave electromagnetic simulation. However, it requires fully trained AI/ML models and a certain amount of data to do so. Nevertheless, this approach

not only improves design efficiency but also reduces the development cost, accelerating the transformation from concept to implementation in RF and antenna design [25].

4.2 Optimal Design and Innovation

AI offers advanced optimization capabilities for RF and antenna design. For instance, a multi-objective AI model can take many design objectives into account, such as performance, cost, and size, to find the optimal design. AI also offers advanced data analysis capabilities for fine-grained control of design details, achieving global optimization and further enhancing design quality. Furthermore, AI provides innovative design capabilities to inspire novel design ideas and approaches.

In the design optimization of complex RF devices, AI models trained on numerical and experimental data offer excellent optimization capabilities, which can be used to design complex antennas, arrays, and RF devices — a task difficult for conventional design approaches. Designing complex RF devices is equivalent to solving high-dimensional and non-convex optimization problems and usually involves time, frequency, and spectral domains and many constraints. Such problems are challenging due to the characteristics of nonlinearity, multiple scales, and strong mutual coupling. Optimization algorithms often require numerous simulation results to achieve the performance goal, and the number of iterations depends on which optimization algorithm is used and the complexity of the problem. Full-wave electromagnetic simulation is a suboptimal approach because it is both energy and computationally inefficient, especially for fast simulation and optimization of electrically large electromagnetic models. ML provides an advanced solution to these problems by achieving the same precision as an electromagnetic model does without time-consuming simulation, significantly improving the efficiency and feasibility of optimization [26].

4.3 Able to Solve Complex Problems

AI technologies can be used to find optimal solutions for RF and antenna design, especially for design scenarios with high dimensions and complex constraints. AI algorithms excel in identifying the characteristics and behavior of non-linear problems and provide more accurate analysis results, which conventional approaches are incapable of.

AI also excels in analyzing multi-physical field coupling by taking different information about physical fields into account, such as electromagnetic, thermal, and structure information, comprehensively evaluating the performance of RF and antenna systems. This interdisciplinary analysis approach provides more in-depth design insights to ensure optimal system performance from various perspectives. Additionally, AI technologies deliver significant value in managing uncertainties, such as ever-changing material parameters and working environments, achieving a robust design and assuring high stability and reliability for RF and antenna systems under various conditions. AI has become indispensable in RF and antenna design due to its advanced capabilities, facilitating innovation in design approaches and improving design quality.

In model analysis of complex RF/microwave structures, two representation approaches are often used: the electromagnetic model (fine model) and the lumped parameter equivalent circuit model (rough model). The fine model is more accurate but is computationally more expensive. The rough model is computationally efficient but inaccurate. Traditional spatial mapping technology optimizes geometric parameters and achieves performance goals by converting parameters between these models [27]. This parameter conversion process can be completed more efficiently by AI/ML models in fewer computational steps. These models can identify the complex relationships between the fine and rough models by learning a large amount of data to quickly and accurately adjust parameters in the optimization process. This approach requires fewer computing resources and accelerates design iteration, enabling designers to design high-performance RF/microwave structures more efficiently in a cost-effective manner [28].

4.4 Adaptive and Dynamic Adjustment

Adaptive and dynamic adjustment is key to significantly advancing the applicability and robustness of RF and antenna systems. Multi-function AI systems can be deployed to monitor and adjust RF and antenna systems in real time, significantly improving system performance from various perspectives, as shown in Figure 7.

These AI systems can monitor system parameters in real time and automatically adjust RF circuit and antenna parameters through intelligent algorithms under

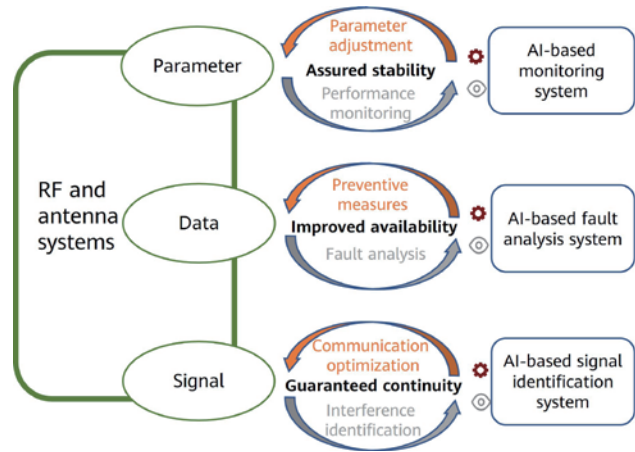


Figure 7 AI-based adaptive adjustment in RF circuit and antenna design

ever-changing working conditions, guaranteeing the performance and stability of the communications system. Such real-time monitoring and adjustment capabilities significantly enhance the flexibility and response speed of RF circuits and antenna systems.

AI technologies also play a significant role in intelligent fault diagnosis by analyzing system running data to detect faults and forecast problems quickly. Based on the analysis result, preventive or corrective measures can be taken to improve system availability and reliability. This approach reduces the labor and time costs of system maintenance and prevents service interruptions caused by faults.

Furthermore, AI technologies enable RF and antenna systems to better adapt to dynamic environments, for instance, mobile communications scenarios, where signal interference and other environmental factors may affect communication quality. Such environmental changes can be analyzed in real time using AI to automatically adjust the beam status of antenna arrays, optimize signal transmission and reception performance, and ensure continuous and stable communication.

4.5 Interdisciplinary Integration and Personalization

One of the advantages of AI technologies in circuit and antenna design is promoting interdisciplinary integration and personalized design. AI can be used to integrate advanced technologies and concepts from different fields, such as material science, electronic engineering, and computer science, to achieve converged innovation, while offering customized RF and antenna design solutions based on specific requirements and application scenarios.

Interdisciplinary integration creates new perspectives and solutions to RF circuit and antenna design, facilitating technological innovation and development.

In personalized design, the application of AI technologies improves the flexibility and degree of customization of the design process. AI can be used to generate customized design solutions by analyzing the requirements of specific applications, such as the communication range, frequency bandwidth, and environmental factors. This capability meets a diverse range of market requirements and offers optimal solutions for specific applications, enhancing the performance and applicability of RF systems.

5 AI Model Design Examples

As AI technologies evolve, they demonstrate significant advantages in RF circuit and antenna design, one of the most crucial and challenging tasks in wireless communications systems. In this section, we discuss several filter design cases to highlight the application of AI in RF and antenna design.

5.1 Bandpass and High-Pass Filters

The topology of a device can be pixelated to achieve a high degree of freedom for device variables and improve the odds of achieving the performance goal. This approach enjoys widespread attention from the microwave research community due to its advanced performance. It can be used to implement all kinds of microwave circuits, including power amplifiers and antennas [15]. Based on the idea of pixelation, we designed different types of microwave filters that can operate on the Ka frequency band (26.5 GHz to 40 GHz) and meet specific frequency selectivity requirements. These filters have been widely used in autonomous driving vehicles. The first example includes bandpass and high-pass filters. We used alumina with a thickness of 0.127 mm as the substrate and gold with a thickness of 4 μm as transmission cables to design the filters. The design process is as follows: (1) Create an electromagnetic model of a solid patch on commercial software such as CST and HFSS. The patch has two ports with 50 ohm microstrips. (2) Divide the patch into grids of different sizes based on the quantity of resonators. (3) Set each grid to metal (1) or non-metal (0) for the neural network's training

process. (4) Configure the Python environment to run the electromagnetic software automatically. (5) Generate data using the electromagnetic software. (6) Train a CNN model using binary sequences (1: metal; 0: non-metal) as the input and S_{11} and S_{21} parameters as the output. (7) Replace the electromagnetic model with the trained CNN model, which is computationally more efficient. (8) Develop an optimization algorithm based on the CNN model. The algorithm uses the passband return loss and stopband insertion loss as the input to generate the required filter shape.

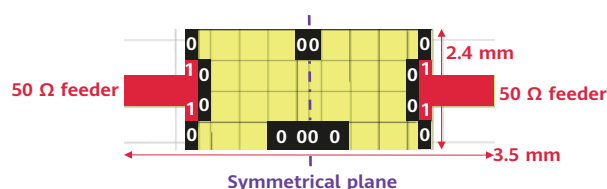


Figure 8 Simulation diagram of a pixelated filter

Figure 8 illustrates the primary filter structure configured on CST. To accelerate data generation, we used a symmetrical plane in the target structure, where the right half of the structure is a mirror of the left half. We divided the left half plane into 20 rectangular grids (4 x 5). Based on the experience we amassed from coupling design, we set some of the grids to 1 (red) or 0 (black) before the optimization, as shown in Figure 8. We also designed different grid sizes for the input/output (I/O), inter-resonator (I/R) coupling, and resonator areas to accelerate full-wave simulation without compromising model precision. The size of each rectangular grid in the I/O and coupling areas is 0.12 mm x 0.3 mm. The size of each grid in the resonator area is 0.27 mm x 0.3 mm. Because the coupling area (coupling gap) is too small compared to the resonator size in most designs, we chose a smaller grid size for the potential coupling area and a larger grid size for the resonator area to reduce the total number of variables in the target design space. In our design, 15 grids are considered variables. We found 32,768 possible structures in total by specifying metal and non-metal grids, performed 32,768 full-wave electromagnetic simulations on frequencies ranging from 34 GHz to 46 GHz, and obtained 51 samples. We used 15 binary variables and frequency variables as the input, and the real and imaginary parts of S_{11} and S_{21} as the output to train an EM-CNN model that can replace full-wave simulation.

Additionally, we integrated an optimization solution based on a genetic algorithm (GA) to develop filters with different filtering performance by modifying the layout and combination (distributions) of metal and non-metal grids (distributions). The stopbands of the bandpass filter are defined as follows:

- Stopband 1 (SB-1) ≤ 40 GHz
- Stopband 2 (SB-2) ≥ 44 GHz
- Passband: $41 \text{ GHz} \leq \text{Passband (PB)} \leq 43 \text{ GHz}$
- Passband insertion loss: $\max[S_{21}(\text{PB})] > -1 \text{ dB}$
- Maximum passband return loss: $\max[S_{11}(\text{PB})] < -10 \text{ dB}$
- Stopband suppression: $\max[S_{21}(\text{SB})] < -25 \text{ dB}$

The stopbands of the high-pass filter are defined as follows:

- Stopband 1 (SB-1) ≤ 40 GHz
- Passband: ≥ 42 GHz
- Minimum in-band insertion loss: $\max[S_{21}(\text{PB})] > -1 \text{ dB}$
- Maximum passband return loss: $\max[S_{11}(\text{PB})] < -10 \text{ dB}$
- Stopband suppression: $\max[S_{21}(\text{SB})] < -25 \text{ dB}$

Based on these design objectives, the loss function of the filters can be expressed as:

$$K = \max[(S_{11})_{PB}, -RL] + w * \max[(S_{21})_{SB}, -IL] \quad (4)$$

where w indicates the weighting factor between the passband return loss (RL) and the stopband insertion loss (IL). The weighting factor w is a key factor for adjusting the passband return loss and the stopband insertion loss. The factor is determined according to the sensitivity of the optimization parameters and their collective impacts on the return loss and insertion loss in the passband and stopband. The EM-CNN model is then used to optimize the loss function for the bandpass filter, generating a binary sequence [1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1], which is converted into a geometric shape on CST, as shown in Figure 9a. The binary sequence generated for the high-pass filter is [0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1], which is also converted to a geometric shape on CST, as shown in Figure 9b. Figure 10a and Figure 10b illustrate the full-wave simulation performance of these filters, which deliver optimal in-band and out-of-band performance that meet the requirements of our design.

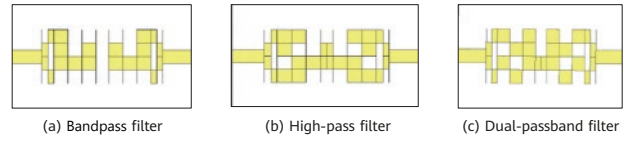
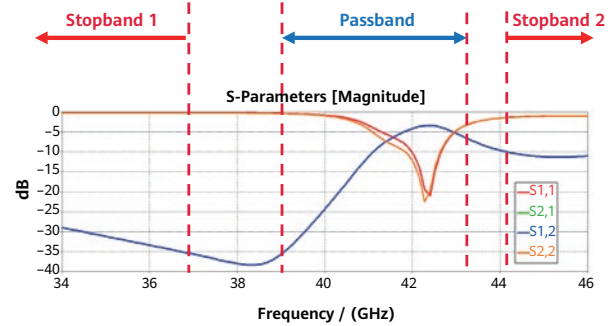
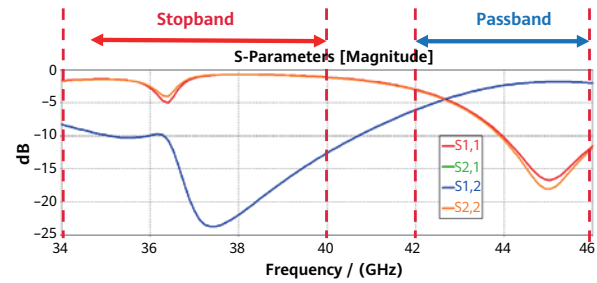


Figure 9 Optimization results of pixelated filters



(a) Bandpass filter



(b) High-pass filter

Figure 10 Simulation results of optimized filters

5.2 Dual-Passband Filter

We also used the EM-CNN model to design a dual-passband filter, which requires a more complex objective function than single-passband filters.

- Stopband 1: $36 \text{ GHz} \leq \text{Stopband (SB-1)} \leq 38 \text{ GHz}$
- Stopband 2: $44 \text{ GHz} \leq \text{Stopband (SB-2)} \leq 46 \text{ GHz}$
- Passband 1: $34 \text{ GHz} \leq \text{Passband (PB1)} \leq 36 \text{ GHz}$
- Passband 2: $38 \text{ GHz} \leq \text{Passband (PB2)} \leq 42 \text{ GHz}$
- Passband insertion loss: $\max[S_{21}(\text{PB})] > -2 \text{ dB}$
- Maximum in-band return loss: $\max[S_{11}(\text{PB})] < -15 \text{ dB}$
- Out-of-band rejection: $\max[S_{21}(\text{SB})] < -25 \text{ dB}$

We set w to 0.8 to balance the passband return loss and stopband insertion loss, minimizing the loss. Similar to the previous examples, we used a GA-based optimization algorithm to obtain a binary sequence [1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1] that characterizes the geometric shape of the filter and the corresponding target response. Figure 9c shows the geometric shape of the

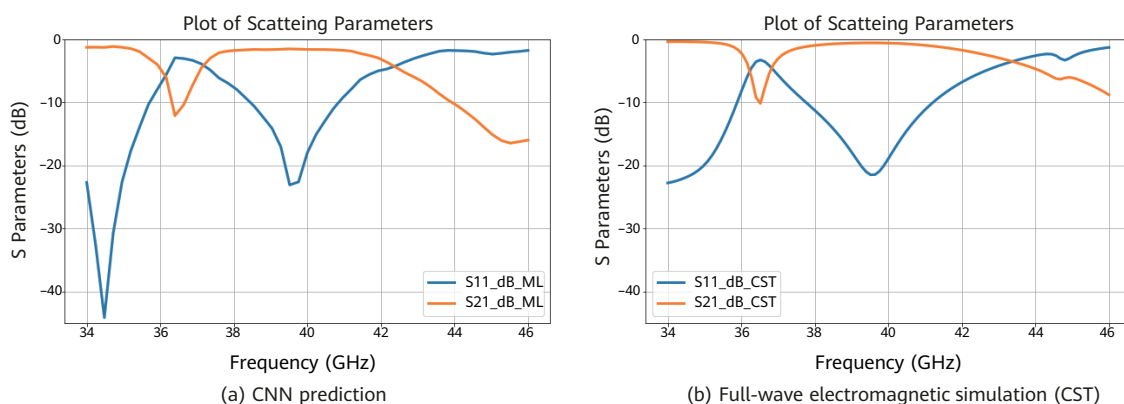


Figure 11 Reponse comparison of optimized filters

filter corresponding to the optimized binary sequence. We compared the CNN prediction results with the full-wave simulation result. Figure 11a shows the S parameter response of the filter predicted by the CNN model. Figure 11b shows the S parameter response calculated by CST. We can see that the response calculated by CST is highly similar to that predicted by the CNN model and fulfills the filter performance goal. This example is only used to explain the feasibility of the design, which does not meet all performance requirements due to a short design time. Increasing the degree of freedom and the density of grids can further improve filter performance.

6 Challenges and Prospects

6.1 Data Quality and Availability

The generation, quality, and availability of data are critical to the successful application of AI models in antenna design. Large-scale and high-quality data is the cornerstone of AI model training. However, RF and antenna design faces several data challenges: lack of data, heavy reliance on expertise, and high complexity. Data used in RF and antenna design often comes from electromagnetic model simulation, which is computationally expensive and makes data obtaining more difficult.

An accurate AI/ML model needs to be trained on adequate training, validation, and test datasets that fully characterize the problem to be solved. For RF/microwave structures, these datasets are usually generated through electromagnetic model simulation, which is an expensive and time-consuming process. Additionally, improving the accuracy and generalization capability of a model

involves data collection, organization, and labeling. This process requires careful selection and sampling of data to ensure that the quality and availability of the datasets can characterize the key areas in the design space.

To address the data challenges faced by AI model training in RF and antenna design, we can implement various policies to strike a balance between data generation time and model accuracy: (1) Active learning: Selectively generates the most informational data points, reducing the total amount of required data while maintaining high model accuracy. (2) Transfer learning: Creates models based on pre-trained models that have learned features and patterns from related tasks, significantly reducing the amount of new data required for model training. (3) Data augmentation: Increases the scale of datasets, reduces the demand for extra electromagnetic simulation, and maintains the diversity of training data. (4) Dimensionality reduction: Focuses on the most crucial parameters, simplifies the data generation process, and reduces the computational cost. A strategic trade-off is required between minimum data generation time and maximum model accuracy.

In addition to data collection and organization, data can be generated using AI technologies. For instance, diffusion models and other generative models offer an innovative data generation approach by learning the distribution features of current data and generating new data samples that are similar to the current data. These models can be used to generate more training data from limited measurement or simulation data for RF and antenna design. These modes can also identify complex data structures, improving the diversity of the generated data while maintaining consistency with the original dataset. GAN is also an advanced data generation approach that can generate realistic data samples through adversarial training. In RF and antenna design, GANs can be used to

generate electromagnetic simulation data to help designers evaluate the performance of different design solutions in the early design phase. In conclusion, the policies mentioned above can be used to reach an optimal trade-off between minimum data generation time and maximum model accuracy and develop effective AI/ML models for RF/microwave structures, accelerating the development process of AI models and offering innovative and efficient RF and antenna design solutions.

6.2 Model Selection and Parameter Configuration

In RF/microwave structure design, AI/ML model development faces a daunting challenge: model selection and hyperparameter configuration. Because different problems have different characteristics, no model can meet the requirements of all scenarios. We need to select the most suitable model architecture based on the problem description. Additionally, selecting hyperparameters, such as the number of layers, number of neurons, data segmentation ratio, and activation function, has a significant impact on model performance. Current hyperparameter selection approaches heavily depend on expertise and trial-and-error tests, which are laborious and time-consuming, and may cause uncertainties in the model development process. Efficiently and accurately selecting models and configuring hyperparameters are the essential challenges for the development of high-performance AI/ML models.

To address these challenges, we can implement the following policies: (1) Develop automated model selection tools that can recommend or select the most suitable model architecture based on the characteristics and data features of the problem. (2) Use hyperparameter optimization techniques, such as grid search, random search, or Bayesian optimization, to systematically explore the hyperparameter space and find the optimal hyperparameter combination. (3) Develop an automated development process integrating data pre-processing, model training, hyperparameter optimization, and model evaluation to reduce manual intervention and improve development efficiency.

6.3 Algorithm Complexity and Computing Resources

Algorithm complexity and computing resources are the major obstacles to the application of AI/ML algorithms in

RF and antenna design. Advanced AI algorithms, especially DL models, often require a large number of computing resources for training and inference. These resources include but are not limited to high-performance GPUs, large storage spaces, and fast data processing capabilities, which are expensive and difficult to obtain, especially for research teams and small development teams. Additionally, highly complex algorithms may increase the training time, slowing down design iterations and innovation processes. Balancing algorithm complexity and computing resources has become a key issue for efficient AI-assisted design.

To address these challenges, we can implement the following solutions: (1) Algorithm optimization: More efficient algorithms can be developed to reduce computing steps and resource consumption while maintaining or even improving model performance. (2) Model simplification: Technologies such as model pruning and knowledge distillation can be used to simplify the model structure and reduce model parameters, thereby lowering the computing workload. (3) Hardware acceleration: Exclusive hardware, such as tensor processing units (TPUs) and field programmable gate arrays (FPGAs), can be used to accelerate AI algorithms, offering optimal computing capabilities for DL. (4) Cloud computing resources: The advanced computing resources provided by cloud computing services can be allocated on demand to reduce investment costs for local hardware. (5) Parallel computing: Parallel computing technology can be used to allocate training tasks to different processors or devices to reduce the training time. (6) Resource scheduling and management: An intelligent resource scheduling system can be developed to optimize the allocation and use of computing resources, and improve resource efficiency. (7) Lightweight models: Lightweight AI models, such as MobileNet and ShuffleNet, can be developed for environments with limited resources. (8) Model quantization: Model quantization technology can be used to lower the requirements on model precision, memory, and storage and reduce computational complexity. (9) Heterogeneous computing resources: Different types of computing resources, such as CPUs, GPUs, and application-specific integrated circuits (ASICs), can be integrated to achieve optimal resource allocation for computing tasks. These solutions can be employed to reduce the complexity of AI algorithms, accelerate the design process, and improve the efficiency and feasibility of RF and antenna design when computing resources are limited.

6.4 Explainability and Transferability

The explainability of AI models is an important issue that is often neglected in RF and antenna design. Although some multi-objective, multi-functional AI models, especially DL models, deliver excellent performance in solving certain complex problems, they are often considered "black boxes" because it is difficult to understand how these models make decisions. In RF and antenna design, designers need to understand the decision-making process of the model to ensure that the design meets the requirements of the principles of physics and the application scenarios. Lack of explainability not only increases design risks, but also limits the application of AI models in key application scenarios.

To improve model explainability in RF and antenna design, we can adopt the following approaches: (1) Develop and apply explainable AI technologies, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHAPley Additive exPlanations (SHAP), which can explain model prediction results. (2) Use visualization tools to demonstrate the working principles of models, including the significance of features and decision boundaries. (3) Use simpler and easy-to-understand models, such as decision trees or linear models, although they may underperform complex models in certain scenarios. (4) Use model distillation technology to migrate the knowledge of complex models to a simpler model. (5) Implement post-processing technologies, such as rule extraction, to extract explainable rules from black-box models. (6) Take explainability into account in the model design phase and select naturally transparent model architectures. These approaches can improve the explainability of AI models in RF and antenna design, enhance designers' trust in models, and promote the application of AI technologies in this field.

However, there is a trade-off between model explainability and transferability. For instance, complex models may yield good performance in new environments (good transferability), but they may be difficult to explain (suboptimal explainability). Simple models are easy to explain but they may deliver unsatisfactory performance in diversified environments or scenarios with a diverse range of data (suboptimal robustness and transferability). To compensate for the lack of transferability, multiple models must be designed for different scenarios in the early

design phase, increasing the design cost and maintenance difficulty. These models may fail to adapt to ever-changing application scenarios. Consequently, new and customized AI models need to be developed for RF circuit networks, just as the world of physics needs a general theory. These models must have high explainability and advanced generalization capability (transferability). Developing such models is one of the key challenges for the extensive application of AI technologies in RF and antenna design.

6.5 Multi-Domain Convergence

RF and antenna design is a cross-disciplinary field involving many disciplines, such as electromagnetics, material science, and electronic engineering. The application of AI models in this field requires the processing and integration of complex data and knowledge from different disciplines. However, multi-domain convergence faces many challenges in AI model development. Data from different disciplines may have different characteristics and formats, making it difficult to integrate. Additionally, expertise and theories in different domains must be integrated to ensure that AI models can thoroughly understand and solve problems. Current AI models often focus on only one domain and lack interdisciplinary integration and collaboration, limiting their performance and application in RF and antenna design.

To address these challenges, we can implement the following approaches: (1) Set up an AI model development team of experts from different domains to ensure that multi-disciplinary knowledge and data are considered. (2) Develop a data standardization process to unify the formats of data collected from different sources, facilitating data processing and analysis by AI models. (3) Use multi-task learning technology to enable AI models to simultaneously learn multiple tasks, promoting the convergence of knowledge from different domains. (4) Implement transfer learning technology to migrate the knowledge learned in one domain to another, improving the adaptability and performance of models in the new domain. (5) Design domain-specific model architectures that can better characterize and process data and knowledge from specific disciplines. (6) Develop cross-disciplinary knowledge graphs to integrate professional knowledge in different domains and provide rich background knowledge for AI models.

7 Conclusion

In this paper, we discussed in detail the application, advantages, challenges, and future development of AI in RF and antenna design. AI technologies have demonstrated significant potential in antenna optimization, RF circuit design, and electromagnetic simulation by improving design efficiency, optimizing design results, and solving complex problems. Although AI faces many challenges in RF and antenna design, such as data quality, model selection, algorithm complexity, model explainability, and model transferability, AI still enjoys a promising prospect in this field. We believe that the continuous research and innovation on AI will facilitate more significant breakthroughs and progress in RF and antenna design. AI has opened up a new horizon for RF and antenna design by improving design efficiency and optimizing performance, facilitating the development of wireless communications. However, certain technological and data challenges need to be addressed before AI can fully unlock its potential in the field. Future research is expected to focus on improving data quality, developing more efficient algorithms, enhancing model explainability, and promoting multi-domain convergence. Such research will advance the development of wireless communications systems and deliver significant benefits to our daily lives.

References

- [1] W. Tong and P. Zhu, Eds., "6G: The next horizon: From connected people and things to connected intelligence," Cambridge: Cambridge University Press, 2021. doi: 10.1017/9781108989817.
- [2] S. K. Goudos, P. D. Diamantoulakis, M. A. Matin, P. Sarigiannidis, S. Wan, and G. K. Karagiannidis, "Design of antennas through artificial intelligence: State of the art and challenges," *IEEE Commun. Mag.*, vol. 60, no. 12, pp. 96–102, Dec. 2022, doi: 10.1109/MCOM.006.2200124.
- [3] N. Sarker, P. Podder, M. R. H. Mondal, S. S. Shafin, and J. Kamruzzaman, "Applications of machine learning and deep learning in antenna design, optimization, and selection: A review," *IEEE Access*, vol. 11, pp. 103890–103915, 2023, doi: 10.1109/ACCESS.2023.3317371.
- [4] H. J. KELLEY, "Gradient theory of optimal flight paths," *ARS J.*, Jun. 2012, doi: 10.2514/8.5282.
- [5] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022, doi: 10.1109/TPAMI.2021.3117837.
- [6] "Time-delay neural networks for control," *IFAC Proc. Vol.*, vol. 27, no. 14, pp. 967–972, Sep. 1994, doi: 10.1016/S1474-6670(17)47423-4.
- [7] "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017, doi: 10.1016/j.neucom.2016.12.038.
- [8] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.
- [9] Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, Oct. 2020, doi: 10.1145/3422622.
- [10] "Knowledge-based artificial neural networks," *Artif. Intell.*, vol. 70, no. 1–2, pp. 119–165, Oct. 1994, doi: 10.1016/0004-3702(94)90105-8.

- [11] F. Feng, W. Na, J. Jin, J. Zhang, W. Zhang, and Q.-J. Zhang, "Artificial neural networks for microwave computer-aided design: The state of the art," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 11, pp. 4597–4619, Nov. 2022, doi: 10.1109/TMTT.2022.3197751.
- [12] S. Haykin, "Neural networks: A comprehensive foundation," Subsequent. Upper Saddle River, N.J: Prentice Hall, 1998.
- [13] C. Roy and K. Wu, "Homotopy optimization and ANN modeling of millimeter-wave SIW cruciform coupler," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 11, pp. 4751–4764, Nov. 2022, doi: 10.1109/TMTT.2022.3200040.
- [14] A. Pietrenko-Dabrowska and S. Koziel, "Low-cost design optimization of microwave passives using multifidelity EM simulations and selective broyden updates," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 11, pp. 4765–4771, Nov. 2022, doi: 10.1109/TMTT.2022.3207482.
- [15] E. A. Karahan, Z. Liu, and K. Sengupta, "Deep-learning-based inverse-designed millimeter-wave passives and power amplifiers," *IEEE J. Solid-State Circuits*, vol. 58, no. 11, pp. 3074–3088, Nov. 2023, doi: 10.1109/JSSC.2023.3276315.
- [16] C. Roy and K. Wu, "A review of electromagnetics-based microwave circuit design optimization," *IEEE Microw. Mag.*, vol. 25, no. 7, pp. 16–40, Jul. 2024, doi: 10.1109/MMM.2024.3387036.
- [17] B. Liu *et al.*, "An efficient method for complex antenna design based on a self adaptive surrogate model-assisted optimization technique," *IEEE Trans. Antennas Propag.*, vol. 69, no. 4, pp. 2302–2315, Apr. 2021, doi: 10.1109/TAP.2021.3051034.
- [18] D. Shi, C. Lian, K. Cui, Y. Chen, and X. Liu, "An intelligent antenna synthesis method based on machine learning," *IEEE Trans. Antennas Propag.*, vol. 70, no. 7, pp. 4965–4976, Jul. 2022, doi: 10.1109/TAP.2022.3182693.
- [19] Z. Ma, K. Xu, R. Song, C.-F. Wang, and X. Chen, "Learning-based fast electromagnetic scattering solver through generative adversarial network," *IEEE Trans. Antennas Propag.*, vol. 69, no. 4, pp. 2194–2208, Apr. 2021, doi: 10.1109/TAP.2020.3026447.
- [20] Z. Wei and X. Chen, "Deep-learning schemes for full-wave nonlinear inverse scattering problems," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 1849–1860, Apr. 2019, doi: 10.1109/TGRS.2018.2869221.
- [21] R. Guo, M. Li, F. Yang, S. Xu, G. Fang, and A. Abubakar, "Application of gradient learning scheme to pixel-based inversion for transient EM data," in *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, Mar. 2018, pp. 1–3. doi: 10.1109/COMPEN.2018.8496518.
- [22] L. Li, L. G. Wang, F. L. Teixeira, C. Liu, A. Nehorai, and T. J. Cui, "DeepNIS: Deep neural network for nonlinear electromagnetic inverse scattering," *IEEE Trans. Antennas Propag.*, vol. 67, no. 3, pp. 1819–1825, Mar. 2019, doi: 10.1109/TAP.2018.2885437.
- [23] Y. Sun, Z. Xia, and U. S. Kamilov, "Efficient and accurate inversion of multiple scattering with deep learning," *Opt. Express*, vol. 26, no. 11, pp. 14678–14688, May 2018, doi: 10.1364/OE.26.014678.
- [24] "Physics-inspired convolutional neural network for solving full-wave inverse scattering problems | IEEE Journals & Magazine | IEEE Xplore," Accessed: Jul. 10, 2024. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8741152>
- [25] Q.-J. Zhang, K. C. Gupta, and V. K. Devabhaktuni, "Artificial neural networks for RF and microwave design - from theory to practice," *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 4, pp. 1339–1350, Apr. 2003, doi: 10.1109/TMTT.2003.809179.
- [26] C. Roy, W. Lin, and K. Wu, "Swarm intelligence-homotopy hybrid optimization-based ANN model for tunable bandpass filter," *IEEE Trans. Microw. Theory Tech.*, vol. 71, no. 6, pp. 2567–2581, Jun. 2023, doi: 10.1109/TMTT.2023.3236676.
- [27] J. W. Bandler, R. M. Biernacki, S. H. Chen, P. A. Grobelny, and R. H. Hemmers, "Space mapping technique for electromagnetic optimization," *IEEE Trans. Microw. Theory Tech.*, vol. 42, no. 12, pp. 2536–2544, Dec. 1994, doi: 10.1109/22.339794.
- [28] C. Roy and K. Wu, "Surrogate model-based filter optimization by a field-circuit model mapping," *IEEE Trans. Microw. Theory Tech.*, vol. 72, no. 5, pp. 3144–3157, May 2024, doi: 10.1109/TMTT.2023.3318692.



Physics-inspired Intelligent Communication: Opportunities, Advances, and Trends

Ziqing Xing, Ridong Li, Zirui Chen, Zhaohui Yang, Zhaoyang Zhang
College of Information Science and Electronic Engineering, Zhejiang University

Abstract

Artificial intelligence (AI) has become increasingly important in the evolution of wireless networks. In wireless domains, applications of deep neural networks (DNNs) have been proven effective in improving the performance of many tasks and even leading to new wireless use cases. However, the existing wireless AI is mainly driven by wireless data based on conventional deep learning models, limiting its ubiquity, reliability, and scalability. In contrast to data-driven communication, physics-inspired intelligent communication (PIC) integrates prior physical laws in wireless issues into a deep learning process, significantly improving the quality and availability of wireless AI. In this paper, we introduce the significance of PIC in the development of wireless AI. First, we explain the core idea of PIC, and demonstrate its typical technical schemes in channel state information (CSI) compression and feedback, channel prediction, and user positioning tasks. Then, we describe how to use multiple PIC technologies to solve complex wireless issues, using mobile channel prediction as an example. Finally, we discuss the prospects and challenges of PIC in wireless systems.

Keywords

wireless artificial intelligence, physics-inspired intelligent communication

1 Introduction

Artificial intelligence (AI) technologies represented by deep neural networks (DNNs) are boosting the development of numerous scientific and industrial technologies by effectively extracting implicit features, representing high-dimensional data, and making complex decisions. This has led to the convergence of AI and communication being seen as a promising vision and a necessity to solve many key technical challenges in the evolution to 6G networks [1]. Over the past few years, deep learning technologies have been applied to a plurality of typical wireless tasks, including channel state information (CSI) compression and feedback [2], channel prediction, and high-precision wireless positioning [3], bringing significant gains compared with conventional signal processing.

Despite this, most existing wireless AI models are designed based on superficial characteristics of wireless channels in terms of data structure and numerical feature. These models are analogous to conventional images or sequences and therefore processed like computer vision models and natural language processing models. Leveraging the neighborhood characteristic of an angular-delay domain channel, the study in [2] processes CSI feedback tasks as image compression by using convolutional neural networks (CNNs). The study in [4] takes a time-varying channel as a time series, and uses a long short-term memory (LSTM) network to process a channel prediction task. And taking a time-varying channel as a time series, the study in [4] uses an LSTM network to process a channel prediction task. In essence, wireless channels are determined by the base station location, user mobility, and signal propagation mechanism. These channels have clear internal physical characteristics and are significantly different from the prior conditions of classical image and sequence models. Each index of a channel matrix corresponds to a clear physical quantity, such as an antenna and subcarrier, and therefore does not have translational invariance in an image. But because the interval and wavelength of an electromagnetic wave are typically far smaller than the scale of spatial-temporal sampling on a channel, a channel sequence has fluctuating phase changes and is extremely unsmooth. However, prediction precision of a conventional sequence model usually depends on the numerical correlation and smoothness of sequence data. The mismatch between the intrinsic characteristics of wireless channels and the priors of AI models limits the performance of AI models in wireless tasks. As a result, existing wireless AI technologies may fail to meet expectations in terms of training costs,

generalization, and reliability, hampering the application of AI technologies in wireless systems.

In order to reduce the reliance of deep learning on data and training processes and to further improve the performance and generalization capability of wireless AI technologies, researchers have turned to physics-inspired methods. One of the core ideas of such methods is to integrate a learning process with universal physical laws, thereby integrating the advantages of data-driven DNNs and model-driven physical laws into the learned intelligence. Specifically, physics-inspired intelligent communication (PIC) does not select or combine DNNs based on data structures and data forms in tasks; instead, it explores the physical significance (also known as inductive biases) of tasks and data in wireless systems. These biases are transferred to neural networks through a prior design, thereby guiding the networks to learn wireless intelligence that more closely represents the physical characteristics. This approach fundamentally meets the diverse requirements of wireless systems [5]. Several studies [6–11] have been conducted to preliminarily explore and analyze the design criteria and utility of PIC on some specific wireless tasks, demonstrating the significant potential of physics-inspired wireless AI technologies in terms of performance, generalization, and application flexibility.

In this paper, we provide a comprehensive analysis, evaluation, and prediction of PIC technologies based on existing technical works. Specifically, we analyze representative design ideas and technical schemes, further evaluate the significance of PIC in wireless systems, and predict the trends and future research directions of physics-inspired wireless AI.

This paper is organized as follows:

- **Section 2: Technical Framework of PIC.** We describe how to design a PIC scheme covering the connection structure, activation function, loss function, and non-fully black-box model of neural networks, using specific wireless AI tasks (i.e., CSI compression and feedback, static channel prediction, and user positioning) as examples.
- **Section 3: Comprehensive Application of PIC.** We introduce how to use multiple PIC technologies to handle complex wireless tasks based on the physical characteristics of the wireless environment, using mobile channel prediction — a challenging task — as an example. We demonstrate the performance advantages of PIC compared with traditional wireless AI schemes through experiments.
- **Section 4: Challenges Faced by PIC and Future Research Directions.**

2 Technical Framework of PIC

Given the significant potential of PIC technologies in improving the performance and efficiency of wireless communication, it is crucial to explore the in-depth physical mechanism of wireless issues and design a neural network architecture that adapts to these issues.

This section will elaborate on the technical framework of PIC, as shown in Figure 1. Considering the internal physical mechanism of a wireless communication scenario, a physics-inspired design should take the following four aspects into account:

- **Connection structure design.** For example, a neural network structure is designed by using a two-dimensional (2D), sequence property of a spatial-frequency domain channel.
- **Activation function design.** For example, an interval activation function is designed to capture the spatio-temporal phase change of a channel.
- **Loss function design.** For example, a cumulative loss function is designed to converge positioning information of CSI at a plurality of frequency levels.
- **Non-fully black-box model design.** For example, an ordinary differential equation (ODE) is used to model a spatial domain change of a channel, and learn a spatial gradient of the channel using DNNs.

In the following sections, we will analyze the internal physical mechanisms of critical wireless tasks such as CSI feedback, channel prediction, and user positioning, and describe PIC design schemes covering the neural network connection structure, activation function, loss function, and non-fully black-box model.

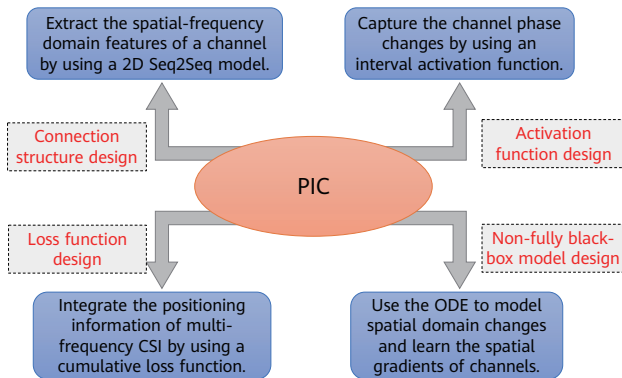


Figure 1 Technical framework and design cases of PIC

2.1 Connection Structure Design

In a multi-input multiple-output orthogonal frequency-division multiplexing (MIMO-OFDM) communication system, the base station side needs to obtain the CSI of downlink channels in order to perform beamforming, carrier allocation, and power control. In frequency division duplex (FDD) mode, the downlink CSI is estimated on the user equipment (UE) side and fed back to the BS side. However, as wireless communication systems evolve, the number of antennas and subcarriers in the MIMO-OFDM system increases, resulting in high CSI feedback overhead. Effectively compressing CSI to reduce feedback overhead has therefore become a major topic of focus.

The study in [2] proposes CsiNet, which utilizes CNNs to extract and compress CSI features for representation by converting a channel from the spatial-frequency domain to the angular-delay domain. This design leverages channel sparsity and correlation of the angular-delay domain. However, unlike natural images, the indices of CSI elements in either the spatial-frequency domain or the angular-delay domain have clear physical meanings, thus making the index location information critical to the CSI. But when the CSI is fed to a CNN, the index location information is lost due to the translational invariance of CNNs. In addition, although the angular-delay domain of channels is correlative to other domains, it does not support smooth translation like an image does, posing a challenge to the CNN-based feature extraction.

According to the expression form of the MIMO-OFDM multipath channel, the element of the CSI matrix $\mathbf{H} \in \mathbb{C}^{N_t \times N_c}$ in the spatial-frequency domain is expressed as

$$\mathbf{H}_{n,m} = \sum_{p=1}^P \alpha_p e^{-j2\pi(f_0 + (m-1)\Delta f)\tau_p + j\varphi_p} e^{-j\chi(n-1)\cos\theta_p} \quad (1)$$

where $\alpha_p e^{j\varphi_p}$, τ_p , and θ_p respectively refer to the complex coefficient, the latency, and the arrival angle of the p th path. In $\chi = 2\pi d f_c / c$, d indicates the antenna spacing, and Δf refers to the subcarrier spacing. In Equation 1, considering the m th column of \mathbf{H} , (spatial domain CSI on the m th subcarrier), the CSI difference between any two adjacent antennas is caused by the spatial domain phase change $\chi \cos\theta_p$ of P paths. Similarly, for the n th column of \mathbf{H} (frequency domain CSI on the n th antenna), the CSI difference between any two adjacent subcarriers is caused by the frequency domain phase change $2\pi\Delta f\tau_p$ of P paths.

Therefore, the spatial-frequency domain CSI matrix \mathbf{H} has the feature of a unique 2D sequence. To extract the features, we divide \mathbf{H} into $L_{\text{ver}} \times L_{\text{hor}}$ cells, and design a model of 2D LSTM cells, as shown in Figure 2. Each 2D LSTM cell can transfer information in two physical dimensions of space and frequency [6].

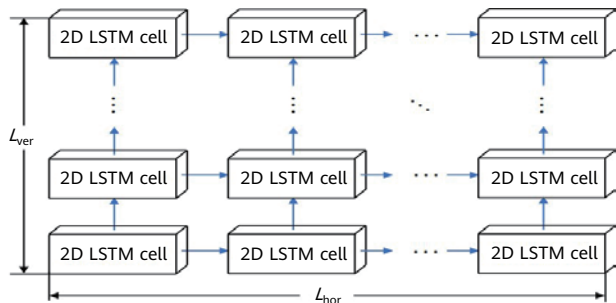


Figure 2 CSI feature extraction based on 2D Seq2Seq of channels

Unlike the neighborhood connection mode of CNNs in CsiNet, this model designs a new 2D sequence-to-sequence (Seq2Seq) connection structure based on the physical model of multipath channels to extract sequence features in the spatial domain and frequency domain. This model serves as the body of the codec in CSI feedback tasks to facilitate efficiency.

2.2 Activation Function Design

Due to the complexity of scattering environments, channels usually contain numerous unknown parameters. Accurately predicting CSI in real time will therefore incur excessively high signaling overhead. However, in practical communication scenarios, a base station always serves a specified area, and a major scatterer (e.g., a building) in the area does not change significantly. We can therefore leverage historical communication data to train a static channel prediction model, and use a neural network to implicitly represent a scattering environment and electromagnetic wave propagation. This model can predict a channel impulse response (CIR) according to a user's location coordinates, significantly reducing signaling overhead for obtaining the CSI.

A channel prediction task is challenging because it needs to recover a high-dimensional CIR vector from a low-dimensional coordinate. Even a minor spatial offset can cause a huge change in a phase of the CIR. Furthermore, it is difficult for a neural network using a conventional activation function such as ReLU to fit a spatial and temporal high-frequency change of the CIR.

In a static channel, the CIR between any TX-RX pair can be expressed as

$$h(f, \tau) = \sum_{p=1}^P \alpha_p \delta\left(\tau - \frac{d_p}{c}\right) e^{-j2\pi f \frac{d_p}{c}} \quad (2)$$

In propagation loss $\alpha_p \propto 1/d_p$, d_p indicates the length of the p th path. The formula features inverse proportion in amplitude and interval in phase of the CIR. In the far field, inverse proportional functions fit well with a series of radial basis functions (RBFs). The study in [12] shows that the multilayer perceptron (MLP) using the interval activation function is more suitable for implicit neural representation in scenarios with high-frequency details. Based on these properties, we propose a static channel prediction model, called Cosine-Gaussian radial basis function network (C-GRBFNet), in [7]. We extract the coordinates of image points in the scattering environment through the MLP and use the following Cosine-Gaussian radial basis function (C-GRBF) as an activation function:

$$\phi(\mathbf{x}) = \cos(\omega|\mathbf{x} - \mathbf{a}| + b) \cdot \exp(\beta|\mathbf{x} - \mathbf{c}|^2) \quad (3)$$

to simulate channel amplitude and phase responses brought by different propagation paths. To process the complex-valued CIR, the activation function outputs in-phase and quadrature components, as shown in Figure 3.

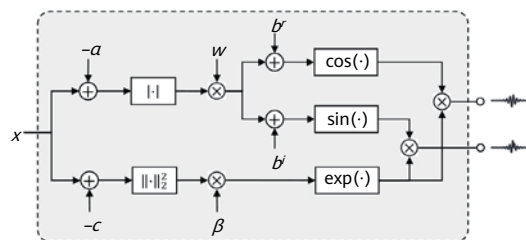


Figure 3 C-GRBF activation function based on channel attenuation and fluctuation

Because a prior structure of a channel response is introduced in the design of the activation function, the C-GRBFNet can capture the channel attenuation and phase characteristics by using only one layer of the C-GRBF function. In the case of low sampling density, the C-GRBFNet outputs at a higher prediction accuracy than the conventional self-coding-based channel prediction scheme and the sinusoidal representation network (SIREN) in [12].

2.3 Loss Function Design

In the next-generation wireless systems, high-precision wireless positioning will become a critical service. Deep

learning technologies are perfectly suited to inferring location information from a CSI fingerprint as it is regarded as a pattern recognition task. In a massive MIMO system, multipath separation can be effectively performed on the CSI of a single subcarrier,

$$\varphi_1: \mathbf{h}(f) \rightarrow \theta_p, \alpha_p \quad (4)$$

According to an ideal reflection model, the path attenuation factor is

$$\alpha_p = \frac{c \cdot \prod_{i=1}^k \gamma_i}{4\pi f d_p} \quad (5)$$

where γ_i indicates the attenuation of the i th reflection. For a fixed scenario and an angle of arrival i , as the total length of a path d_p increases, the quantity of reflection times k increases, and therefore α_p decreases monotonically compared with d_p . This process can be expressed as

$$\varphi_2: \theta_p, \alpha_p \rightarrow \theta_p, d_p \quad (6)$$

Based on the known base station location, angle of arrival θ_p , and propagation path length d_p in a fixed scenario, the user location can be uniquely determined as

$$\varphi_3: \theta_p, d_p \rightarrow \mathbf{x} \quad (7)$$

Therefore, we find a one-to-one mapping relationship between the user location and the single-carrier CSI,

$$\varphi: \mathbf{h}(f) \rightarrow \theta_p, \alpha_p \rightarrow \theta_p, d_p \rightarrow \mathbf{x} \quad (8)$$

Although non-ideal factors such as noise and scattering make it challenging to build an accurate mapping relationship in practical applications, the idea of using single-carrier CSI to locate a user is still valuable.

In [13], we proposed converging the fingerprint positioning results of multiple subcarriers rather than using the entire MIMO-OFDM CSI matrix \mathbf{H} as a fingerprint. Based on this design idea, we proposed the multi-carrier cumulative learning neural network (MCCNet). The LSTM is used to transfer the fingerprint features extracted by each subcarrier and output the positioning results after multi-carrier convergence. In the training phase, to ensure that the accumulated positioning results $\mathbf{x}_1, \dots, \mathbf{x}_{N_c}$ gradually approach the real coordinate \mathbf{x} , we designed the following cumulative learning loss function:

$$Loss(\Theta_{MCC}) = \sum_{n=1}^{num} \sum_{i=1}^{N_c} \|w_i \cdot (\mathbf{x}_i - \mathbf{x})\|_2^2 \quad (9)$$

where the weighting coefficient is

$$w_i = \frac{2i}{N_c + 1}, i = 1, 2, \dots, N_c \quad (10)$$

In the inference phase, we take \mathbf{x}_{N_c} , which accumulates the best subcarrier fingerprint information, as the final positioning result. Figure 4 shows the overall training and inference process.

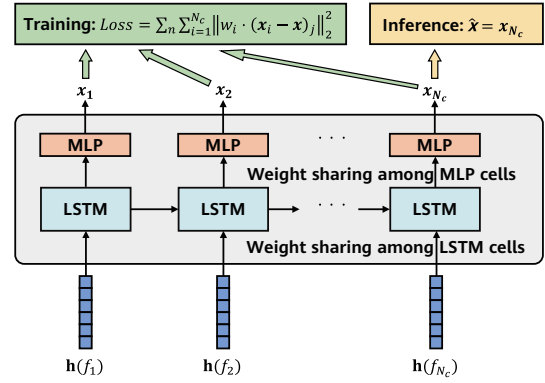


Figure 4 Positioning scheme based on single-carrier fingerprint extraction and multi-carrier convergence

Thanks to the potential mapping between massive MIMO channels and user locations, the single-carrier feature extraction and multi-carrier cumulative learning schemes are integrated with more sufficient physical priors. In this way, the MCCNet has significant performance advantages in positioning tasks compared with the scheme in [3], in which a CNN is used to perform feature extraction on an angular-delay domain channel.

2.4 Non-fully Black-Box Model Design

The PIC schemes we introduced earlier for the connection structure, activation function, and loss function of wireless AI models make the model prior more suitable for the physical characteristics of wireless issues, resulting in improved performance and generalization capabilities. However, due to the black-box training mode of neural networks, performing physics-inspired design on only some modules of neural networks cannot ensure that the neural networks operate in compliance with physical laws. This also affects the explainability of wireless AI models. To further integrate physical information into the design of the wireless AI models, we can directly embed parsed or partially parsed physical equations into the architecture of the wireless AI models, thereby implementing the non-fully black-box PIC scheme.

In a quasi-static scattering environment, there is a one-to-one mapping relationship between user coordinates and channel responses. Therefore, the derivative $\frac{\partial \mathbf{H}}{\partial \mathbf{m}}$ of the channel matrix \mathbf{H} to the spatial displacement \mathbf{m} is related only to the current channel \mathbf{H} and the user movement direction θ_m . This drives us to use the ODE to model the spatial change rate of the channel

$$\frac{\partial \mathbf{H}}{\partial \mathbf{m}} = f(\mathbf{H}, \mathbf{m}) \quad (11)$$

However, due to numerous unknown parameters in the scattering environment, the spatial gradient function $f(\cdot)$ cannot be explicitly represented. In the study of [9], we use the neural ODE to implicitly identify the spatial gradient $\frac{\partial \mathbf{H}}{\partial \mathbf{m}}$ of the channel \mathbf{H} , which is called the spatial channel gradient network (SCGNet). As shown in Figure 5, after the network completes training, a channel at the current location can be predicted from a historical channel by forward integration as follows:

$$\mathbf{H}_{\text{pred}} = \mathbf{H}_0 + \int_0^S f_{\theta}(\mathbf{H}, \mathbf{m}) d\mathbf{m} \quad (12)$$

In this design, the channel prediction model does not directly learn the high-dimensional mapping function from the spatial location \mathbf{x} to the channel matrix \mathbf{H} . Instead, it learns the spatial change rate of the channel, and then completes the channel prediction task through the parsed integral operation based on the historical CSI and user locations. This design significantly reduces the model learning difficulty.

3 Comprehensive Application of PIC

The PIC technical routes we discussed in the previous four sections complement — rather than being isolated from — each other. To address real-world wireless AI issues, various PIC technical schemes need to be used based on specific physical models and application scenarios.

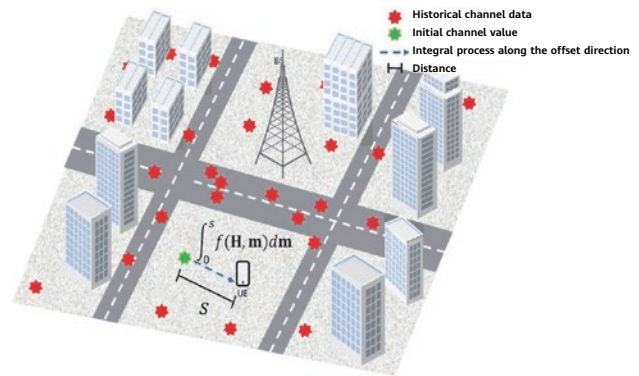


Figure 5 Channel prediction scheme based on spatial gradient learning

In a specific MIMO-OFDM mobile communication scenario, the quantity and locations of main scatterers are fixed. However, users' movements cause Doppler frequency shifts on a channel. Furthermore, Doppler frequency shifts of different moving speeds and different subcarrier frequencies are different. This makes it far more complex to predict a mobile channel than it is to predict a static one. To improve channel prediction, multiple PIC technologies need to be comprehensively used, and the physical characteristics of a MIMO channel and users' motion information need to be fully explored.

3.1 Physics-inspired Mobile Channel Prediction Scheme

Using the channel sequence characteristics [6], we perform physics-inspired design for the connection structure. We directly embed the physical equations of channel spatial domain changes into the architecture of the wireless AI model based on the neural ODE [9]. This approach implements the non-fully black-box PIC scheme. Figure 6 shows the overall procedure, which includes three main steps: iterative positioning and motion information extraction, static channel prediction based on the neural ODE, and angular-delay domain Doppler compensation.

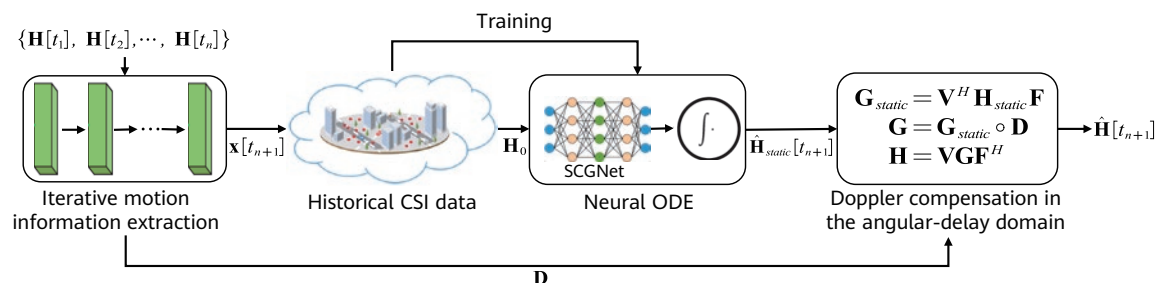


Figure 6 Neural ODE-based mobile channel prediction

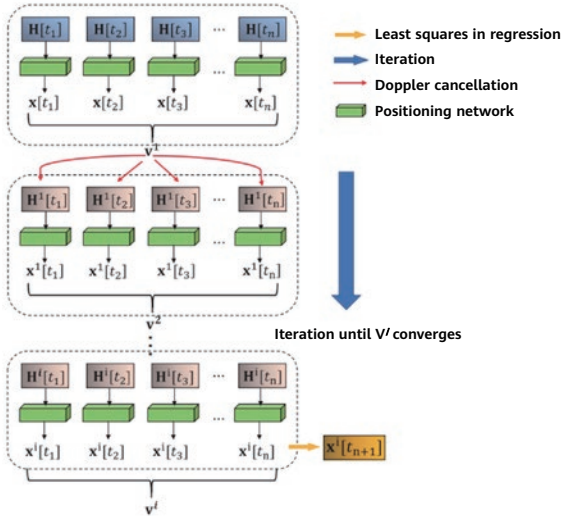


Figure 7 Iterative update of users' positions and velocity vectors

Figure 7 illustrates the iterative positioning and motion information extraction network. By using the LSTM-based static channel CSI fingerprint positioning network in [13], we can establish a time-related location sequence based on the CSI in different timeslots. If a user's motion in the timeslot is modeled as uniform linear motion, the position sequence can be fitted through least squares in regression as follows:

$$\mathbf{x} = \mathbf{v}^i t + \boldsymbol{\sigma} \quad (13)$$

where \mathbf{v}^i is a velocity vector of the i th iteration, and $\boldsymbol{\sigma}$ is a constant vector. Then, Doppler cancellation is performed on the channel response according to the estimated velocity vector. The CSI sequence obtained after the Doppler cancellation is then input into the positioning network again. This makes it possible for us to obtain a new location sequence, thereby achieving the iterative update of users' positions and velocity vectors. The CSI fingerprint positioning network introduces the concepts of single-carrier feature extraction and multi-carrier cumulative learning. The iterative algorithm uses the time series characteristics of channel data and combines the prior knowledge of users' motion modes, improving the accuracy and robustness of positioning and speed estimation.

The design of the SCGNet complies with the physical differential equations of channel variations in the angular-delay domain

$$\frac{\partial g_{\tau,\theta}^{real}}{\partial \mathbf{m}} = \left(-\frac{1}{d_p} g_{\tau,\theta}^{real} + \rho g_{\tau,\theta}^{imag} \right) \frac{\partial d_p}{\partial \mathbf{m}}, \quad (14)$$

$$\frac{\partial g_{\tau,\theta}^{imag}}{\partial \mathbf{m}} = \left(-\frac{1}{d_p} g_{\tau,\theta}^{real} - \rho g_{\tau,\theta}^{imag} \right) \frac{\partial d_p}{\partial \mathbf{m}}. \quad (15)$$

where $g_{\tau,\theta}^{real}$ and $g_{\tau,\theta}^{imag}$ respectively represent the real part and the imaginary part of a matrix element whose angle of arrival is θ and whose delay of arrival is τ in the angular-delay domain channel matrix. Specifically, the SCGNet consists of the scattering learning network and the direction embedding network, as shown in Figure 8. The scattering learning network adopts a fully connected structure to learn the mapping from the CSI matrix to $-\frac{1}{d_p}$. The direction embedding network takes the direction vector as input to obtain direction derivative $\frac{\partial d_p}{\partial \mathbf{m}}$. The outputs of these two networks are combined through the matrix operation to obtain the spatial gradient of the channel matrix. Therefore, the SCGNet is a neural ODE that performs physics-inspired design on a connection structure. After the training is completed, forward integration may be performed on the SCGNet's output in order to predict a static channel at a current user location from historical CSI.

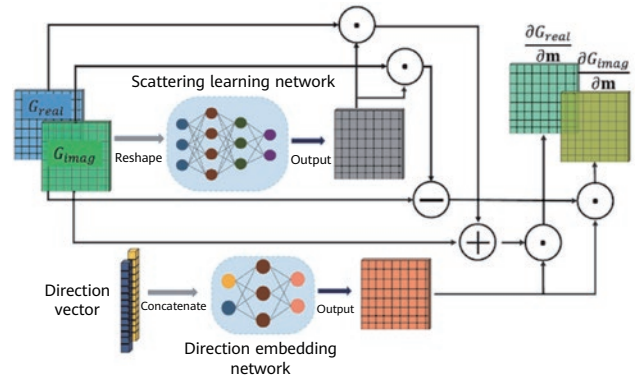


Figure 8 Structural diagram of the SCGNet

Finally, Doppler compensation is performed on the static channel prediction result in the angular-delay domain based on the user's velocity vector estimated in the first step. This enables us to obtain the prediction result of the user's mobile channel.

3.2 Simulation Results

We modeled the surrounding environment of the Xindian Building (High-tech Building) of Zhejiang University in 3D (shown in Figure 9) and used the Wireless InSite of Remcom to perform ray tracing to generate a channel dataset for simulation verification. Building 1 is 25 m high, building 2 is 35 m high, and buildings 3 and 4 are 8 m high. The building material is set to concrete. Area 5 is a wood. Users are distributed in a 120 m x 60 m area. The frequency

at the electromagnetic wave center is set to 3.5 GHz. The BS is located 10 m above building 2 and is equipped with a uniform linear array (ULA). The OFDM system bandwidth is 100 MHz, and the maximum number of calculation paths is 25. Considering the channel acquisition time interval in a system is relatively short, it may be assumed that a user moves in any random direction in uniform linear motions within the interval.

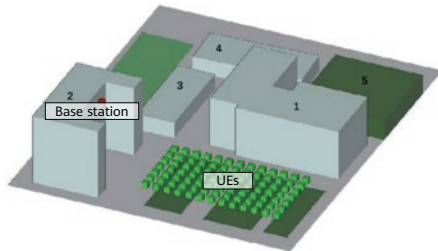


Figure 9 3D model for simulation

To verify the functions of the physics-inspired neural ODE and SCGNet, we compared the proposed scheme with the static database method and with the neural ODE+MLP method under different historical data sampling densities. Figure 10 shows the comparison result. It can be seen that the neural ODE-based learning structure can significantly improve the performance. While the prediction accuracy significantly decreases when the channel sampling density decreases from 50 to 25 in the comparison scheme, it does not decrease significantly in the neural ODE-based method, even at a low sampling density. Moreover, when the network is switched from the MLP to the SCGNet, the prediction accuracy is significantly improved under all sampling density settings. This is because the proposed network integrates prior information of a physical channel change into the network design, rather than directly

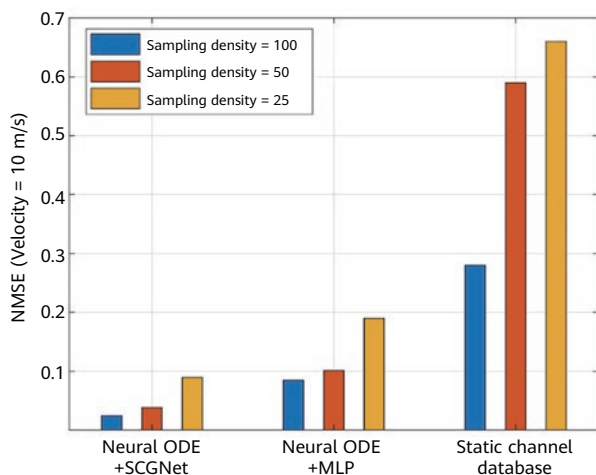


Figure 10 Predicted NMSEs using three methods at different sampling density levels

learning a mapping between CSI and its spatial gradient. Consequently, the difficulty involved in network learning is dramatically lowered.

In order to compare the proposed scheme and the LSTM channel prediction network more comprehensively, we use three different lengths in the LSTM network in the comparison scheme. A sequence interval of a training dataset of the LSTM is 1 ms, and all networks output an angular-delay domain channel matrix. Figure 11 compares the predicted normalized mean square errors (NMSEs) between the proposed scheme and the comparison scheme at different UE speeds when the sampling density is 50. Thanks to the comprehensive use of multiple PIC technologies, the proposed scheme can achieve far higher prediction accuracy than the comparison scheme in any case. In addition, the experiment result proves that prediction performance of a sequence-based network structure is highly sensitive to the UE moving speed. When the UE speed increases, the prediction performance of these networks significantly decreases. This is mainly due to the performance of these networks being highly dependent on the correlation between sequence data. As the UEs move faster, the spatial interval between sequence sampling points also increases correspondingly, and a spatial correlation between sequences decreases. In contrast, the proposed scheme can maintain high prediction accuracy even in the case of fast-moving UEs.

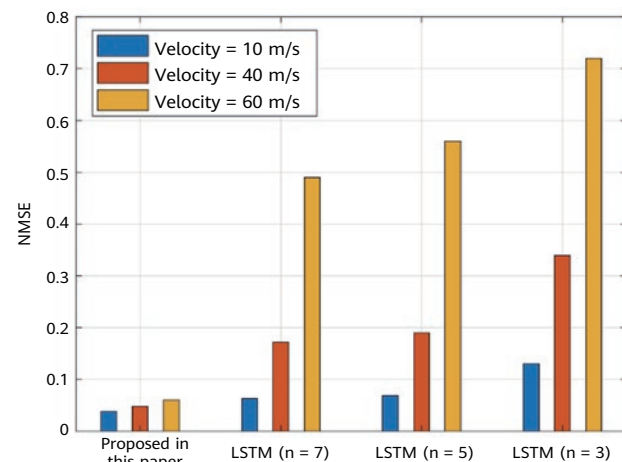


Figure 11 Comparison of predicted NMSEs between schemes at different UE speeds with a sampling density of 50

4 Challenges and Future Research Directions

Although the existing physics-inspired method significantly improves the performance and availability of the wireless AI technology, the PIC technologies are still far from perfect.

First, the physics-inspired method and data-driven methods are typical DNNs and have a similar overall deployment paradigm and working mode even though they differ in learning. This means that some technologies proven to improve the deployment quality of DNNs in wireless systems (e.g., neural network quantization and adaptive inference) can also be integrated with existing PIC schemes to reduce the inference overhead of the PIC technologies and curb costs in AI usage. In addition, more advanced AI architectures, such as Transformer and MLP-Mixer, have been proved capable of describing the physical attributes of wireless signals after proper modeling and variation [10, 11], further improving the utilization efficiency of physical priors.

Second, the physics-inspired method is not limited to improving existing typical wireless subtasks. Instead, utilizing deep learning's exceptional information convergence capabilities, future research can be based on the basic requirements and overall functions of wireless systems. The research should not be limited to replacing existing functional modules but rather should develop a new functional system based on end-to-end communication. Physics inspiration will be one of the guiding principles of this evolution. To build a simpler and more efficient wireless system, we must be closer to the physical nature of electromagnetic transmission and make full use of the physical characteristics of wireless information. This will boost the application of physics inspiration from specific structure design to overall architecture design, and will further unleash the potential of AI methods for wireless network evolution.

Research on the existing PIC technologies focuses on how to properly introduce physics inspiration and adopts the task-specific and scenario-specific intelligent paradigm of the wireless AI technology in an overall framework. This narrow generalization poses challenges to the availability, scalability, and performance of intelligence on difficult usage cases. To overcome the generalization limitations of the existing wireless AI framework design, researchers have envisioned the wireless big artificial intelligence models (wBAIMs) [14]. These models aim to integrate multiple tasks, unify application scenarios, and implement integrated scheduling. The PIC technologies are expected to work perfectly with the wBAIMs. In addition to PIC promoting the establishment of cross-task and cross-scenario wireless AI thanks to the universality of physical laws, the larger learning capacity and more complex inference mechanism of large models can further improve the performance and availability of the PIC technologies.

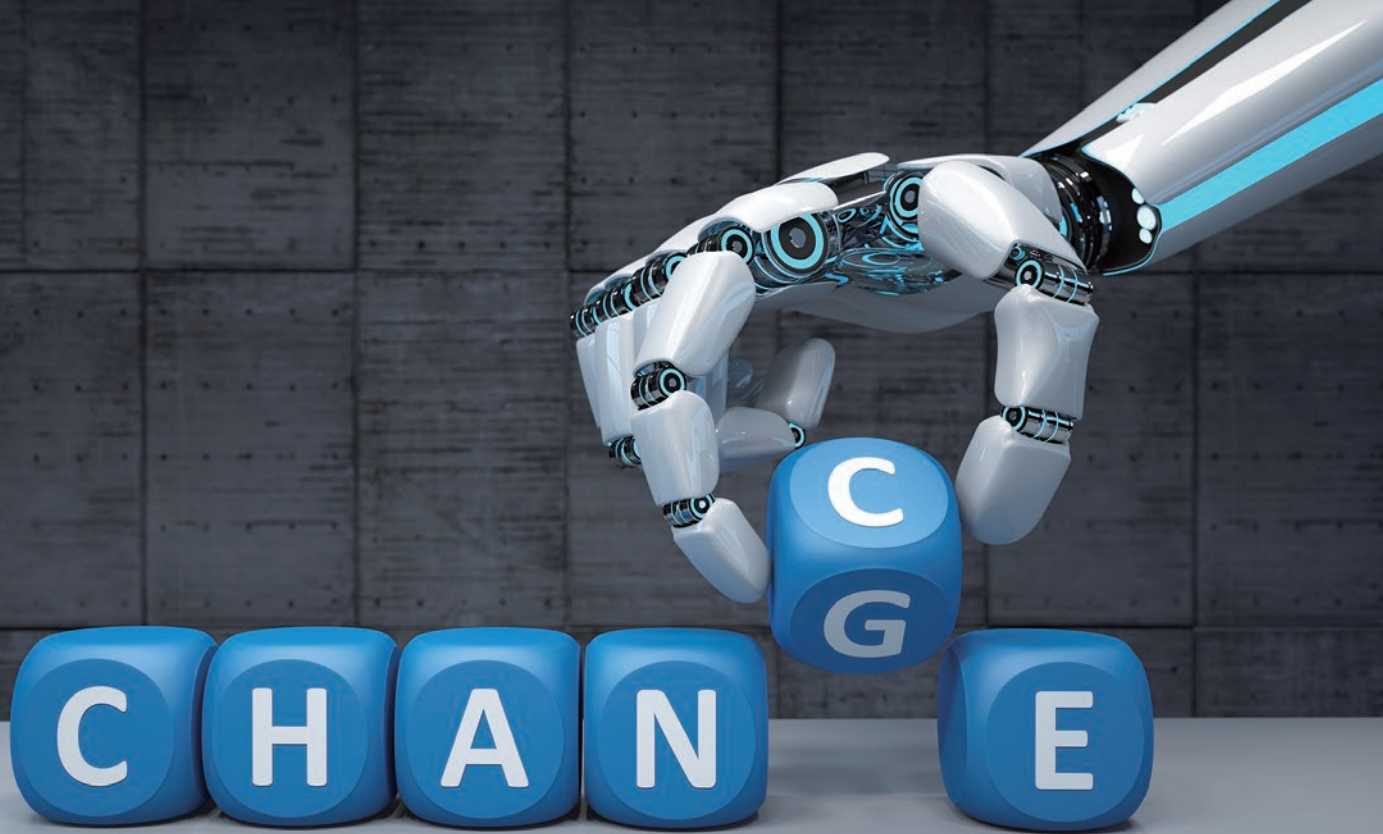
Meanwhile, the PIC technologies have significant potential in emerging technologies such as metaverse and digital twin. By incorporating the prior physical laws of wireless issues in the deep learning process, the PIC technologies can improve communication efficiency, prediction accuracy, positioning precision, and system convergence, laying a solid technical foundation for more abundant and more realistic digital experience.

5 Conclusion

This paper introduces the critical role that DNNs play in 6G networks and points out the generalization and reliability issues caused by DNNs' dependency on massive training data. By introducing the PIC technologies, we can combine physical laws with data-driven AI models to enhance the performance of wireless AI. We discuss the technical framework of PIC, including optimizing the connection structure, activation function, and loss function of neural networks, and designing non-fully black-box models. In addition, we demonstrate the application cases of the PIC technologies in wireless tasks such as CSI compression and feedback, channel prediction, and user positioning. The PIC technologies outperform the traditional wireless AI, proving the application availability of the PIC technologies in practical communication systems. Although the PIC technologies bring significant improvements to wireless systems, it still faces challenges such as model optimization and system complexity. Future research should focus on optimizing the deployment of PIC and redesigning the functional modules of the current wireless systems based on the PIC technologies to further unleash its potential.

References

- [1] I. Recommendation, "Framework and overall objectives of the future development of IMT for 2030 and beyond," *Int. Telecommun. Union ITU Recomm. ITU-R*, 2023.
- [2] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018, doi: 10.1109/LWC.2018.2818160.
- [3] J. Vieira, E. Leitinger, M. Sarajlic, X. Li, and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017, pp. 1–6. doi: 10.1109/PIMRC.2017.8292280.
- [4] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 320–332, 2020, doi: 10.1109/OJCOMS.2020.2982513.
- [5] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019, doi: 10.1109/TCOMM.2019.2924010.
- [6] Z. Chen, Z. Zhang, Z. Xiao, Z. Yang, and K.-K. Wong, "Viewing channel as sequence rather than image: A 2-D Seq2Seq approach for efficient MIMO-OFDM CSI feedback," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 11, pp. 7393–7407, Nov. 2023, doi: 10.1109/TWC.2023.3250422.
- [7] Z. Xiao, Z. Zhang, C. Huang, X. Chen, C. Zhong, and M. Debbah, "C-GRBFnet: A physics-inspired generative deep neural network for channel representation and prediction," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2282–2299, Aug. 2022, doi: 10.1109/JSAC.2022.3180800.
- [8] Z. Chen, Z. Zhang, Z. Xiao, Z. Yang, and R. Jin, "Deep learning based multi-user positioning in wireless FDMA cellular networks," *IEEE J. Sel. Areas Commun.*, p. 1, 2023, doi: 10.1109/JSAC.2023.3322799.
- [9] Z. Xiao, Z. Zhang, Z. Chen, Z. Yang, C. Huang, and X. Chen, "From data-driven learning to physics-inspired inferring: A novel mobile MIMO channel prediction scheme based on neural ODE," *IEEE Trans. Wirel. Commun.*, p. 1, 2023, doi: 10.1109/TWC.2023.3338419.
- [10] Z. Chen, Z. Zhang, Z. Yang, and L. Liu, "Channel mapping based on interleaved learning with complex-domain MLP-mixer," *IEEE Wirel. Commun. Lett.*, p. 1, 2024, doi: 10.1109/LWC.2024.3370303.
- [11] Z. Chen, Z. Zhang, Z. Yang, C. Huang, and M. Debbah, "Channel deduction: A new learning framework to acquire channel from outdated samples and coarse estimate," Mar. 28, 2024, *arXiv*: arXiv:2403.19409. doi: 10.48550/arXiv.2403.19409.
- [12] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 7462–7473. Accessed: Jul. 03, 2024. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html>
- [13] Z. Chen, Z. Zhang, Z. Xiao, C. Zhang, and Z. Yang, "CSI of each subcarrier is a fingerprint: Multi-carrier cumulative learning based positioning in massive MIMO systems," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2023, pp. 1–7. doi: 10.1109/PIMRC56721.2023.10294059.
- [14] Z. Chen, Z. Zhang, and Z. Yang, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *IEEE Wirel. Commun.*, pp. 1–9, 2024, doi: 10.1109/MWC.015.2300404.



Robots Empowered by AI Foundation Models and the Opportunities for 6G

Massimiliano Maule, Anh Vu Vu, Hanwen Cao, Tingzhong Fu, Mohamed Gharba, Daniel Gordon, Joseph Eichinger, Shenfei Zhang, Yiqun Wu, Xueli An, Lei Lu
Advanced Wireless Technology Laboratory

Abstract

Global initiatives highlight the importance of AI in robots. While deep learning (DL) offers effective training, it lacks the ability to generalize. To address this limitation, foundation models (FMs) with massive parameters and pre-training have been developed. By leveraging FMs, robots can autonomously understand and execute tasks from natural language instructions, dynamically decompose complex tasks, and adapt actions based on real-time feedback, minimizing human intervention. This paper analyzes current research and industry efforts towards integrating FMs into robotics, explores the potential of 6G technology for robotics applications, and introduces a prototype 6G robotic system utilizing AI and FMs.

Keywords

robot, foundation model, 6G, ISAC, AlaaS, 3GPP

1 Introduction

The global vision for developing robotic technologies demonstrates the crucial importance of integrating artificial intelligence (AI) into robots. In the United States, the 2024 edition of "A Roadmap for US Robotics: Robotics for a Better Tomorrow" [1] by the National Robotics Initiative (NRI) highlights AI as a pivotal force. The roadmap outlines advancements in machine learning (ML), artificial general intelligence (AGI) research, pervasive automation, and the convergence of AI with robotics. It also emphasizes personalized AI, AI ethics, and AI-driven scientific discovery, all aimed at shaping the economy, workforce, and national security.

The European Union's joint "Strategic Research Innovation and Deployment Agenda (SRIDA)" for the AI, data, and robotics partnership [2] underscores a human-centric and trustworthy approach to AI and robotics. This agenda focuses on fostering collaboration among industry, academia, and policymakers to drive research, development, and deployment. It aims to establish Europe as a global leader in AI and robotics by stimulating investment and tackling key challenges, thereby enhancing economic, societal, and environmental outcomes in alignment with European values and rights.

China's "14th Five-Year Plan for Robot Industry Development" [3] emphasizes the need to enhance the intelligence and networking capabilities of robots through the integration of AI, 5G, big data, and cloud computing. This plan ensures the functionality, network, and data security of robotic systems, thereby advancing the nation's technological capabilities and industrial applications.

To achieve their perception capabilities, classic AI robotic systems use deep learning (DL) methods deployed in a controlled environment. Although this approach provides an effective way of learning multiple skills, it not only requires significant training time and extensive engineering effort to set up each task, but also lacks distribution shifts and generalizability.

While this might sound reasonable for a single task, the learning costs and effort could exponentially increase when multitasking on a real-world experiment is performed, introducing new challenges inside the robotic domain.

Building generalizable robotic systems faces several challenges. At the same time, however, a novel field of study

has emerged that could help enhance robotic systems. A foundation model (FM) is a large-scale AI model that serves as a versatile and general-purpose framework for various downstream tasks by being adapted to specific applications. FMs are pre-trained on internet-scale data, presenting superior generalization capabilities and extending the concepts of transfer learning (TL) and model scaling [4].

They enable robots to autonomously understand and execute tasks from high-level natural language instructions, dynamically decompose complex tasks, and adjust actions based on real-time feedback, minimizing human intervention. Furthermore, situation awareness is enhanced by enabling semantic understanding of the environment using multimodal data from commonly used sensors such as cameras, LiDAR, and microphones.

These advancements shift robots away from rigid, predefined operations and narrowly focused models, moving them towards dynamic, intelligent task execution and environmental understanding, significantly enhancing their autonomy, flexibility, and efficiency.

In this paper, we analyze the current efforts of academics and industries and the future directions they will take in applying FMs to robotics. Furthermore, we analyze the impact of 6G technology on robotics, highlighting the future applications, integration with AI-FM, and networking requirements. This paper is structured as follows: Section 2 provides the state-of-the-art (SOTA) analysis of the current FMs for robotics. Section 3 provides a brief overview of the standardization effort by the main entities. Section 4 broadly illustrates the market and research opportunities of 6G and AI applied to robotics. Section 5 introduces our 6G robotic prototype. And finally, Section 6 presents conclusions, remarks, and future research directions.

2 SOTA Foundation Models for Robots

This section provides an overview of the types, roles, and capabilities of FMs specific for the robotic domain. We use terminology that is consistent with the ISO standard 8373:2021 [5] for robots and robotic devices. This international standard is key for ensuring that communication is clear and consistent across different industries, academic fields, and geographic regions involved in robotics.

2.1 FM Enablers for Robotics

The key benefits of FMs for robots are summarized as follows:

- **Comprehensive knowledge base:** FMs provide robots with extensive, multi-domain knowledge, enabling them to understand and execute a wide range of tasks. This knowledge base allows robots to perform complex operations across various fields, without needing extensive reprogramming for each specific task.
- **Natural language understanding:** FMs possess strong natural language processing (NLP) abilities, allowing robots to comprehend and interact using human language. This simplifies task instruction and communication, enabling users to provide commands and receive feedback in a natural language.
- **Multimodal situation awareness:** FMs enhance robots' multimodal situation awareness by enabling semantic understanding of their surroundings using various sensors, such as RGB cameras, LiDAR, and microphones. Robots can understand the logical and geometrical connections between objects, assess current situations, interpret events, and predict future occurrences.
- **Zero-shot and few-shot learning:** FMs excel in zero-shot and few-shot learning, enabling robots to perform tasks with minimal to no task-specific training. This enhances flexibility and adaptability, allowing robots to handle new tasks and environments without needing extensive retraining.

2.2 FM Macro Typologies for Robotics

FMs have the potential to unlock new possibilities in the robotics domain. Among FMs, a subclass of pre-trained models can be utilized to improve various tasks such as perception, prediction, planning, and control [6]:

- **Large language models (LLMs)** [7]: These models would enable robots to understand natural language instructions and potentially respond with natural language.
- **Vision transformers (ViTs) or multimodal transformers** [8]: These models would be crucial for enabling robots to interpret visual data from their environment through cameras and LiDAR sensors.
- **Embodied multimodal language models** [9]: This is a broader category that could potentially combine the functionalities of LLMs and ViTs in order to allow robots

to understand not only natural language instructions but also the visual context of those instructions.

- **Visual generative models (VGMs):** In terms of the evolution behind the diffusion models [10], VGMs trained on massive datasets can create realistic scenarios for robots to virtually practice tasks. This enhances perception, refines movement, and provides diverse training data [11].

These advancements highlight the potential of using FMs in the field of robotics for the development of models that are more specific to this field rather than just combining existing vision and language models.

2.3 Robotic FMs: Intent Recognition and Visual Reasoning

There has been growing interest recently in transformed-based robotic AI for its strong capabilities of intent recognition and visual reasoning [12]. This architecture uses language embeddings and observations as inputs, and outputs predicted actions. To achieve long-horizon robust and generalizable policies, a vision-language-action (VLA) model applied to language-conditioned robotic manipulation (LcRM) has been introduced for visuomotor control inputs. This approach further reduces the gap between robot physics and AI, improving two main aspects:

- **High-level planning:** A complex language instruction can be converted and divided into a sequence of basic action primitives, which are then executed by low-level controllers. PaLM-E [9], a combination of PaLM [13] and ViT [14], consists of up to 562B parameters and serves as a high-level policy for planning and reasoning.
- **End-to-end learning:** An LLM can be trained to directly generate actions based on instructions and observations. RT-1 [15] and RT-2 [16] are examples of multitask models that tokenize robot inputs and output actions to enable efficient inference at runtime. Such an approach makes real-time control feasible. Similarly, Octo [17] provides training and fine-tuning of generalist robotic policies (GRPs) using transformer-based diffusion methods. Out of the box, Octo supports multiple RGB camera inputs and multi-arm robots, and can be instructed via language commands or goal images. Furthermore, Octo uses a modular attention structure in its transformer backbone. This allows it to be effectively fine-tuned to robot setups with new sensory inputs, action spaces, and morphologies, using only a small target domain dataset and accessible compute budgets.

2.4 Joint Robotic and AI-FM Simulation Platforms

Several frameworks have been developed to simulate robots powered by AI-based planning, control algorithms, or both. Two main families of frameworks have been identified as possible platforms on which to base our analysis. There is a third platform, NVIDIA Isaac Lab [18], but due to the need for a proprietary commercial license, it is not considered.

- **RoboCasa** [19] is a simulation framework for training robots to perform everyday tasks. Methods are provided to train transformer-based models on a combination of proprioceptive robot data (e.g., joint encoder readings) and images (e.g., from a camera on the robot or in the world).
- **MuJoCo** (Multi-Joint dynamics with Contact) [20] is a physics engine specifically designed for simulating physical systems, particularly robots. MuJoCo's realistic simulations can be used to train FMs for various robotic tasks. Such FMs can learn by interacting with the virtual environment, manipulating virtual objects, and receiving feedback on their actions. This training data can then be transferred to the real robot, allowing it to perform similar tasks in the physical world.
- **HABITAT** [21] is a high-performance 3D simulation environment designed specifically for training embodied AI agents, such as robots and virtual assistants. HABITAT simulates various sensors (e.g., RGB-D cameras) commonly used in robots, providing FMs with diverse sensory information for perception and decision-making.

3 Use Cases of Robots Empowered by 6G and AI

In the telecommunication industry, R&D and standardization efforts have been exploring the applications of mobile network in robotics. The 3rd Generation Partnership Project (3GPP) System Aspect 1 (SA1) has studied service robots [22] and identified eight use cases. These include real-time cooperative safety protection, smart communication data collection and fusion using multimodal sensors on multiple robots, and autonomous and teleoperated robots working on mining actuation and delivery. Some technical aspects have been discussed, such as tactile and multimodality communication, integrated sensing and communication (ISAC), metaverse, and high-level communications.

The one6G association aims to evolve, test, and promote next-generation cellular and wireless communication

solutions. It envisions that robotic applications will penetrate several application areas and societal sectors. In addition, it has published a series of openly available whitepapers on 6G and robotics, providing in-depth discussions of 6G's enabling functions to robots (e.g., communication, AI/ML, and ISAC)¹. Furthermore, several use cases of robots empowered by 6G are proposed, such as collaborative robots, disaster relief, action planning, industrial robots, and healthcare assistance.

The EU-funded flagship 6G research projects, Hexa-X and Hexa-X-II [23], have discussed and analyzed various 6G use case and requirements, focusing on autonomous robots that can communicate with each other, with other machines, and with nearby humans to perform individual tasks that contribute to a common cooperative objective. One of those that was discussed and analyzed was cooperating mobile robots (CMRs).

4 Opportunities for 6G

Robot control is commonly divided into four levels: task-level, action-level, primitives-level, and servo-level [24, 25]. With the integration of the AI and sensing capabilities of 6G, robots are poised to achieve an even higher level of intelligence, surpassing traditional task-level control. We envision these enhanced capabilities as part of a new level named meta-level. At this level, robots will be able to — in a fully autonomous manner — identify problems, define their tasks, and adapt to dynamic environments based on meta-definitions of their roles, missions, and rules, in addition to possessing real-time situation awareness. Figure 1 illustrates the interoperability between the levels, ISAC functionalities, and native AI infrastructure.

Our vision of how the control levels will be defined for future intelligent robots is as follows:

- **Meta-level control:** This level empowers robots to autonomously identify problems, define tasks, and adapt to dynamic environments based on meta-definitions of their roles, missions, and rules, with real-time situation awareness.
- **Task-level control:** This level defines the overall goals and missions of robots, involving high-level planning, decision-making, and task decomposition. Examples include "Clean the kitchen floor" and "Serve a low-calorie sparkling drink."

¹ <https://one6g.org/resources/publications/>

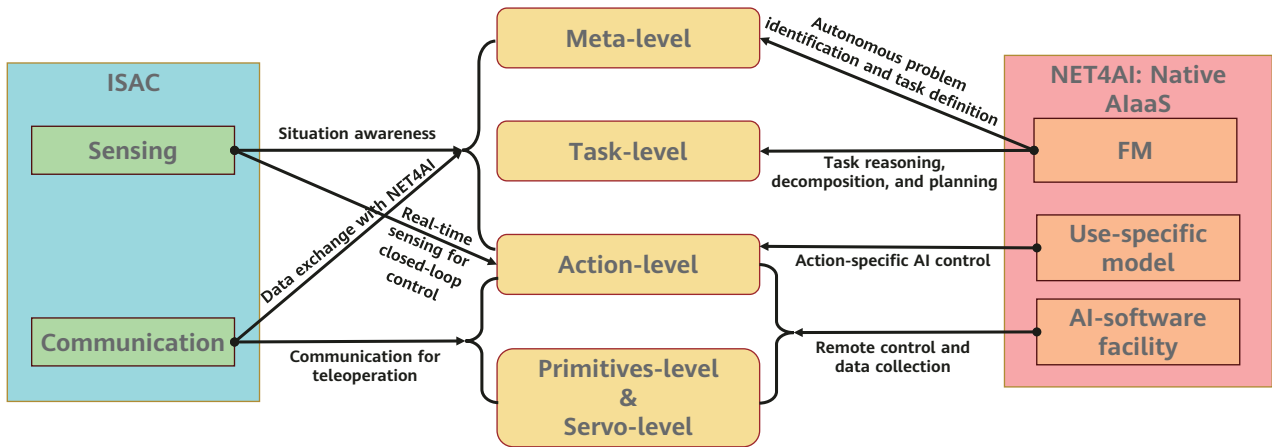


Figure 1 6G capabilities applied to different control levels of robots

- **Action-level control:** This level converts task-level commands into specific movement sequences, including trajectory planning and path generation. An example is planning a path to navigate from the living room to the kitchen without running over a child's toys.
- **Primitive-level control:** This level involves direct control of the robot's actuators to follow planned trajectories, generating commands for joint positions, velocities, and forces. An example is controlling the arm to move precisely along a path to pick up an object.
- **Servo-level control:** This level, the lowest one, focuses on maintaining precise control of actuators through feedback loops. It ensures the execution of commands with high accuracy and stability.

The new ISAC and Network-for-AI (NET4AI) [26] features, borne from the 6G vision and the initial research and standardization efforts, could become important enablers for future robots that are empowered by AI FMs.

4.1 Native AI as a Service, Accommodating AI Models and Computing Facilities

6G aims to provide AI as a service (AlaaS) enabled by NET4AI, embedding FMs and other specific AI models directly within the network infrastructure. This integration provides several key benefits:

- **Low-latency performance:** Embedding AI models within the 6G network significantly reduces latency. Processing data close to the source within the radio access network (RAN) and core network (CN) minimizes the need to

transmit data to external servers for processing, resulting in faster response times.

- **Access to rich data:** AI models within the 6G framework have access to a wealth of data from the RAN and CN, as well as from ISAC. Access to the extensive volume of data enables more accurate and context-aware AI decision-making, enhancing the performance of AI-driven applications.
- **Enhanced data integration:** The seamless integration of sensing and communication in 6G allows AI models to utilize diverse data sources for more robust and holistic analysis. This integration supports advanced applications like real-time environmental monitoring, adaptive robotic control, and dynamic resource management.

Compared to conventional multi-edge computing (MEC), 6G AlaaS offers improved latency and bandwidth efficiency. It achieves this by embedding AI capabilities directly within the network infrastructure, thereby reducing additional data routing between edge servers and the cellular system. Furthermore, 6G native AI models can access a broader range of data from across the entire network (including ISAC data), leading to more informed AI processing and improved service delivery. The 6G framework also supports dynamic allocation of AI resources in different RAN and CN entities, enabling AI service deployment to be more scalable and flexible. And in comparison to onboard robotic AI systems, 6G native AI offers significant advantages. Specifically, running AlaaS in the network typically offers better computing performance and therefore faster system responsiveness than running AI locally does. Offloading intensive AI computations to the network reduces the power consumption and heat dissipation problems associated with onboard processing, extending the operational life of robots

and reducing costs. Moreover, given the large amount of data available from the network, 6G native AI models provide more accurate and context-aware decision-making. To conclude, AlaaS should have the ability to deploy parts of the "brain" flexibly across the local and network nodes depending on the given needs, such as needing to meet challenging safety requirements [27].

4.2 ISAC for Robot Comprehensive Situation Awareness

The 3GPP has begun a study of ISAC, recognizing its potential to revolutionize various applications, including robotics. The SA1 has completed its study on ISAC (FS_Sensing), resulting in 32 ISAC use cases detailed in TR22.837 [28] and service requirements specified in TR22.137 [29]. These documents consider the comprehensive ISAC approach incorporating sensing based on both 3GPP radio networks and non-3GPP sensors, such as cameras and LiDAR.

The ISAC of the future mobile network is beneficial to robot applications in the following aspects:

- **Integrated sensing, communication, and AI in the same standardized network architecture:** Integrating sensing, communication, and AI FMs into a unified 6G network architecture offers transformative benefits for future intelligent robots. This approach enhances real-time decision-making and situation awareness by providing robots with immediate access to comprehensive, real-time data.
- **Networked sensing for comprehensive situation awareness:** Integrating sensing, communication, and AI FMs into a unified 6G network architecture allows robots to achieve comprehensive situation awareness through networked sensing. Instead of relying solely on a robot's onboard sensors, ISAC provides access to a richer array of data from various sensing nodes on the network, including other robots and environmental sensors.
- **Integrated sensing and positioning:** Mobile robots require positioning capabilities in order to find objects and perform navigation. ISAC can be used to improve positioning accuracy by fusing passive sensing and active positioning functions of the mobile network.
- **Sensing digital twin (DT) construction:** Real-time and accurate sensing data is needed to construct DTs for robots. In the future, ISAC might support the creation of precise and dynamic virtual replicas for effective DTs, improving collaboration among multiple robots.

4.3 Enhancing Future Robots with 6G Communication

The advent of 6G communication will significantly enhance future robots by leveraging hyper reliability, ultra-low latency, advanced quality of service (QoS) provisions, and interworking with robotic software and protocols.

- **Hyper reliable and low-latency communication (HRLLC):** HRLLC has become crucial for industrial applications if robot control will be centralized. 6G provides hyper-reliable and stable communication channels with minimum jitter, ensuring smoother operation and synchronization of robotic systems. This is vital for tasks that require high precision and reliability.
- **Advanced QoS framework:** 6G introduces advanced QoS frameworks that dynamically allocate network resources based on the specific needs of AI FMs and specialized robotic applications. Through its enhanced data throughput capabilities, 6G enables efficient transmission of AI training data, sensor data, and real-time analytics, supporting complex decision-making processes and learning algorithms.
- **New protocols for interworking:** 6G's support for seamless interworking with robotic software and communication protocols such as Data Distribution Service (DDS) [30], Open Platform Communications Unified Architecture (OPC UA) [31], Message Queuing Telemetry Transport (MQTT) [32], and Zenoh [33] allow robots to benefit from its capabilities without requiring an extensive redesign of existing systems.
- **Real-time closed-loop teleoperation and training:** 6G enables real-time closed-loop teleoperation of robots by humans or AI. This is crucial for solving unknown complex tasks as well as training AI models to acquire new skills through imitation learning. Through 6G's robust communication infrastructure, operators can remotely control robots in real time, providing hands-on training that accelerates AI learning and adaptation.
- **New business opportunities:** The power of AI, coupled with 6G sensing capabilities of both the robot and the network, unlocks new business opportunities for network owners and robot service providers. Real-time robot operations necessitate the integration of sensing, AI, and control functions with low latency and high data throughput to ensure seamless and efficient performance. Depending on the deployment of AlaaS agents, timely integration of sensing data from various sources is essential. Additionally, robotic operators and

vendors may also favor resource-intensive services through an integrated mobile network solution that ensures contracts and trust for uninterrupted and reliable operation.

5 MELISAC — FM-powered Robot for 6G Proof-of-Concept

In this section, we introduce MELISAC (Machine Learning Integrated Sensing and Communication), our proof-of-concept (PoC) compound robot that integrates several advanced technologies, including intelligent robotic control, online robot training, and ISAC.

5.1 Hardware Setup

MELISAC is a dual-arm compound robot consisting of two industrial articulated collaborative robots (cobots), the UR5e² and an automated guided vehicle (AGV). The UR5e is mounted on an aluminum frame atop the AGV. This configuration enables autonomous navigation and precise object

manipulation. For end-effectors, MELISAC is equipped with MiaHand³, a pair of anthropomorphic robotic hands that allow it to perform tasks in a manner similar to human hands. This capability is particularly beneficial for training AI models that control robots by demonstrating human task execution.

Additionally, an ISAC-capable sub-THz radio system is deployed on the robot, with its antenna mounted either on the body frame or as an end effector. A local computer handles onboard computation for action control and signal processing.

5.2 Software Architecture

In our deployment, sensor data processing and action planning are managed by the local computer, while computationally intensive tasks (e.g., AI inference) are offloaded to edge servers, as illustrated in Figure 2.

- Cobot arms and AMR controllers:** These are the native controllers provided by the robot manufacturers. They expose application programming interfaces (APIs) for executing low-level robot functions (e.g., emergency stop, obstacle detection, and kinematics/inverse kinematics).

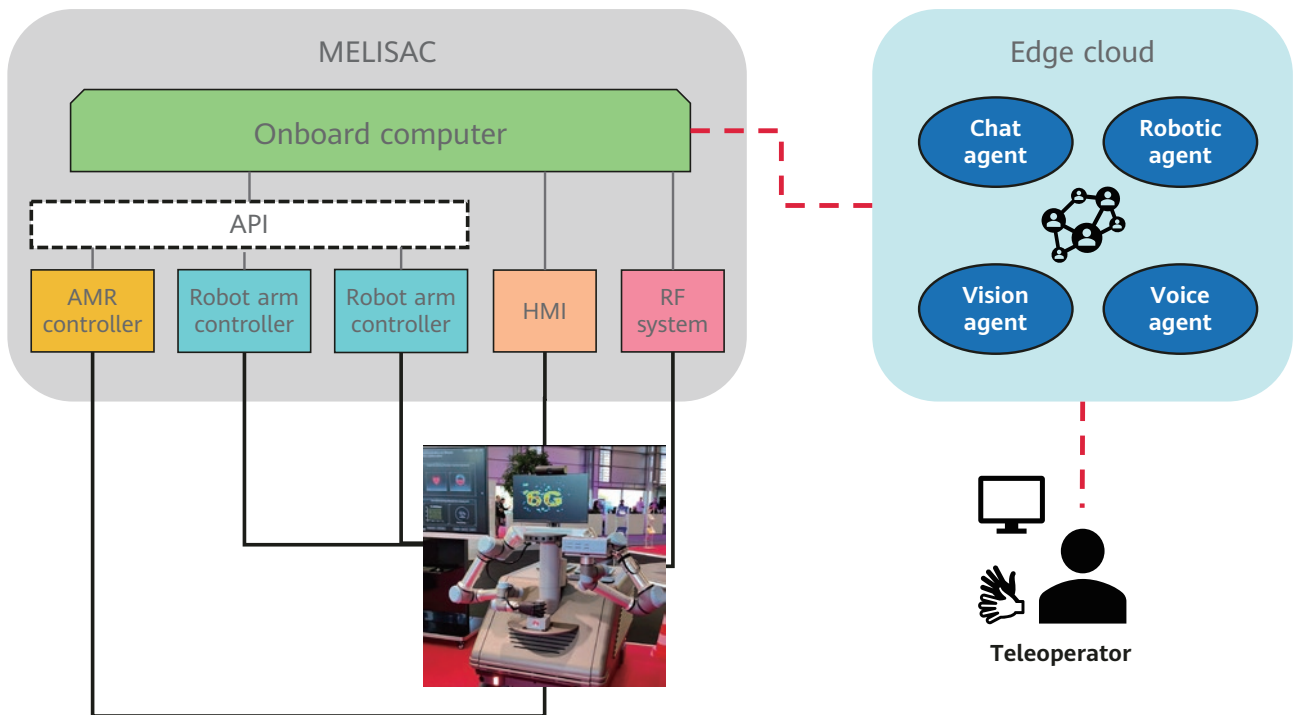


Figure 2 MELISAC in Hannover Messe 2023 and its software architecture

² <https://www.universal-robots.com/products/ur5-robot>

³ <https://www.mia-hand.com>

- **Adaptation API:** This is an adaptation layer that abstracts low-level control for the high-level controller. It is essential for hardware-agnostic FM-based control functions.
- **Human-machine interfaces (HMIs):** These are modalities for human-robot interactions, such as speech and gestures.
- **Radio frequency (RF) sensing:** This refers to the RF system for integrated radio sensing and communication. Radio sensing provides an additional perception layer alongside RGB-D cameras and microphones.

Due to their computation and memory requirements, SOTA FMs need to be deployed on powerful servers located on the edge cloud. Each FM is loaded into an AI agent, which combines its FM with necessary software stacks. The AI agents interact with each other in a text-based multi-agent system located on the edge cloud. The local computer communicates with robot components and the AI agent in the edge cloud using ROS2.

- **Chat agent:** an AI agent powered by an LLM with a large vocabulary and general knowledge capable of engaging in conversations with humans on various topics.
- **Vision agent:** a vision-language FM agent specializing in extracting semantics from video and image inputs, as well as classifying and localizing objects of interest.
- **Robotic agent:** a robotic-FM agent responsible for high-level planning of robot actions based on inputs from the chat agent (user requests) and vision agent (environmental context).
- **Voice agent:** provides real-time speech-to-text and text-to-speech conversion.

Robotic FMs may often struggle with unfamiliar tasks in unstructured environments. In such cases, a human operator can step in to demonstrate the task. MELISAC allows a teleoperator to control it over the network using teleoperation data for training. This human-in-the-loop online training adds an adaptation layer to the pre-trained FM and should be continuous on the cloud.

5.3 Technical Discussions

End-to-end models vs. chain of models: A key question in building FM-controlled robots is whether to use a single end-to-end FM for all input modalities or a pipeline of multiple models. The single-model approach, seen in Octo,

RT-1, and RT-2, often has better generalization because all modalities are trained together and allows real-time control with one inference. The model-pipelining approach, despite offering flexibility, transparency, and customization, incurs extra inference time and integration complexity. Existing frameworks such as Promptflow⁴ and DSPy⁵ can help manage these challenges. The choice depends on data availability and hardware suitability. A domain-specific task with confidential data might benefit from a pipelined vision model with language and action models, whereas an end-to-end model trained on large Internet datasets is better for general tasks.

Integration with robot manufacturers' APIs: Currently, manufacturers provide control stacks with high-level APIs for capabilities like simultaneous localization and mapping (SLAM) and movement, while low-level action control remains restricted for safety compliance. Integrating AI into robots requires extended access to sensors and actuators. Given that full replacement of low-level control by FM-based solutions is unlikely, an integration scheme is needed to embed FM functionalities within existing systems. Retrieval-augmented generation (RAG) [34] can help FMs learn control using standard low-level API documentation. A logical transition step is to define common interfaces between high-level functions (potentially FM-based) and low-level APIs, ensuring both safety and functionality. This requires collaboration between manufacturers and FM developers, with standardization of these interfaces being beneficial but not essential.

6 Conclusions and Remarks

Robotic FMs, despite their impressive ability to grasp basic objects and movements, struggle with complex tasks. They lack a nuanced understanding of real-world physics, hindering their ability to perform actions that require subtle manipulation. Furthermore, high precision and dexterity remain out of reach for current robotic FMs. In addition to these physical limitations, FMs need more than basic instructions for complex tasks and cannot learn intricate skills simply by observing. These shortcomings are compounded by slow control frequencies that restrict their ability to operate in real-time, high-speed environments. Even for tasks requiring smooth, precise movements, FMs

⁴ <https://github.com/microsoft/promptflow>

⁵ <https://github.com/stanfordnlp/dspy>

are not well-suited. On top of all this, training them for entirely new actions without prior examples remains a significant challenge. These limitations, coupled with the lack of reliable and safe robot control systems, highlight the need for significant advancements in robotic FMs.

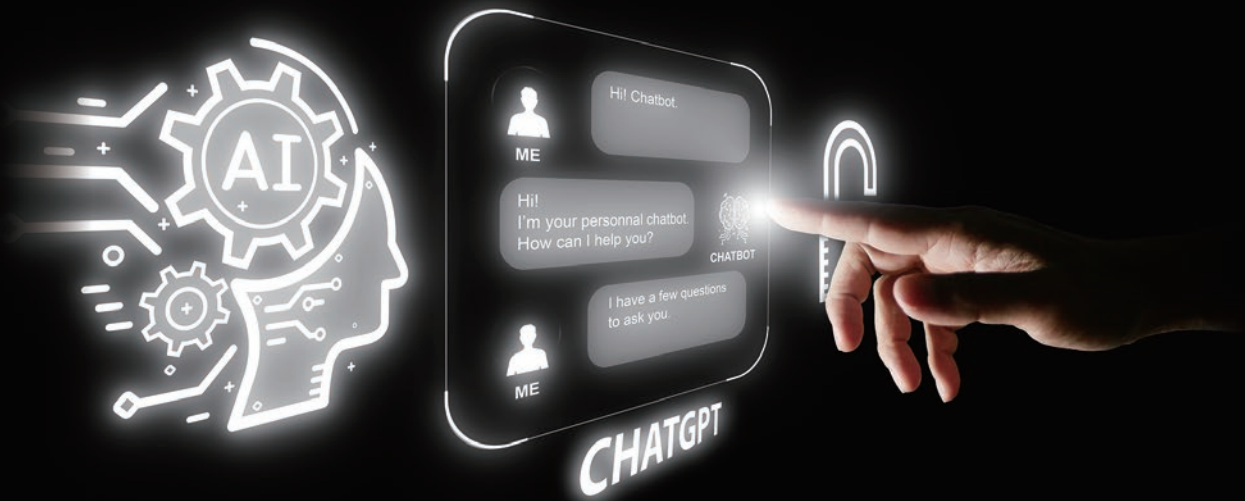
To adapt to future advancements, it is necessary to augment FMs with task-specific AI models, DT technology, and high-performance computing resources. Integration of specialized AI promises to improve precision and dexterity, while DT technology offers advanced physical simulations and AI training. This fosters a deeper understanding and prediction of physical interactions. The development of intelligent hybrid control systems that incorporate high-level planning from FMs, task-specific AI for specialized skills, and traditional methods for low-level execution will ensure smoother and more efficient operations. Additionally, leveraging advanced computing and programming tools to elevate control frequencies and real-time responsiveness will enable robots to handle dynamic tasks more effectively. This comprehensive approach will significantly enhance robotic autonomy, flexibility, and efficiency, empowering them to navigate complex real-world scenarios with greater competence. The advent of 6G, with its advanced AI and sensing capabilities, promises to propel robots beyond traditional task-level control, enabling them to operate at a new meta-level. This will empower them with autonomous problem identification, task definition, and adaptation to dynamic environments. And by leveraging 6G's ISAC and AlaaS, these robots can identify tasks and solve problems with greater autonomy and efficiency, guided by the meta-definitions of their roles, missions, and rules in addition to possessing real-time situation awareness.

References

- [1] "A roadmap for US robotics: Robotics for a better tomorrow," National Robotics Initiative (NRI), 2024.
- [2] "Joint strategic research innovation and deployment agenda (SRIDA) for the AI, data, and robotics partnership," European Union, 2020.
- [3] "14th five-year plan for robot industry development," Chinese Government, 2021.
- [4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.
- [5] I. O. for Standardization, "ISO 8373: 2021 robotics-vocabulary," 2021.
- [6] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, *et al.*, "Foundation models in robotics: Applications, challenges, and the future," arXiv preprint arXiv:2312.07843, 2023.
- [7] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 3, pp. 1–45, 2024.
- [8] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 1, pp. 87–110, 2022.
- [9] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Xu, *et al.*, "PaLM-E: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.
- [10] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

- [11] N. Gothoskar, M. Lázaro-Gredilla, A. Agarwal, Y. Bekiroglu, and D. George, "Learning a generative model for robot control using visual feedback," arXiv preprint arXiv:2003.04474, 2020.
- [12] K. Kawaharazuka, T. Matsushima, A. Gambardella, J. Guo, C. Paxton, and A. Zeng, "Real-world robot applications of foundation models: A review," arXiv preprint arXiv:2402.05741, 2024.
- [13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "PaLM: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
- [14] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian, "Vision transformers for action recognition: A survey," arXiv preprint arXiv:2209.05700, 2022.
- [15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "RT-1: Robotics transformer for real-world control at scale," arXiv preprint arXiv:2212.06817, 2022.
- [16] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," arXiv preprint arXiv:2307.15818, 2023.
- [17] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," arXiv preprint arXiv:2405.12213, 2024.
- [18] F. F. Monteiro, A. L. B. Vieira, J. M. X. N. Teixeira, V. Teichrieb, *et al.*, "Simulating real robots in virtual environments using NVIDIA's Isaac SDK," in *Anais Estendidos do XXI Simpósio de Realidade Virtual e Aumentada*. SBC, 2019, pp. 47–48.
- [19] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlikar, and Y. Zhu, "RoboCasa: Large-scale simulation of everyday tasks for generalist robots," arXiv preprint arXiv:2406.02523, 2024.
- [20] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.
- [21] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "HABITAT: A platform for embodied AI research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9339–9347.
- [22] 3GPP, "3rd generation partnership project; technical specification group TSG SA; study on network of service robots with ambient intelligence (Release 19)," 3GPP, Tech. Rep. TR TR 22.916 V1.0.0 (2023-12), 2023.
- [23] S. Kerboeuf, P. Porambage, A. Jain, P. Rugeland, G. Wikström, M. Ericson, D. T. Bui, A. Outtagarts, H. Karvonen, P. Alemany, *et al.*, "Design methodology for 6G end-to-end system: Hexa-X-II perspective," *IEEE Open Journal of the Communications Society*, 2024.
- [24] D. Kortenkamp, R. Simmons, and D. Brugali, "Robotic systems architectures and programming," *Springer handbook of robotics*, pp. 283–306, 2016.
- [25] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.
- [26] X. Li, "Net4AI: Supporting AI as a service in 6G," *HuaweiTech*, 2022. [Online]. Available: <https://www.huawei.com/en/huaweitech/future-technologies/net4ai-supporting-ai-as-a-sevice-6g>
- [27] C. E. DE NORMALISATION and E. K. F. NORMUNG, "Guidelines for the development and use of safety testing procedures in human-robot collaboration," 2022.
- [28] 3GPP, "3rd generation partnership project; technical specification group TSG SA; feasibility study on integrated sensing and communication (Release 19)," 3GPP, Tech. Rep. TS 22.837 V19.3.0 (2024-03), 2024.

- [29] —, "3rd generation partnership project; specification group TSG SA; service requirements for integrated sensing and communication (Release 19)," 3GPP, Tech. Rep. TS 22.137 V19.1.0 (2024-03), 2023.
- [30] Object Management Group (OMG), "Data Distribution Service (DDS)," 2015. [Online]. Available: <https://www.omg.org/spec/DDS/1.4/PDF>
- [31] Y. S. Bareedu, T. Frühwirth, C. Niedermeier, M. Sabou, G. Steindl, A. S. Thuluva, S. Tsaneva, and N. Tufek Ozkaya, "Deriving semantic validation rules from industrial standards: An OPC UA study," *Semantic Web*, vol. 15, no. 2, pp. 517–554, 2024.
- [32] Organization for the Advancement of Structured Information Standards (OASIS), "MQTT version 3.1.1," October 2014. [Online]. Available: <https://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html>
- [33] Eclipse Zenoh Project, "Zenoh: A protocol for data-centric, resource-efficient, and location-transparent data sharing," May 2020. [Online]. Available: <https://zenoh.io/>
- [34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474.



LLM Application in Wireless Communication Knowledge Management

Hongwei Hou, Chixiang Ma, Lihong Du, Junhui Li

Abstract

Large language models (LLMs) have undergone significant development in recent years, leading to a paradigm shift in the realm of technology. While LLMs have enormous potential in knowledge management due to their advanced capabilities, they face several challenges. First, they are trained solely on general-purpose data from the Internet to maximize accessibility and applicability. This results in suboptimal LLM performance in professional fields due to the lack of professional data in the training process. Second, LLMs often generate seemingly convincing but inaccurate responses, known as hallucinations. To address these challenges, the industry has developed two common solutions: fine-tuning and retrieval augmented generation (RAG). In this paper, we detail the use of RAG technology to design a question and answer (Q&A) solution for wireless communication knowledge bases and an accompanying evaluation solution. The evaluation solution can be used to select the optimal combination of an LLM, embedding model, and rerank model. This model combination is used to successfully implement Q&A for wireless communication knowledge bases with the help of many open-source tools. While the RAG-based integration of an LLM and databases significantly reduces hallucinations, it faces several challenges and limitations concerning multimodal data processing, hyperparameter selection, integration with enterprise knowledge bases and search engines, and introduction of time attributes. Moving forward, we aim to continuously enhance RAG technology and improve its application performance and value in the field of wireless communication knowledge management.

Keywords

large language model, hallucination, fine-tuning, retrieval augmented generation, embedding model, rerank model, database

1 Development of LLMs

In 2005, the use of large n-gram models in machine translation marked the beginning of large language models (LLMs) [1]. In 2017, the Transformer network structure was introduced, which redefined natural language processing (NLP) by incorporating an attention mechanism that significantly improved model performance across multiple tasks [2]. The introduction of Bidirectional Encoder Representations from Transformers (BERT) models in 2018 and 2019 further advanced the development of pre-trained language models (PLMs). BERT effectively utilizes the context information from both the left and right through a bidirectional encoder, achieving state-of-the-art (SOTA) performance on multiple NLP tasks [3]. RoBERTa, an advanced edition of BERT, further improves model performance by adjusting the size of hyperparameters and training data [4].

The launch of GPT-3 in 2020 marked an important milestone in LLM development. GPT-3 enhances an LLM's generalization and few-shot learning capabilities by simply increasing the model size. Additionally, GPT-3 excels in text generation, producing samples of news articles that are indistinguishable from human works [5].

In recent years, LLMs have been increasingly used for multimodal tasks, such as image + text hybrid tasks [6], in addition to conventional text processing tasks. With the rapid advancement of technology, LLMs face new challenges and research interests in terms of adapting to the ever-changing knowledge in real-world applications [7, 8] through knowledge updates.

The evolution of LLMs involves algorithmic and architectural innovation, as well as advanced research on model training, evaluation, and application [9–11], transitioning from simple statistical models to complex neural network models and large pre-trained models. LLMs are expected to advance toward greater explainability, improved efficiency, and optimal integration and processing of multiple data types [6, 12].

2 Essential LLM Technologies in Knowledge Management

LLMs have shown great potential in the field of knowledge management due to their advanced capabilities. However, they also face several significant challenges. First, they are trained using general-purpose data from the Internet

to maximize accessibility and applicability. This lack of professional data in the training process leads to suboptimal LLM performance in professional fields. Second, LLMs often generate seemingly convincing but inaccurate responses, known as hallucinations.

To address these challenges, the industry has developed two common solutions: fine-tuning and retrieval augmented generation (RAG).

2.1 Fine-Tuning

Fine-tuning is a machine learning technology that involves using a small volume of task-specific data to retrain a pre-trained LLM for a new or specific application scenario. This process involves adding one or more output layers to the pre-trained model and using a dataset designed for the task to retrain the model, enabling it to better understand and execute the specific task. Fine-tuning leverages the general knowledge learned by the pre-trained model as a starting point, eliminating the need to train a model from scratch, which can be computationally expensive and time-consuming. BERT is a prime example of fine-tuning. It is pre-trained on a large amount of text data first and then fine-tuned for specific NLP tasks, resulting in significant performance improvements [3].

2.2 RAG

RAG is an innovative approach that combines pre-trained parameterized memory, like LLMs, with non-parameterized memory, such as dense vector indexes from Wikipedia. It dynamically retrieves information from external knowledge resources in language generation tasks, improving the accuracy, diversity, and factuality of the generated content. A typical RAG model includes an LLM as parameterized memory and a retriever that accesses non-parameterized memory such as dense vector indexes [13].

2.3 Advantages and Disadvantages of RAG and Fine-Tuning

We compare RAG and fine-tuning from six dimensions [14]: dynamic data, external knowledge, model customization, reducing hallucinations, transparency, and technical expertise.

Table 1 Advantages and disadvantages of RAG and fine-tuning

Dimension	RAG	Fine-Tuning
Dynamic data	Win	Lose
External knowledge	Win	Lose
Model customization	Lose	Win
Reducing hallucinations	Win	Lose
Transparency	Win	Lose
Technical expertise	Win	Lose

Because RAG demonstrates superior performance in five dimensions, we used RAG in LLMs to improve the performance of wireless communication knowledge management.

3 Solution

3.1 Q&A Solution Design for Wireless Communication Knowledge Bases

The solution comprises two parts: offline construction of wireless communication knowledge bases and online question and answer (Q&A).

3.1.1 Offline Construction of Wireless Communication Knowledge Bases

Figure 1 illustrates the process of offline construction of wireless communication knowledge bases. Initially, users upload different types of documents, such as code files and 3GPP protocols. These uploaded documents undergo parsing, cleaning, and slicing and are sent to the LLM to generate Q&A pairs for each slice, which is optional. Vector indexes and keyword indexes are then created for the Q&A pairs and raw slice data, and are stored in a vector database and common database, respectively. The creation of vector indexes involves an embedding model.

3.1.2 Online Q&A

Figure 1 also illustrates the online Q&A process. A user inputs a question, and the LLM recognizes the user's intent, which is optional. Based on the intent, the LLM selects relevant wireless communication knowledge databases or a handling process. Then, hybrid retrieval is performed to recall the first K knowledge segments that are most

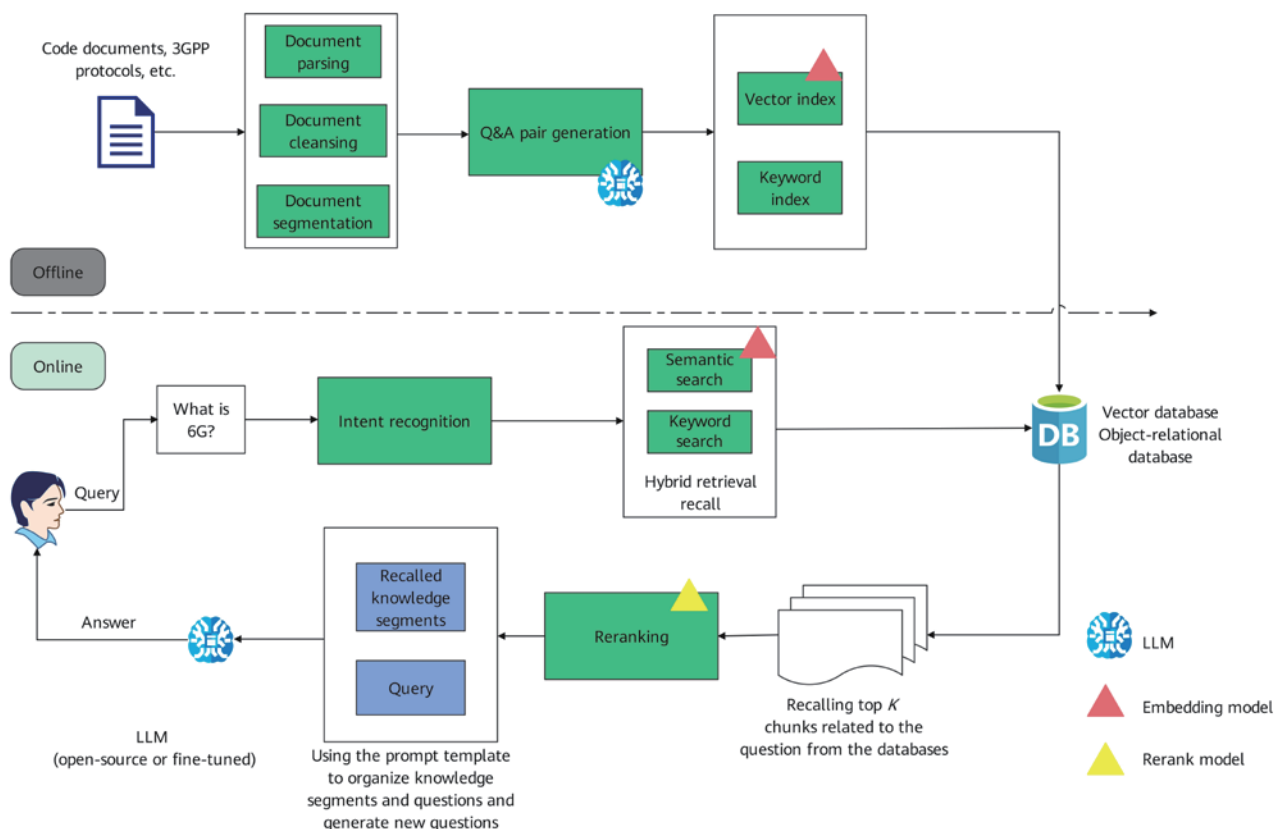


Figure 1 Diagram of using LLMs in wireless communication knowledge management

relevant to the question from the selected databases. These knowledge segments are ranked using a rerank model based on the question to obtain the most relevant N knowledge segments. The question and N knowledge segments are then organized according to a prompt template and sent to the LLM. The LLM provides an answer based on the input and relevant knowledge segments found in the wireless communication knowledge databases.

3.1.2.1 Hybrid Retrieval

Hybrid retrieval involves semantic retrieval and keyword retrieval. In semantic retrieval, an embedding model is used to vectorize the user's question, match the vectors of the question with those in the vector database, and recall K knowledge segments with similar semantics. Keyword retrieval involves searching information from databases based on keywords.

Semantic retrieval supports text with complex semantics and has the following advantages:

- Multi-lingual understanding: English content can be retrieved based on Chinese input.
- Multi-modal understanding: Information can be retrieved for various types of input, such as text, image, audio, and video.
- Advanced fault tolerance: Spelling mistakes and ambiguous descriptions are acceptable.

Despite these advantages, semantic retrieval may deliver suboptimal performance in certain scenarios, for example:

- Searching for a person or item by its name. For instance, the semantic retrieval result of the input "Huawei Mate 60" may include information about Mate 50.
- Searching for an abbreviation or short phrase, for example, "LLM".

In these scenarios, the conventional keyword search approach offers the following advantages:

- Exact match: Product names and person names can be accurately matched.
- Efficient search of short words: Information can be quickly searched based on a few keywords. However, the performance of vector retrieval is unsatisfactory in the case of only a few keywords.
- Able to match words that are used less frequently: Such words often convey more significant information. For instance, in the sentence "Do you want to have a cup of

coffee with me?", the words "have" and "coffee" offer more information than the words "do", "you", or "with".

Hybrid retrieval integrates the unique advantages of vector retrieval and keyword retrieval to search for the most relevant information, which is a major goal in all text search scenarios.

3.1.2.2 Reranking

Hybrid retrieval integrates multiple retrieval technologies to improve the recall rate of search results. It uses a data normalization policy to efficiently process the results of different retrieval technologies. The policy converts data to a standard paradigm or distribution, which can be quickly compared, analyzed, and processed by the LLM. A crucial ingredient to the conversion process is a scoring system — a rerank model.

A rerank model rearranges the retrieval result by measuring the relevance between the documents in the candidate document list and the semantics of the user's query. Relevance is evaluated based on the relevance score of each candidate document. All items are ranked in descending order of the score.

This technique can also be implemented after keyword retrieval in a non-hybrid retrieval system to significantly improve the recall rate. A rerank model can also benefit vector databases, which often trade retrieval accuracy for computational efficiency, leading to uncertainties in the retrieval result. Such uncertainties may disturb the ranking order (descending order by relevance), meaning that the top K segments in the original retrieval result may not be the most relevant ones. In this scenario, a rerank model can be used to reorganize the retrieval result.

Reranking is not a retrieval technology but an enhancement to retrieval systems. With its simplicity and low complexity, it integrates semantic correlations into search systems without requiring any major infrastructure changes.

3.2 Model Combination Evaluation

As shown in Figure 1, we used three types of models in our design: LLM, embedding model, and rerank model. The open-source models were deployed locally, and local documents were used to build wireless communication knowledge bases.

3.2.1 Model Selection

3.2.1.1 LLM

We selected Llama-3-70b-Instruct, Command R+, and Qwen1.5-110B-Chat from the LLM leaderboard "LMSYS Chatbot Arena" [15]. These models support both Chinese and English.

3.2.1.2 Embedding Model

Retrieval is a major indicator for selecting an embedding model, according to the embedding model leaderboard "Massive Text Embedding Benchmark (MTEB) Leaderboard" [16].

We selected 360Zhiniao-search, stella-mrl-large-zh-v3.5-1792d, PEG, and bce-embedding-base_v1 for Chinese and SFR-Embedding-Mistral, gte-large-en-v1.5, GritLM-7B, and bce-embedding-base_v1 for English.

3.2.1.3 Rerank Model

We selected bge-reranker-v2-gemma and bce-reranker-base_v1 by referring to [17, 18]. These models support Chinese and English.

3.2.1.4 Model Combination

To select an optimal combination of Chinese and English models, we evaluated the selected models in terms of Chinese and English: three candidate LLMs, four candidate embedding models, and two candidate rerank models for each language. The candidates formed 24 possible combinations for each language. We used vLLM [19] to run the LLM and Xinference [20] to run the embedding and rerank models.

```
#Create prompt model
template = """You are an assistant for question-answering tasks.
Use the following pieces of retrieved context to answer the question.
If you don't know the answer, just say that you don't know.
Use two sentences maximum and keep the answer concise.
Question: {query}
Context: {context}
Answer:
"""
```

Figure 4 Prompt template

3.2.2 Evaluation Method

We created a Chinese dataset **zh_refine.json** and an English dataset **en_refine.json** based on the open-source project RGB [21]. Figure 2 shows the data format.

```
"id": xxx,
"query": "xxx",
"answer": "xxx",
"positive": "xxx",
"negative": "xxx"
```

Figure 2 Raw dataset format

id indicates the data ID, **query** indicates the question corresponding to the data, **answer** indicates the answer, **positive** indicates the text relevant to the question, and **negative** indicates text irrelevant to the question (interference). The **positive** and **negative** texts of 300 Chinese data records are stored in the same file, which the embedding model sends to the vector database to create a Chinese knowledge base. We used the same approach to create an English knowledge base. The vector database is created based on the open-source vector database Chroma [22].

The evaluation framework is Ragas [23], which requires the data format in Figure 3.

```
data = {
    "question": questions,
    "answer": answers,
    "contexts": contexts,
    "ground_truths": ground_truths
}
```

Figure 3 Data format required by Ragas

question indicates the question, and **ground_truths** indicates the correct answer. The values of these fields can be obtained from either **zh_refine.json** or **en_refine.json**. The LLM generates the values of **answer** and **contexts** according to the prompt template we designed, as shown in Figure 4.

We evaluated the created datasets and the Chinese and English model combinations.

3.2.3 Evaluation Results

The evaluation indicators are **faithfulness**, **answer_relevancy**, **context_precision**, and **context_recall**. For details about each indicator, see [24].

3.2.3.1 Chinese Model Combinations

The total score of each of the 24 Chinese model combinations equals the sum of the scores of **faithfulness**, **answer_relevancy**, **context_precision**, and **context_recall**.

In Figure 5, the horizontal coordinate indicates the name of each model combination, in the format of zh_x_y_z. x indicates the LLM (0: Command R+, 1: Llama-3-70b-Instruct, 2: Qwen1.5-110B-Chat). y indicates the embedding model (0: stella-mrl-large-zh-v3.5-1792d, 1: bce-embedding-base_v1, 2: 360Zhiniao-search, 3: PEG). z indicates the rerank model (0: bce-reranker-base_v1, 1: bge-reranker-v2-gemma).

The optimal Chinese model combination deployed for Chinese scenarios is Command R+, stella-mrl-large-zh-v3.5-1792d, and bge-reranker-v2-gemma.

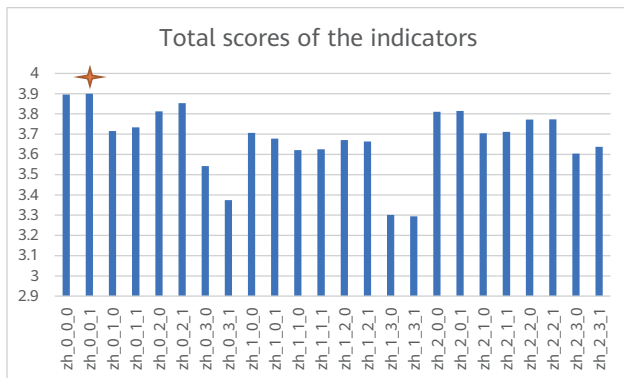


Figure 5 Total score of the indicators of each Chinese model combination

3.2.3.2 English Model Combinations

The total score of each of the 24 English model combinations is equal to the sum of the scores of **faithfulness**, **answer_relevancy**, **context_precision**, and **context_recall**.

In Figure 6, the horizontal coordinate indicates the name of each model combination, in the format of en_x_y_z. x indicates the LLM (0: Command R+, 1: Llama-3-70b-Instruct, 2: Qwen1.5-110B-Chat). y indicates the embedding model (0: SFR-Embedding-Mistral, 1: bce-embedding-base_v1, 2: gte-large-en-v1.5, 3: GritLM-7B). z indicates the rerank model (0: bce-reranker-base_v1, 1: bge-reranker-v2-gemma).

The optimal English model combination deployed for English scenarios is Llama-3-70b-Instruct, SFR-Embedding-Mistral, and bce-reranker-base_v1.

3.3 Implementation Results

We used Dify [25] as the bottom-layer framework to implement the Q&A solution for wireless communication knowledge bases. We employed the selected optimal model combinations to create wireless communications knowledge bases based on Huawei documents. The RAG processes are integrated into workflows, which form the LLM application software.

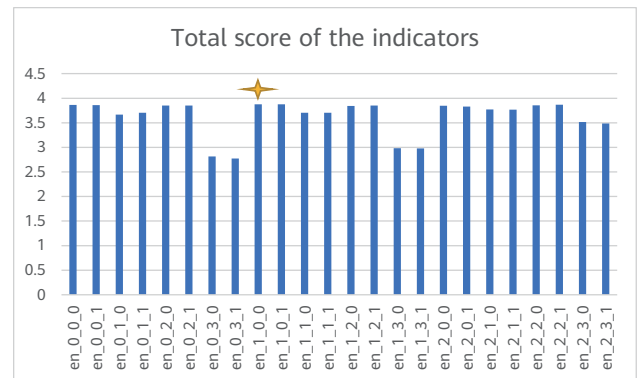


Figure 6 Total score of the indicators of each English model combination

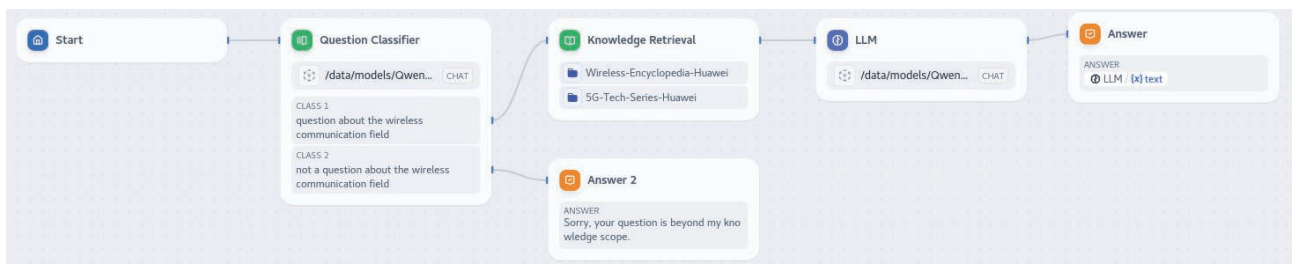


Figure 7 Implementation of the Q&A solution for wireless communication knowledge bases

The question classifier is responsible for intent recognition. If the question concerns 5G, the LLM application software performs RAG until an answer is generated. If the question is unrelated to wireless communications, the software politely refuses to answer.

3.3.1 Questions Related to Wireless Communications

In Figure 8, the user asks a question about wireless communications. The LLM application software performs

RAG to generate an answer and provide reference documents.

3.3.2 Questions Unrelated to Wireless Communications

In Figure 9, the LLM application software politely refuses to answer any question unrelated to wireless communications.

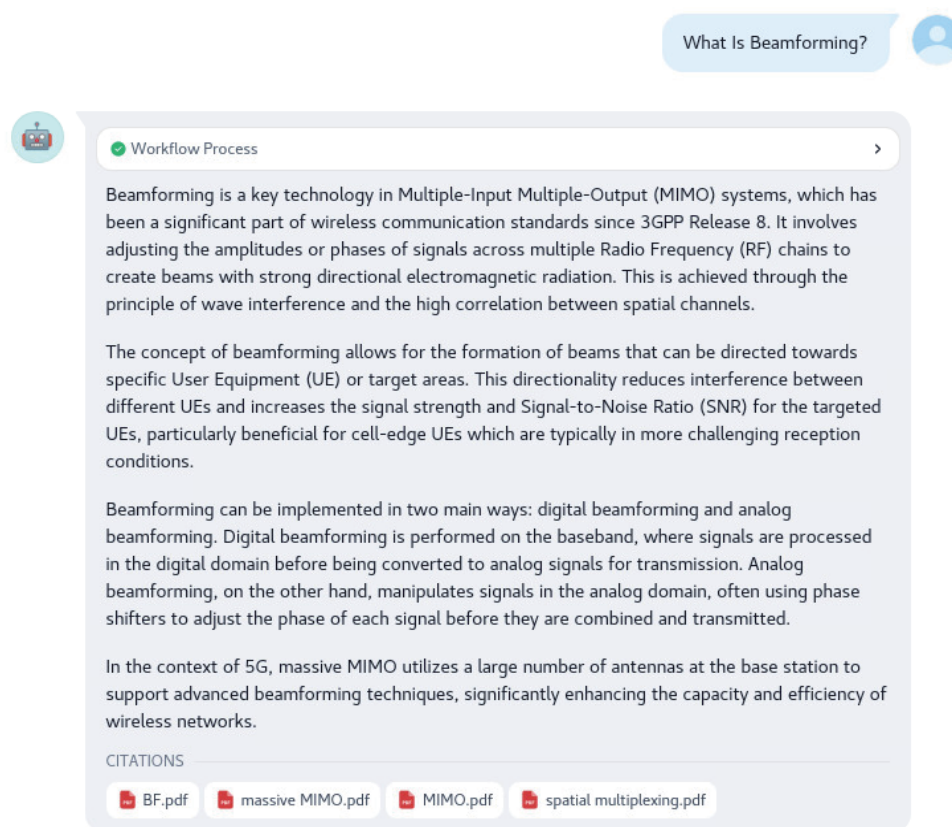


Figure 8 Wireless communications-related question answered by the LLM application

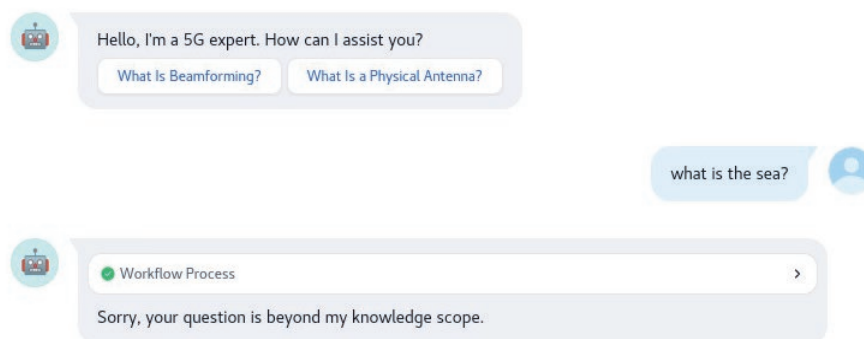


Figure 9 Unrelated question answered by the LLM application

4 Prospects

Although the integration of RAG with LLMs and databases significantly reduces hallucinations in the generated content, RAG still faces many challenges. This section outlines these challenges and RAG's future research interests.

- Multi-modal data processing

A large proportion of enterprise digital assets are PowerPoint and PDF files, which include a massive volume of unconstructed data, such as images and tables. Only a small proportion is text files. However, text is the most stable information source for LLMs. Extracting accurate key information from unstructured data is critical to improving model performance.

- Numerous RAG components and hyperparameters

A RAG application involves many components. For instance, LlamaIndex has more than 80 RAG components that can be selected and integrated based on different scenario requirements. Consequently, developing a RAG application may involve adjusting and optimizing many parameters, including the slicing mode, recall mode, pre-processing, post-processing, and routing parameters. Finding the optimal combination among these components in a huge parameter space is a significant challenge for developers and researchers, leading to a large number of experiments and trial-and-error tests. It is critical to simplify the development process and improve the efficiency of developing RAG applications.

- Integration with enterprises' knowledge bases and search engines

RAG can be integrated with enterprise search engines (such as Elasticsearch) and keyword search engines, eliminating the need for developing knowledge bases from scratch. To meet the requirements of LLMs, such integration requires interface customization. The customized interfaces must be free from restrictions on the context window, be easy to understand, use simplified words and expressions, and have fewer interface parameters. These interfaces can fully unlock the potential of RAG models by further adapting LLMs to search engines, improving the usage of enterprises' knowledge bases and data resources.

- Lack of time attributes

Conflicting datasets in databases might become a major issue for current RAG strategies. For example, after an enterprise policy document is updated, the new rules

may conflict with old ones. Since the current RAG recall strategy does not take time attributes into account, it may recall old knowledge segments during vector recall, causing the LLM to provide incorrect answers. This problem can be solved by introducing time attributes of data into the RAG strategy and increasing the recall score of data that is updated and more relevant based on the attributes. This approach can reduce the possibility of old knowledge segments being recalled after the data is updated. Additionally, paying more attention to the time attributes during model training and application can improve model accuracy and adaptability.

- Fine-tuning the embedding model and rerank model based on data in professional areas

Most open-source embedding models and rerank models are trained based on common corpora. In professional fields, such as wireless communications, these models may fail to recall the most relevant corpus in the RAG process, delivering suboptimal performance. Consequently, these models need to be fine-tuned based on the data in professional fields to improve RAG performance further.

5 Conclusion

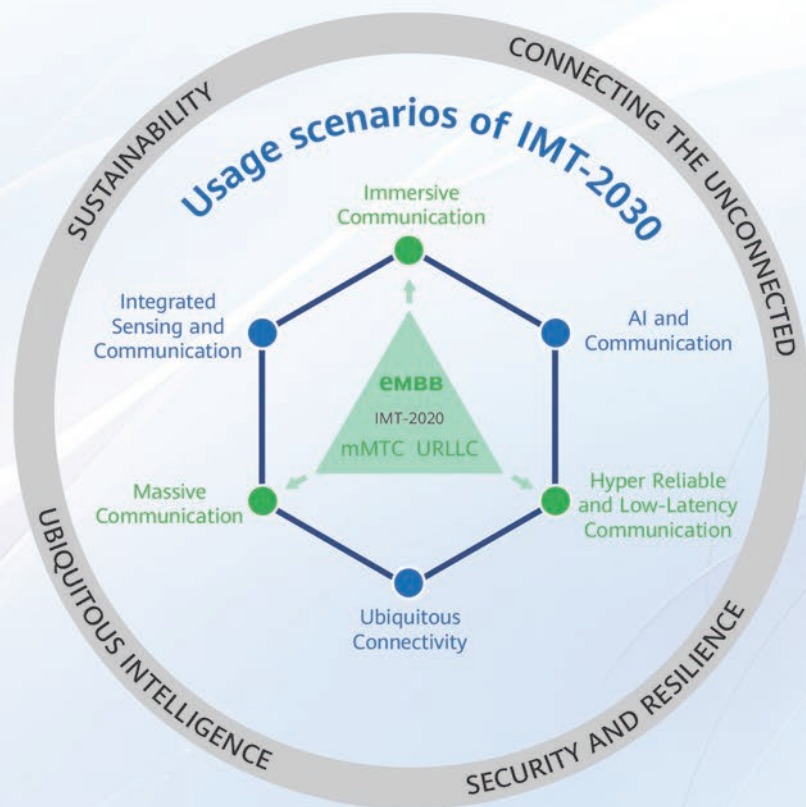
In recent years, the development of LLMs has led to significant innovation in the field of knowledge management. In this paper, we have outlined the challenges and technical schemes involved in knowledge management, including fine-tuning and RAG. We used RAG technology to create a knowledge management solution. We evaluated different LLM models, embedding models, and rerank models to select the optimal combinations for implementing an LLM application that achieves Q&A for wireless communication knowledge bases through many open-source tools. Our model combinations achieved significant reductions in hallucinations through RAG-based integration with databases. However, RAG technology still faces challenges in real-world applications, such as processing multimodal data, selecting complex hyperparameters, and integrating enterprise knowledge bases with search engines.

Our work demonstrates the potential of RAG technology in wireless communication knowledge management and lays a foundation for improving LLM application performance in this field.

References

- [1] Naomi Saphra, Eve Fleisig, *et al.*, "First tragedy, then parse: History repeats itself in the new era of large language models," arXiv.org (2023).
- [2] Ashish Vaswani, Noam M. Shazeer, *et al.*, "Attention is all you need," *Neural Information Processing Systems* (2017).
- [3] Jacob Devlin, Ming-Wei Chang, *et al.*, "BERT: Pre-training of deep bidirectional Transformers for language understanding," *North American Chapter of the Association for Computational Linguistics* (2019).
- [4] Yinhan Liu, Myle Ott, *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv.org (2019).
- [5] Tom B. Brown, Benjamin Mann, *et al.*, "Language models are few-shot learners," *Neural Information Processing Systems* (2020).
- [6] Shukang Yin, Chaoyou Fu, *et al.*, "A survey on multimodal large language models," arXiv.org (2023).
- [7] Zihan Zhang, Meng Fang, *et al.*, "How do large language models capture the ever-changing world knowledge? A review of recent advances," *Conference on Empirical Methods in Natural Language Processing* (2023).
- [8] Boxi Cao, Hongyu Lin, *et al.*, "The life cycle of knowledge in big language models: A survey," Machine Intelligence Research (2023).
- [9] Lizhou Fan, Lingyao Li, *et al.*, "A bibliometric review of large language models research from 2017 to 2023," arXiv.org (2023).
- [10] Michael R Douglas, "Large language models," *Communications of the ACM* (2023). 7-7.
- [11] Shu Wentao, Li Ruixiao, Sun Tianxiang, *et al.*, "Large language models: Principles, implementation, and progress," [J], *Journal of Computer Research and Development*, 2024, 61(02): 351-361.
- [12] Christopher Akiki, Giada Pistilli, *et al.*, "BigScience: A case study in the social construction of a multilingual large language model," arXiv.org (2022).
- [13] Patrick Lewis, Ethan Perez, *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Neural Information Processing Systems* (2020).
- [14] <https://www.rungalileo.io/blog/optimizing-llm-performance-rag-vs-finetune-vs-both>
- [15] <https://chat.lmsys.org/?leaderboard>
- [16] <https://huggingface.co/spaces/mteb/leaderboard>
- [17] https://github.com/FlagOpen/FlagEmbedding/tree/master/FlagEmbedding/llm_reranker
- [18] https://huggingface.co/maidalun1020/bce-reranker-base_v1
- [19] <https://github.com/vllm-project/vllm>
- [20] <https://github.com/xorbitsai/inference>
- [21] <https://github.com/chen700564/RGB?tab=readme-ov-file>
- [22] <https://github.com/chroma-core/chroma>
- [23] <https://github.com/explodinggradients/ragas>
- [24] <https://docs.ragas.io/en/stable/concepts/metrics/index.html>
- [25] <https://github.com/langgenius/dify>

From Connected People and Things to Connected Intelligence



6G-Related Issues of *Communications of HUAWEI RESEARCH*



Issue 1
General



Issue 2
6G



Issue 5
ISAC



Issue 7 (Latest)
AI and Communication







HUAWEI TECHNOLOGIES CO., LTD.

Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129, PRC
Tel: +86-755-28780808

Trademark Notice

 **HUAWEI**, **HUAWEI**, and  are trademarks or registered trademarks of Huawei Technologies Co., Ltd.

All other trademarks and product, service, and company names mentioned in this journal are the property of their respective owners.

General Disclaimer

The information in this journal may contain predictive statement including, without limitation, statements regarding the future financial and operating results, future product portfolios, and new technologies. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

Copyright © 2024 Huawei Technologies Co., Ltd. All Rights Reserved.

No part of this journal may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.