Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Position Article:

AI-agent communication network for 6G: vision, architecture, and key technologies*

Xiaodong DUAN, Zhenglei HUANG, Shiyu LIANG, Shaowen ZHENG, Lu LU, Tao SUN†‡

China Mobile Research Institute, Beijing 100032, China

†E-mail: suntao@chinamobile.com

Received Aug. 19, 2025; Revision accepted Oct. 21, 2025; Crosschecked Nov. 24, 2025

Abstract: The booming of artificial intelligence (AI) agents has brought about promising business scenarios for sixth-generation (6G) mobile networks, while simultaneously posing significant challenges to network functionalities and infrastructure. These AI agents can be deployed on end devices (e.g., intelligent robots and intelligent cars) or as digital entities (e.g., personal AI assistants). As novel service entities with autonomous decision-making and task execution capabilities, AI agents introduce potential risks of uncontrollable actions and privacy disclosures. AI agents also require new 6G capabilities beyond traditional communication, including multimodality information interaction (e.g., AI models and tokens) and support for service requirements (e.g., computing and sensing of data). In this article, we introduce the concept of AI-agent communication network (ACN), a new paradigm to enable global information interaction and on-demand capability provisioning for single or multiple AI agents. We first introduce the vision and architectural framework of ACN. Then, key technologies and future research directions related to ACN are discussed. Furthermore, we provide potential use cases to elaborate on how ACN can expand the service capabilities of 6G networks.

Key words: Artificial intelligence agent; Sixth-generation mobile networks; Network architecture; Multimodality interaction; Multi-agent coordination

https://doi.org/10.1631/FITEE.2500582 **CLC number:** TN929.5

1 Introduction

In recent years, artificial intelligence (AI)-based large models have developed rapidly, evolving from single-modal large language models (LLMs) such as ChatGPT to multimodal vision language models (VLMs) such as GPT-4o, as well as vision–language–action (VLA) large models. AI agents built on large models, which are autonomous systems capable of executing certain tasks by interpreting intents, sensing the environment, planning, making decisions, and using tools, constitute an emerging technology (Wang L et al., 2024). There are various AI agents with different capabilities, including embodied AI agents (such as intelligent service robots and robotic dogs) and digital AI agents (such as virtual intelligent assistants). Multi-agent systems are widely applied in various industries and fields. Currently, these AI agents of diverse types and capabilities have been widely applied across various industries, effectively enhancing work efficiency (Huang et al., 2025; Jiang XY et al., 2025).

The integration of AI with communication is one of the six most promising business scenarios for sixth-generation (6G) mobile networks, as presented in the International Telecommunication Union (ITU) 6G report (ITU, 2023). Numerous studies have been conducted on the application of AI and large models in telecommunication networks (Mahmoud et al., 2024; Shahid et al., 2025). Zhou et al.

‡ Corresponding author

* Project supported by the National Science and Technology Major Project of China on Mobile Information Networks (No. 2024ZD1300400)

ORCID: Tao SUN, https://orcid.org/0009-0003-3491-8813

(2025) reviewed key techniques and opportunities in the context of LLMs for telecommunications, including LLM-enabled generative applications with telecommunication domain knowledge, LLM-based classification applications involving network security and traffic classification, LLM-enabled optimization techniques, and LLM-aided prediction for telecommunications. Chen ZR et al. (2024) proposed an architecture design and system evaluation for large AI models in 6G wireless networks. Similarly, the fundamental principles, diverse applications, key challenges, and future research directions of wireless large-scale AI models were elaborated in detail by Zhu et al. (2025).

Specifically, large-scale AI has shown tremendous potential in network management and optimization (Shahid et al., 2025). Some research works focus on integrating large-scale AI with network management. Xu YF et al. (2023) developed an LLM-aided network configuration benchmark called CloudEval-YAML, providing a realistic and scalable assessment framework for YAML configurations in cloud-native applications. Dzeparoska et al. (2023) proposed an advanced method to automatically translate high-level user intent into executable policies via LLM's learning capabilities, eliminating pre-configuration. Besides, how to use large-scale AI for network optimization has been widely studied. Du et al. (2024) proposed an innovative LLM-enabled mixture of experts (MoE) approach for network optimization, which uses the powerful reasoning capabilities to analyze the objectives and constraints of users, select specialized deep reinforcement learning (DRL) experts, and determine their decision weights.

One notable potential research direction for future communication networks focuses on generative AI (GAI). Bariah et al. (2024) discussed the potential applications of GAI models with critical thinking capabilities in the field of telecommunications networks. Chai et al. (2024) proposed a GAI-driven mobile network digital twin (DT) paradigm, which uses the capability of GAI to implicitly learn the complex distribution of network data and further generate DT data. Xu MR et al. (2024a) outlined a method for deploying AI-generated content (AIGC) applications on mobile edge networks, which enables the real-time provision of personalized AIGC services while safeguarding user privacy. Similarly, a reinforcement learning method with an LLM interaction framework for distributed GAI services has been proposed by Du et al. (2025), which enables generative agents to mimic varied user personalities and provide subjective quality of experience feedback reflecting personal preferences.

Another potential direction for future communication networks lies in embodied AI (Bai et al., 2024; Liu Y et al., 2025), which centers on exploring the effective integration of large-scale model capabilities with expert datasets, simulation platforms, and task suites of embodied robots. Wang XQ et al. (2025) noted that existing models predominantly rely on offline datasets and thus exhibit limitations in adapting to real-time, dynamic, and non-stationary wireless environments; they therefore proposed a wireless embodied large-scale AI paradigm. This paradigm achieves a shift from passive observation to active embodiment and significantly enhances its adaptability to complex wireless systems.

However, large models lack closed-loop action capabilities and cannot directly convert the inference results into network configurations. Therefore, AI agents with action capabilities are further applied in 6G networks (Xu MR et al., 2024b). A multi-agent system named CommLLM with customized communication knowledge and tools has been proposed by Jiang FB et al. (2024), to solve complex tasks such as semantic communication and resource allocation in 6G networks through knowledge retrieval, collaborative planning, and dynamic evaluation mechanisms. Chen ZQ et al. (2024) analyzed the typical mobile AI agent use cases in 6G, consisting of AI agent-based 6G network automation, handheld personalized agents, connected robotics and autonomous systems, and wearable AI agents.

To provide a more intuitive illustration of the current research status, we present Table 1 to summarize all the references mentioned earlier and compare them with this paper. Aiming to address both the new requirements introduced by AI agents and the intrinsic limitations of existing networks, in this paper, we innovatively propose the concept of AI-agent communication network (ACN) and comprehensively expound on ACN from aspects such as its vision, end-to-end architecture, key technologies, experiments, and potential application scenarios.

Table 1 Classification of research on AI in 6G telecommunication networks

Reference	Network architecture	Trusted access	Flexible networking	Intelligent interaction	Intelligence enablement	Experiment
Mahmoud et al. (2024)	×	✓	×	×	×	×
Zhou et al. (2025)	×	×	×	×	×	×
Chen ZR et al. (2024)	×	✓	×	×	×	×
Zhu et al. (2025)	×	✓	✓	✓	✓	×
Shahid et al. (2025)	×	✓	×	×	✓	✓
Xu YF et al. (2023)	×	×	×	×	×	✓
Dzeparoska et al. (2023)	×	×	×	×	×	✓
Du et al. (2024)	×	×	×	×	×	✓
Du et al. (2025)	×	×	×	✓	×	✓
Xu MR et al. (2024a)	×	×	×	×	✓	✓
Bariah et al. (2024)	×	×	×	×	✓	×
Chai et al. (2024)	×	×	×	×	✓	✓
Bai et al. (2024)	×	×	×	✓	✓	×
Wang XQ et al. (2025)	×	×	×	×	✓	✓
Xu MR et al. (2024b)	×	×	×	✓	✓	✓
Jiang FB et al. (2024)	×	×	×	✓	×	✓
Chen ZQ et al. (2024)	✓	×	×	✓	✓	×
Our paper	✓	✓	✓	✓	✓	✓

2 Demands and challenges for 6G networks in empowering AI agents

Existing research works mainly focus on how to use AI-based large models and AI agent technology in mobile networks. Meanwhile, we have observed that AI agents introduce new challenges for future 6G networks from the perspective of an end-to-end system, especially considering how 6G can empower the various AI agents.

The first challenge is that the service objects are shifting from people to heterogeneous AI agents. AI agents introduce potential risks due to the weak interpretability and high unpredictability of their behaviors, including uncontrollable actions, personal privacy leaks, data security issues, and system vulnerabilities (Deng et al., 2025).

Furthermore, existing security and identity mechanisms are predominantly based on subscriber identity module (SIM)/embedded SIM (eSIM) technologies (Silva et al., 2021). However, these mechanisms are ill-suited for virtual AI agents that lack dedicated and trusted storage capabilities. In addition, the static nature of conventional registration mechanisms further impedes their ability to accommodate the dynamic attribute management requirements of AI agents, whose permissions often vary depending on the task. Therefore, it is crucial to design new security and trust mechanisms for AI agents in 6G networks.

The second challenge is that the interaction content and mode of AI agents differ from those of humans. On one hand, the interaction content of AI agents is shifting from video/audio/text to AI models/tokens/feature vectors, which introduces new traffic characteristics and requires new

transmission schemes. Specifically, the communication traffic of AI agents exhibits prominent dynamic complexity, which is manifested in a set of distinctive features, including multimodal data transmission, short-lived connection links, irregular traffic bursts, and low-latency real-time interaction capabilities (Li et al., 2024). It is necessary to solve the problems of real-time perception and modeling of traffic characteristics, as well as that of deterministic low-latency quality of service (QoS) guarantee for bursts. On the other hand, the interaction mode of AI agents is shifting from a fixed and preconfigured mode to a dynamic and flexible one to enable multi-agent coordination for specific tasks, necessitating new communication protocols and networking schemes.

The third challenge concerns the new capability requirements of AI agents beyond connectivity, such as computing and sensing capabilities. Regarding computing capability, some intelligent terminals with limited computing capability exhibit inherent constraints in supporting large-scale models such as VLMs or VLA models (Jiang FB et al., 2024), rendering them reliant on supplementary computational resources provisioned by the network. In contrast, embodied intelligent robots, while boasting robust computing capabilities sufficient to support large models with billions of parameters, incur substantial deployment and operational costs (Adornetto et al., 2025). This economic constraint drives their demand for computational offloading to the network.

Concerning sensing capabilities, the processes of data collection and storage for training large-scale models of AI agents consume significant resources and time. For instance, training the RT-1 VLA model requires extensive data collected from 13 robots over 17 months, encompassing more than 700 distinct tasks (Brohan et al., 2023). Therefore, leveraging the advanced sensing capabilities of 6G networks emerges as a critical strategy to mitigate these challenges.

Considering the above challenges of 6G networks in empowering AI agents, we propose the concept of ACN, which is a new paradigm to enable secure global information interaction and on-demand capability provisioning for single or multiple AI agents. Based on the endogenous capabilities of 6G, including communication, computing, sensing, and security, ACN can provide new types of services beyond connectivity for various virtual and physical AI agents. In this paper, we first introduce the vision of ACN and then discuss its architecture framework, key technologies, and potential use cases.

3 Vision of ACN as a new paradigm of services

By interconnecting, collaborating, and empowering various AI agents, an ACN aims to enable agents to better assist users with their daily life and work tasks, providing a new experience. AI agents of various forms (physical or virtual) with diverse capabilities (high-end, mid-range, and low-end) are supported with secure information transmission and on-demand capability provisioning. ACN, as depicted in Fig. 1, includes the following capabilities:

1. Trusted ubiquitous access. ACN can support ubiquitous access for AI agents based on multiple access technologies, including the 3rd Generation Partnership Project (3GPP) cellular networks and wireless fidelity (Wi-Fi) on a fixed network. The ubiquitous access facilitates seamless connectivity regardless of whether physical or virtual AI agents are stationary, moving within outdoor cellular networks, or roaming in other countries. Digital identities are provided by the network, derived from and associated with the communication identifiers of the corresponding individual or enterprise users (tenants). These digital identities play a crucial role in

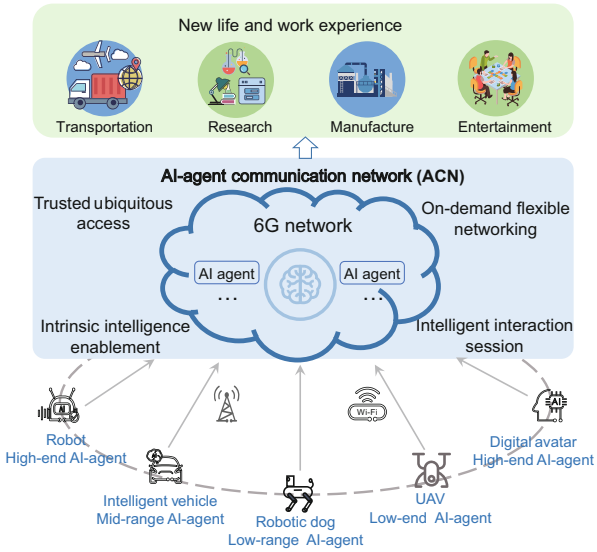

Fig. 1 Vision of ACN. AI: artificial intelligence; ACN: AI-agent communication network; UAV: unmanned aerial vehicle; 6G: sixth-generation

behavior authorization and security authentication during inter-agent communication.

2. On-demand flexible networking. To support enhanced security isolation and completion of complex tasks, ACN must support user-level (or tenant-level) subnets, enabling users or AI agents (under authorization) to create dedicated subnets. Within these subnets, AI agents for these users can communicate in a secure and trusted manner, allowing for interaction and collaboration with each other to share data and resources. User authorization, facilitated through digital identity authentication and attribute association, regulates interactions among AI agents both within and across subnets, thereby achieving on-demand isolation in ACN.

3. Intelligent interaction session. An AI agent can initiate a direct communication session with one or more other AI agents in ACN using a unique and routable digital identity. Such a session can support point-to-point or point-to-multipoint communications. ACN is responsible for establishing the optimal traffic route with ensured QoS across network nodes, as well as providing data buffering and replication services to redistribute the data packets. This approach can eliminate unnecessary transmission delays, ensure end-to-end bandwidth, and mitigate potential security risks associated with the public network. Furthermore, communication among AI agents can be confined within the aforementioned subnets to achieve stricter security isolation or enhanced traffic efficiency.

4. Intrinsic intelligence enablement. ACN provides diverse network resources and capabilities to empower the intelligence and multidimensional capabilities of AI agents. For resource-limited AI agents, such as AI wearable devices and unmanned aerial vehicles (UAVs), ACN can enhance their inference capabilities based on device–network coordination. Additionally, the sensing capability of AI agents can be enhanced by ACN to develop environmental awareness and location-positioning capabilities, thereby expanding their application scenarios without increasing the computing resource costs.

ACN will become a new foundational network service provided by operators to society, following short message services (SMSs), voice and data services of mobile internet, the Internet of Things (IoT), and 5G private networks. By interconnecting various virtual and physical AI agents, ACN will bring operators new types of connections and, consequently, larger traffic demands. More importantly, ACN will provide new means for the monetization of computing, AI, and sensing capabilities of future mobile networks.

4 Design and advantages of ACN architecture

4.1 Architecture design of ACN

As shown in Fig. 2, we propose an end-to-end ACN architecture that connects user equipment (UE), a radio access network (RAN), and a core network (CN). From the UE perspective, future user devices will exhibit a high degree of intelligence, enabling users to directly convey their intentions to the network. Such users can establish connections to the CN via both 3GPP (cellular) and non-3GPP access technologies.

From the perspective of RAN, a new task agent is introduced to decompose wireless-oriented tasks with the assistance of wireless LLMs. This task agent also functions as a model context protocol (MCP) client, enabling it to send requests to MCP servers associated with connection, computing, data and model capabilities, thereby facilitating the invocation of corresponding tools. Using the MCP protocol (Hou et al., 2025), each MCP server is capable of flexibly invoking various tools to provide corresponding services. Notably, compared to generic LLM-based agent tool use, the tool-use mechanism for the network environment is strongly related to the network-specific knowledge, providing improvements in reliability, real-time performance, and security.

Similar to the RAN side, we introduce a task agent on the CN side to support intent recognition, task management, and orchestration. As the interface for UE intent on the CN side, the task agent first uses the network's large model to identify user intent and then decomposes the task into subtasks related to the control plane, security plane, computing plane, data plane, and user plane:

1. The control plane mainly focuses on the allocation of users and network resources, consisting of mobility management, session management, policy management, and networking management tools.

2. The user plane focuses on forwarding user messages and ensuring QoS and includes the routing

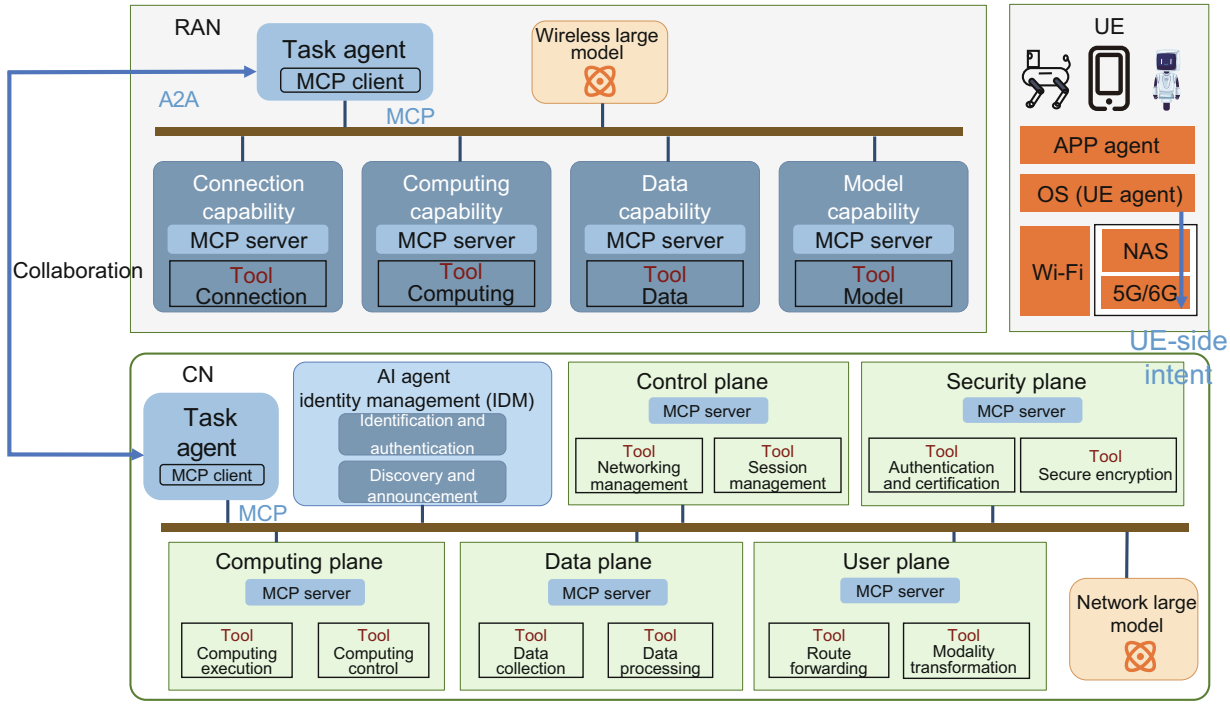

Fig. 2 Architecture of ACN. ACN: AI-agent communication network; A2A: agent-to-agent; MCP: model context protocol; RAN: radio access network; UE: user equipment; CN: core network; IDM: identity management; Wi-Fi: wireless fidelity; APP: application; OS: operating system; NAS: non-access stratum

forwarding, QoS execution, and multimodal transformation tools.

3. The computing plane is responsible for handling computing-related services, including computing control, computing execution, and computing decomposition tools.

4. The data plane is responsible for specific requirements of data and includes the data collection, data processing, and data storage tools.

5. The security plane is responsible for security-related services, including authentication and certification, secure encryption, and related tools.

To flexibly invoke the functions and services of the network, the task agent acts as an MCP client, while each plane is an MCP server, with its network functions encapsulated into diversified tools related to the service. Furthermore, to manage massive AI agents in the future, we introduce AI agent identity management (IDM) on the CN side.

1. IDM: IDM encompasses two core functionalities, identification and authentication, and discovery and announcement. The identification and authentication function validates the digital identity of AI agents and constrains their permission scopes, ensuring secure and trustworthy access and control. Meanwhile, the discovery and announcement function facilitates the retrieval of AI agents that align with task requirements by means of attribute-based queries, such as capability list and service scopes. Through this integrated platform, an underlying connection management can be achieved for both internal and external AI agents within the network, laying the technical foundation for multi-agent interaction and collaboration.

2. Collaboration between two task agents: Notably, two task agents establish the connection between the RAN side and the CN side via collaborative interaction. The agent-to-agent (A2A) protocol (Ray, 2025) is used to support this task collaboration. Concretely, the task agent on the CN side performs overall scheduling and control, and then issues wireless-related tasks to the task agent on the RAN side. Their collaboration is critical to jointly guaranteeing the quality of experience.

Specifically, taking the basic service process as an example, when a UE initiates a service request,

the task agent on the CN side first generates session parameters based on the specific service requirements, and transmits the data-bearer requirements (including QoS constraints) to the task agent on the RAN side. Subsequently, to ensure that resource allocation aligns with service demands, the RAN-side task agent sends a wireless resource allocation request to the MCP server, including QoS parameters provided by the CN. Upon receiving this request, the MCP server invokes the relevant tools to implement the mapping of QoS parameters to the wireless resource policies. After completing resource allocation, the RAN-side task agent returns feedback to the CN-side task agent, typically including information on whether the CN-specified requirements (e.g., latency thresholds and bandwidth limits) are satisfied.

Furthermore, the RAN-side task agent conducts real-time monitoring of its own load status (e.g., wireless resource utilization rate and the number of connected users). When the load exceeds a predefined threshold, it issues a load-balancing request to the CN-side task agent. Upon receiving this request, the CN-side task agent invokes the session management tool via the control-plane MCP server to dynamically adjust QoS parameters, thereby alleviating the load pressure on the access side.

4.2 Advantages of the architecture

Compared with the 3GPP 5G core (5GC), the proposed architecture has two key advantages. One is that it abandons the definition of traditional network functions (e.g., access and mobility management function (AMF), session management function (SMF), network repository function (NRF), unified data management (UDM), policy control function (PCF), and user plane function (UPF)), and instead deploys MCP servers on five planes (i.e., the control plane, user plane, data plane, computing plane, and security plane). Meanwhile, it converts the original atomic services into invokable tools for MCP servers, which are further used to provide services. Besides, task agents are introduced to support user intent recognition and task decomposition, and they act as MCP clients to request various services from the MCP servers deployed on the five planes.

Taking session management as an example, session establishment in 5GC requires collaboration among multiple network functions (e.g., AMF, SMF, PCF, UPF, and UDM). In contrast, in our proposed architecture, the CN-side task agent sends session establishment requests to the MCP servers on the control plane. The MCP servers then complete the session establishment process by invoking the session management tool, which eliminates the cumbersome interaction procedures inherent in 5GC, thereby having significant implications for network simplification.

The other key advantage lies in incorporating new elements absent from 5GC into the proposed network. To facilitate wide-area communication and collaboration among massive heterogeneous and cross-vendor AI agents, an IDM platform is introduced which supports the identification, authentication, authorization, and discovery of AI agents. To address the diverse demands of AI agents, including computing-task offloading, AI model training, and data sharing, computing and data planes are added. Furthermore, an integration mechanism for communication and computing, along with a unified data service framework, is established.

Overall, the proposed architecture enhances resilience in terms of extensibility by using task agents to identify user intentions, decompose tasks, and invoke tools, as new functions can be integrated simply by introducing additional tools. Additionally, the tool invocation mechanism streamlines the implementation of functional modules. Given that the behavior of AI agents is prone to unpredictability and hallucination, we prioritize ensuring the correctness of the decision-making and execution processes of AI agents and implement timely interventions upon detecting abnormal situations, thereby safeguarding the safety and stability of the entire system. There are already some security-related studies that can be used to ensure the security of multi-agent systems (Liu YS et al., 2025a, 2025b). Notably, in terms of standardization, we have incorporated the requirements such as trusted access of agents and flexible networking into the technical report of the 3GPP Service and System Aspects (SA) Working Group 1 and are further striving to include the relevant technical solutions in SA Working Group 2.

5 Key technologies

5.1 Multimodal traffic transmission

The traffic of AI agents has characteristics such as burst data, short connections, multimodality, and

real-time interactions. We propose a method for the real-time perception and transmission of multimodal traffic to achieve service recognition and real-time resource scheduling.

Specifically, as shown in Fig. 3, we develop a transmission assurance framework based on online learning and dynamic policy generation. First, information on the traffic characteristics is exchanged through network-service collaboration. Then, service feature recognition is achieved through AI algorithms such as reinforcement learning. Building on these features, dynamic transmission policies are generated to guide resource scheduling on the RAN side. Through real-time resource allocation based on fine-grained service perception, the exact scheduling requirements of multimodal traffic for AI agents are satisfied. A previous study (Lu and Osorio, 2024) has proposed a link transmission model that significantly improves the computational time performance by optimizing the model formulation. Its efficient computing design, tailored for large-scale networks, can support the coordinated transmission scheduling of multimodal traffic. Additionally, the idea of a dynamic network function provisioning approach (Sun et al., 2021), which enables on-demand deployment and adjustment of network functions, can be effectively adapted to the fluctuating resource needs of multimodal traffic.

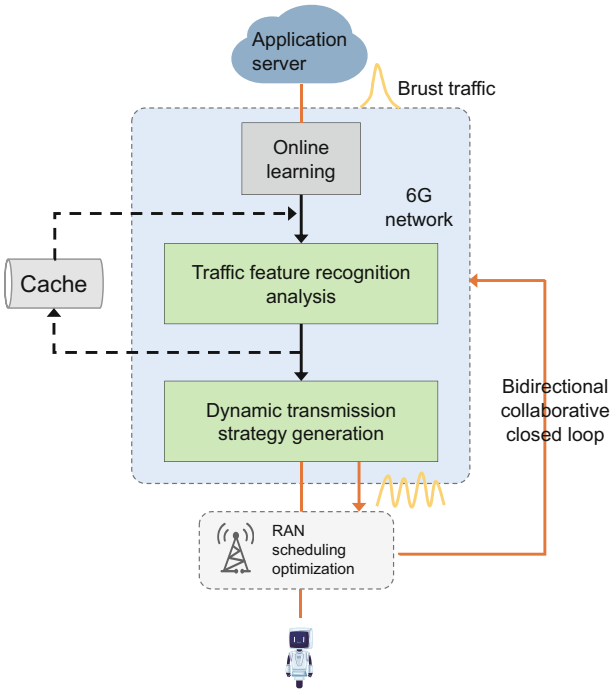

Fig. 3 Framework of multimodal traffic transmission. RAN: radio access network; 6G: sixth-generation

5.2 Design of AI agent identity

AI agents exist in two forms: physical and digital. For physical agents, their security authentication can be in the form of symmetric keys and the identity is bound to the device, similar to the SIM card of mobile phones. In addition, asymmetric keys such as digital identities can be applied. For digital agents, there is no trusted hardware to store symmetric keys; therefore, they can adopt only asymmetric key-based identification (e.g., digital identification) for security authentication. Considering that digital identification can be created and modified remotely, the use of digital identification for various AI agents in 6G networks is suggested. What follows offers a detailed description of the design of AI agent identification.

In the realm of AI agent identification, there are currently various approaches, such as device IDs used in intelligent robots, the generic public subscription identifier (GPSI) defined by 3GPP, and decentralized identifiers (DIDs) defined by World Wide Web Consortium (W3C). Consequently, a unified identity management mechanism is needed to abstract the underlying heterogeneity while reflecting the differentiated characteristics of each AI agent. Inspired by the design principle of the uniform resource identifier (URI) and DIDs, the proposed AI agent identity (ID) syntax is as follows:

Scheme: Method: Method-specific string.

Here, the "Scheme" component indicates the architecture adopted by this ID, which supports the ID framework defined by 3GPP and can be extended to accommodate other schemes (e.g., DIDs) as needed. This component allows for the representation of certain characteristics of the AI agent. The "Method" component specifies the set of schemes (methods) used for defining and operating the ID. Meanwhile, the "Method-specific string" component serves as a unique identification string associated with this method. An example could be "3GPP device: 3GPP: GPSI: 188xxxx8901."

5.3 AI agent registration and discovery

To achieve direct point-to-point communication among AI agents, effective registration and discovery

mechanisms are extremely important. Based on the identification design provided in Section 5.2, we propose a registration and discovery method for AI agents.

5.3.1 AI agent access registration based on the digital ID

As mentioned before, an IDM is introduced into the network to perform identity management. As shown in Fig. 4, the UE can obtain a digital ID for the AI agent from the IDM by providing its own UE ID. After verifying that the UE has subscribed to the AI agent management-related services, the IDM issues a digital identity ID that is associated with the UE ID. The AI agent then uses this digital ID to register with the IDM, submitting relevant attributes such as its ownership, capabilities, and intelligence level. Based on the received information, the IDM proceeds to establish a comprehensive agent profile for the AI agent. After successful registration, the IDM can issue verifiable credentials (VCs) to the AI agent based on the received information or preconfigurations. These VCs include basic information such as authorization claims used for verification and authorization.

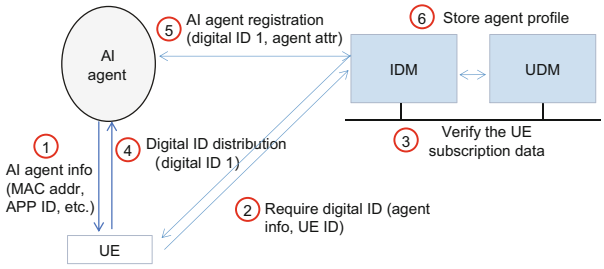

Fig. 4 AI agent access registration based on the digital ID. AI: artificial intelligence; ID: identifier; MAC: media access control; IDM: identity management; UDM: unified data management; UE: user equipment; attr: attribute; addr: address

5.3.2 AI agent discovery

As illustrated in Fig. 5, in the proposed solution, an AI agent sends a discovery request to the IDM for a specific task. If the digital ID of the target AI agent is known, it is included in the request; if not, the request contains filtering information regarding the desired agent. Upon successful verification, the IDM queries its local database to identify agents that fulfill the specified criteria based on the agents' stored profiles. If the query yields results, the IDM provides the AI agent with the address and/or profile of the relevant agents, contingent on the permissions granted. This enables the AI agent to effectively communicate with the target AI agent using the received information.

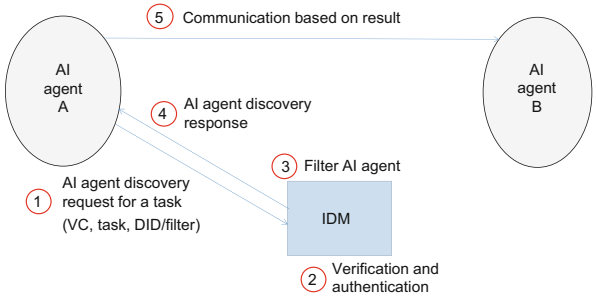

Fig. 5 AI agent discovery. AI: artificial intelligence; IDM: identity management; VC: verifiable credential; DID: decentralized identifier

5.4 Dynamic networking on demand

To achieve dynamic networking, we define the AI agent communication group and its associated information. Specifically, an AI agent communication group refers to a temporary communication group formed by multiple AI agents, which can be user- or task-level groups.

Based on the principle of security isolation and control, individual or enterprise users may create user-level groups. AI agents belonging to the same user can join the group in a secure and trusted manner, to interact, collaborate, and share resources with each other. Through user authorization, it is possible to further achieve intergroup interaction based on digital identity authentication.

Based on the specific objectives of the task, AI agents from different users may be temporarily assembled into a task-level group, so that they can achieve task objectives through information interaction and cooperation. Upon completion of the task, the resources associated with the group can be released, thereby facilitating task-level isolation of AI agent networking.

Both user- and task-level groups are identified by group identifiers, which encompass a list of group members, service scopes (optional), task identifiers of associated services (optional), and other relevant information. Within this framework, users

can create, delete, and update AI agent communication groups via non-access stratum (NAS) signaling, and their information can be routed based on the corresponding group identifiers. The following subsections present two processes to demonstrate how the framework achieves networking and group communication.

5.4.1 Creation of an AI agent communication group

As shown in Fig. 6, the user sends a request for creating a group to the CN through NAS signaling, providing information such as the group's task identifier, service time and scope, and the identity of the AI agent that needs to join the group. Upon receiving the request and associated information, the task agent first sends an authorization request to the security-plane MCP server, confirming whether the user has the right to create the group. After confirming the authority, the task agent forwards the request for creating a group and associated information to the control-plane MCP server. The control-plane MCP server then invokes the networking management tool to assign identifiers to the newly created group and stores relevant information, such as the group member list. Then, the MCP server sends back the group-related information received from the tools to the task agent. Finally, the task agent sends the relevant information of the newly created group to the user.

5.4.2 Establishment of a group communication session

As illustrated in Fig. 7, when an AI agent within a group intends to communicate with another AI agent in the same group, it transmits a communication intent to the CN via NAS signaling, including essential information such as its affiliated group identifier and the digital identity of the target AI agent. After receiving the signal, the task agent recognizes the intent and determines to establish group communication sessions for two AI agents. Therefore, the task agent sends the session establishment request to the control-plane MCP server and the information routing request to the user-plane MCP server. The control-plane MCP server first invokes the session management tool to establish the session for the AI agents, with two AI agents assigned to the same UPF. Then, the user-plane MCP server invokes the route-forwarding tool to route the information based on the group identifier.

6 Experiments and potential use cases of ACN

6.1 Experiments on ACN

Building upon the network architecture and key technologies, this study presents the overall process of ACN, as illustrated in Fig. 8.

To test the latency of the proposed architecture and procedure, we construct a prototype system,

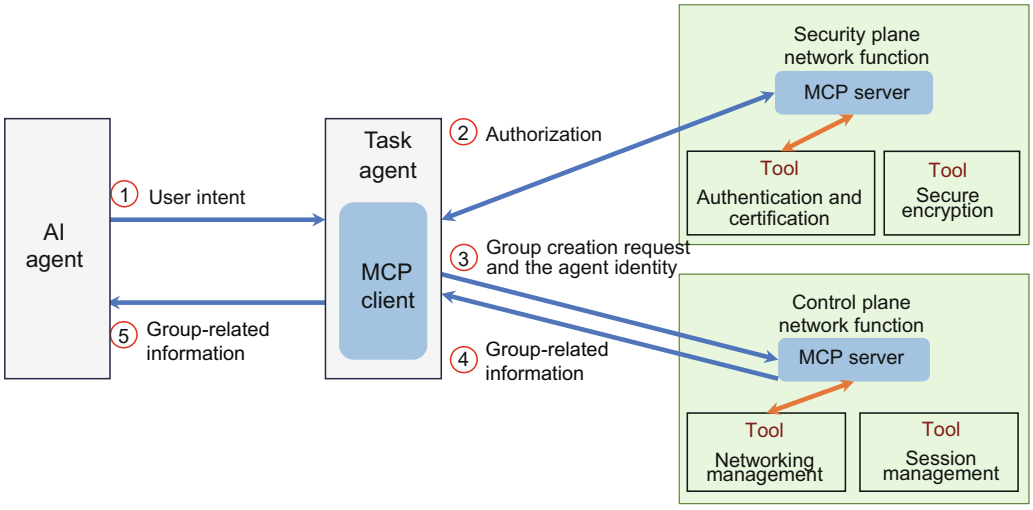

Fig. 6 Procedure of creating a group. AI: artificial intelligence; MCP: model context protocol

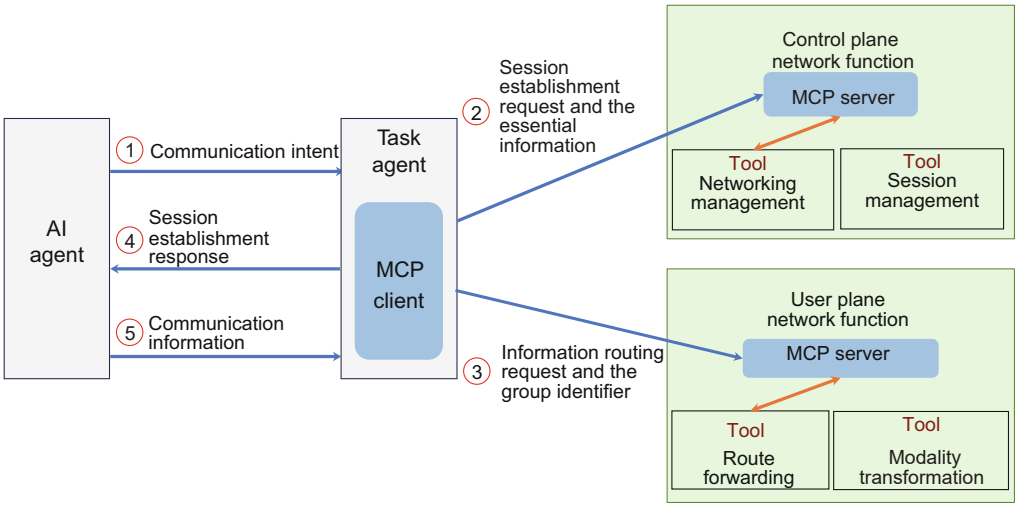

Fig. 7 Procedure for establishing a group communication session. AI: artificial intelligence; MCP: model context protocol

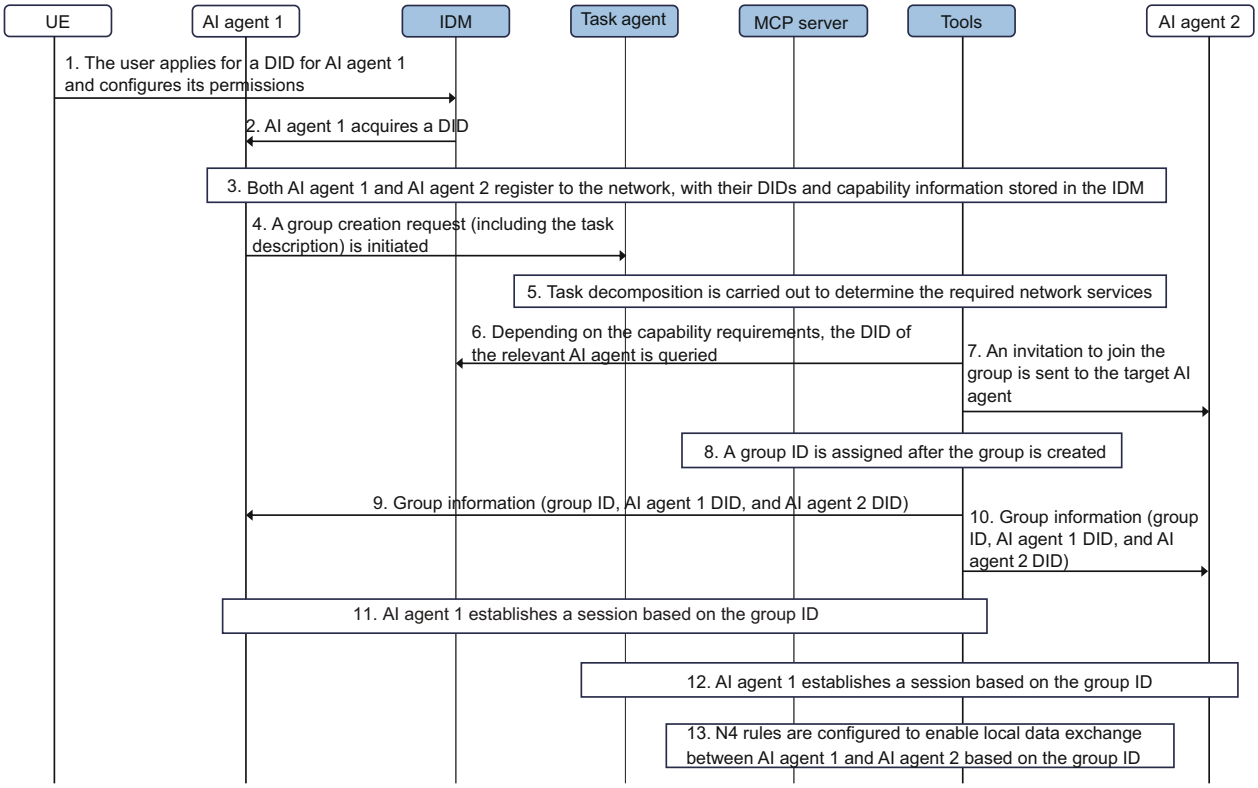

Fig. 8 Overall process of ACN. AI: artificial intelligence; UE: user equipment; IDM: identity management; MCP: model context protocol; DID: decentralized identifier

including robots, routers, and switches. Specifically, robots access the network through an AR611W-S router, which is linked to a CE6860 switch. Downstream of the switch, an identically configured E9000H server is connected which hosts the CN of ACN. Simultaneously, the switch is connected to a large-model server, such as the Atlas 800. The precise configuration and parameters of the devices are listed in Table 2.

Leveraging the aforementioned simulation

Table 2 Configuration and parameters of devices

Device	Configuration
Router (AR611W-S)	Wi-Fi 2.4 GB and 5 GB frequency bands, IPv4 mixed packet forwarding rate of 800 Mbit/s
Switch (CE6860)	25 GE optical ports, packet forwarding rate of 3200 million packets per second
Server (E9000H)	2 Kunpeng 920 processors (64 cores running at 2.6 GHz), 16×32 GB of memory, 1×3.2 TB NVMe SSD storage, and 4×10 GE + 8×10 GE/25 GE network ports
Large-model server (Atlas 800)	2 Kunpeng 920 CPUs and 8 Atlas 300i Pro AI acceleration cards, capable of providing 1120 TOPS of INT8 computing power

Wi-Fi: wireless fidelity; IPv4: Internet Protocol version 4; GE: gigabit Ethernet; SSD: solid-state drive; CPU: central processing unit; TOPS: tera operations per second; INT8: 8-bit integer

environment, we evaluate the latency associated with two procedures for the AI agent: applying for a DID and registering on the network via the acquired DID. These evaluations are performed under two distinct scenarios: with and without the integration of large-model inference executed by the Qwen3 model (Yang et al., 2025).

From Tables 3 and 4, it is evident that considering only the latency of CN transmission and tool invocation meets the stringent low-latency requirements of 6G networks. Although introducing large-model inference can lead to increased latency due to hardware performance bottlenecks, this can be reduced by deploying high-performance graphics processing units (GPUs) or adopting multi-card parallel computing. Meanwhile, future efforts may focus on the joint optimization of network communication latency and inference computation latency.

Table 3 Latency of the robot registration process (without large-model inference)

Procedure	Latency (ms)		
	First test	Second test	Third test
Applying for a DID	157	151	159
Registering with a DID	81	78	82

DID: decentralized identifier

Furthermore, we evaluate the latency from the creation of an AI agent communication group to the point where all group members successfully joined the group. The data presented in Tables 5 and 6 further validate that our architecture and workflow meet stringent low-latency requirements, with no significant latency increase as the number of group members grows. Similarly, the issue of high inference latency can be mitigated via hardware upgrades and joint optimization methods. Existing studies have proposed various optimization methods to address such latency challenges (Wang YD et al., 2023; Yu et al., 2025; Zhang et al., 2025). For example, Yu et al. (2025) proposed a fine-grained expert offloading system for MoE models, which reduces memory usage and achieves a significant reduction in inference latency compared to existing schemes. Wang YD et al. (2023) proposed a multilevel inference system that uses small models to handle the majority of queries and reroutes only high-demand applications to LLMs when necessary. Compared with state-of-the-art schemes, this system achieves a significant reduction in average latency.

Table 4 Latency of the robot registration process (with large-model inference)

Procedure	Latency (s)		
	First test	Second test	Third test
Applying for a DID	9.865	6.207	9.250
Registering with a DID	7.737	8.866	7.492

DID: decentralized identifier

6.2 Potential use cases of ACN

To further elaborate on the potential application scenarios of ACN, this paper explores a typical

Table 5 Latency of creating AI agent communication groups with varying numbers of robots (without large-model inference)

Number of robots	Latency (ms)		
	First test	Second test	Third test
Two robots	445	453	441
Ten robots	479	485	486

AI: artificial intelligence

Table 6 Latency of creating AI agent communication groups with varying numbers of robots (with large-model inference)

Number of robots	Latency (s)		
	First test	Second test	Third test
Two robots	6.245	8.803	7.852
Ten robots	9.248	8.267	7.467

AI: artificial intelligence

application involving two scenarios: (1) communication and collaboration among AI agents within a single group; (2) dynamic networking and communication of AI agents across multiple groups. In this application context, a user may own diverse AI agents (e.g., housekeeping robots, drones, smart cars, and robotic dogs), which can form a user-level AI agent communication group (hereinafter referred to as a user group). The subsequent sections provide a brief introduction to the aforementioned scenarios.

6.2.1 Scenario 1: communication and collaboration among AI agents within a single group

As shown in Fig. 9, taking the communication and collaboration within User group 1 as an example, when approaching a campsite, the housekeeping robot (a high-level AI agent) instructs the drone (a low-level AI agent) to scan a specific area. The instruction includes modal information such as maps, icons (identifying specific areas in the map), and text ("Scan this area"), which can be regarded as a type of high-level instruction.

However, due to the limited battery life, low-level AI agents (such as drones) have limited computing capabilities themselves and are unable to parse high-level instructions composed of natural language, pictures, and so on. Therefore, when the task agent receives high-level instructions and determines that the drone cannot process them, it sends a multimodal information transformation request to the MCP server on the user plane. The MCP server further invokes the modality transformation tool to convert the high-level instructions into low-level instructions that the drone can understand, and then invokes the routing-forwarding tool to forward the information to the drone. Subsequently, the drone flies to the specific area based on the simplified flight control instructions, collects video information, and streams it back to the housekeeping robot. After receiving the video information, the housekeeping robot can then determine the best campsite.

6.2.2 Scenario 2: the dynamic networking and communication of AI agents among multiple groups

As shown in Fig. 10, to detect dangers around the campsite and drive away wild animals at night, a night security joint-defense team is formed, consisting of housekeeping robots, robotic dogs, and drones.

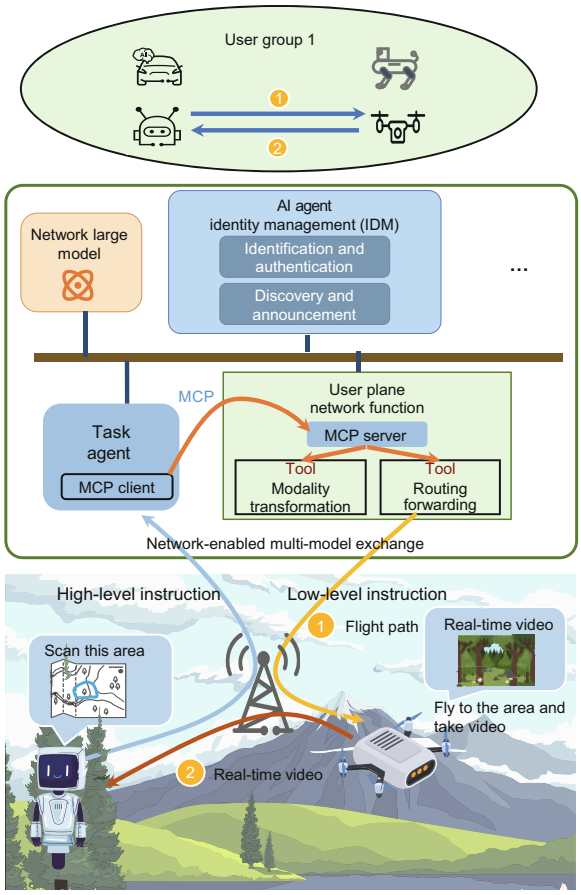

Fig. 9 Communication and collaboration among AI agents within a single group. AI: artificial intelligence; MCP: model context protocol; IDM: identity management

The process begins when the network receives a request for the night security joint defense task from user 1 and user 2. Then, the task agent identifies this request and determines that a temporary task-level group (the validity period of this group is from 19:00 on the same day to 6:00 the next day) needs to be created for the task. To create the group, user 1 first uses the AI agent discovery process described in Section 5.2 to obtain the identities of the AI agents that need to join the group, and then uses the group creation process described in Section 5.4 to complete the group creation.

After creating the task-level group and individual sessions for each AI agent, the task agent further generates the security task plan (including three subtasks):

1. The drone is responsible for collecting video and sending the video to the housekeeping robot.

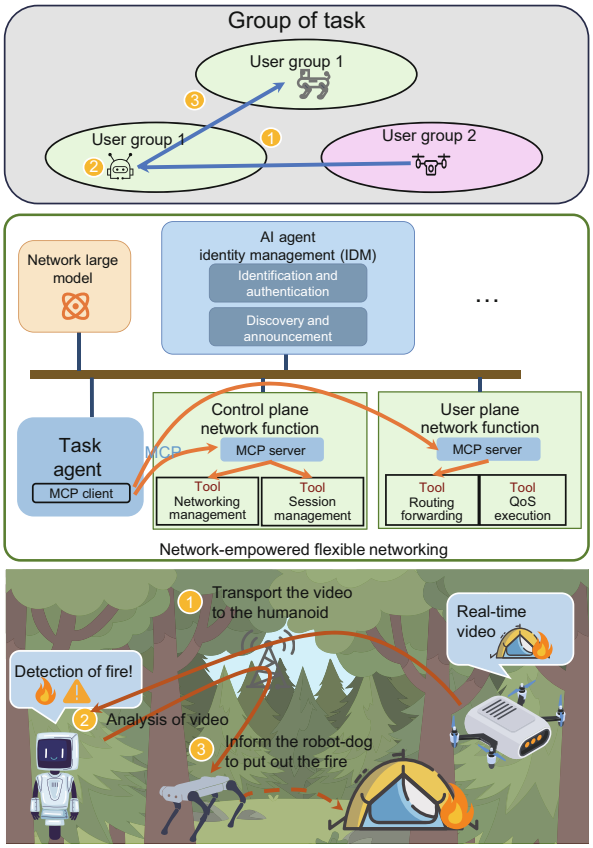

Fig. 10 Dynamic networking and communication of AI agents among multiple groups. AI: artificial intelligence; MCP: model context protocol

2. The housekeeping robot analyzes the video.

3. Once a fire or beast invasion is detected, the housekeeping robot notifies the robotic dog to put out the fire or drive away the wild animals.

To improve the endurance of the housekeeping robot, the robot can request the network to complete the video analysis task. The task agent analyzes the intent and sends the computing request to the MCP server on the computing plane. The MCP server then invokes the computing execution tool to conduct abnormal event analysis. Finally, the AI agent identity management platform can monitor the behavior of the drone. If the drone encounters abnormalities or has insufficient battery life, the AI agent registration management platform can assign a new drone to take over the work.

From the potential use cases, it is evident that ACN has broad application prospects. However, in the actual implementation process, cross-vendor AI agents may use distinct data formats and interaction rules. Under such circumstances, a unified AI agent communication protocol becomes particularly important. In addition, from the standpoint of the industrial ecosystem, the implementation of ACN relies heavily on the in-depth collaboration among operators, AI vendors, and vertical industries.

7 Conclusions

In this article, we introduce the concept of ACN from an end-to-end system perspective, especially focusing on how 6G technology can empower various AI agents. ACN represents a new paradigm designed to enable secure information interaction and on-demand capability provisioning for AI agents. The vision of ACN includes trusted ubiquitous access, on-demand flexible networking, intelligent interaction sessions, and intrinsic intelligence enablement. We also propose an agent-overlay architecture framework and some potential solutions for ACN. Ultimately, ACN is positioned to become a foundational network service in future mobile networks, providing computing, AI, and sensing capabilities to various AI agents.

Our future work will focus on flexible management and network empowerment. For flexible management, ACN will dynamically control the permissions of AI agents based on their operational status (such as indoor/outdoor operation permissions and flexible adjustment of resource usage permissions), thereby ensuring the safety and controllability of AI agents. In terms of network empowerment, ACN will provide integrated sensing and positioning capabilities to AI agents, thereby enabling more accurate environmental detection. In addition, ACN will provide computing resources for AI agents with limited computing capability, allowing them to offload certain computation-intensive tasks to the network.

Contributors

Xiaodong DUAN, Zhenglei HUANG, and Tao SUN designed the research. Shiyu LIANG and Shaowen ZHENG processed the data. Xiaodong DUAN and Zhenglei HUANG drafted the paper. Shiyu LIANG, Shaowen ZHENG, and Lu LU helped organize the paper. Xiaodong DUAN and Zhenglei HUANG revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Adornetto C, Mora A, Hu K, et al., 2025. Generative agents in agent-based modeling: overview, validation, and emerging challenges. *IEEE Trans Artif Intell*, 6(12):3165-3183. https://doi.org/10.1109/TAI.2025.3566362

Bai CJ, Xu HZ, Li XL, 2024. Embodied-AI with large models: research and challenges. *Sci Sin Inform*, 54(9):2035-2082 (in Chinese). https://doi.org/10.1360/SSI-2024-0076

Bariah L, Zhao QY, Zou H, et al., 2024. Large generative AI models for telecom: the next big thing? *IEEE Commun Mag*, 62(11):84-90. https://doi.org/10.1109/MCOM.001.2300364

Brohan A, Brown N, Carbajal J, et al., 2023. RT-1: robotics Transformer for real-world control at scale. https://doi.org/10.48550/arXiv.2212.06817

Chai HY, Wang HD, Li T, et al., 2024. Generative AI-driven digital twin for mobile networks. *IEEE Netw*, 38(5):84-92. https://doi.org/10.1109/MNET.2024.3420702

Chen ZQ, Sun Q, Li N, et al., 2024. Enabling mobile AI agent in 6G era: architecture and key technologies. *IEEE Netw*, 38(5):66-75. https://doi.org/10.1109/MNET.2024.3422309

Chen ZR, Zhang ZY, Yang ZH, 2024. Big AI models for 6G wireless networks: opportunities, challenges, and research directions. *IEEE Wirel Commun*, 31(5):164-172. https://doi.org/10.1109/MWC.015.2300404

Deng ZH, Guo YJ, Han CZ, et al., 2025. AI agents under threat: a survey of key security challenges and future pathways. *ACM Comput Surv*, 57(7):182. https://doi.org/10.1145/3716628

Du HY, Liu GY, Lin YJ, et al., 2024. Mixture of experts for intelligent networks: a large language model-enabled approach. Proc Int Wireless Communications and Mobile Computing Conf, p.531-536. https://doi.org/10.1109/IWCMC61514.2024.10592370

Du HY, Zhang RC, Niyato D, et al., 2025. Reinforcement learning with LLMs interaction for distributed diffusion model services. https://doi.org/10.48550/arXiv.2311.11094

Dzeparoska K, Lin JY, Tizghadam A, et al., 2023. LLM-based policy generation for intent-based management of applications. Proc 19th Int Conf on Network and Service Management, p.1-7. https://doi.org/10.23919/CNSM59352.2023.10327837

Hou XY, Zhao YJ, Wang SN, et al., 2025. Model context protocol (MCP): landscape, security threats, and future research directions. https://doi.org/10.48550/arXiv.2503.23278

Huang SJ, Sun CN, Wang RQ, et al., 2025. Toward adaptive and coordinated transportation systems: a multi-personality multi-agent meta-reinforcement learning framework. *IEEE Trans Intell Transp Syst*, 26(8):12148-12161. https://doi.org/10.1109/TITS.2025.3560227

International Telecommunication Union (ITU), 2023. M. 2160: Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond. ITU, Geneva.

Jiang FB, Peng YB, Dong L, et al., 2024. Large language model enhanced multi-agent systems for 6G communications. *IEEE Wirel Commun*, 31(6):48-55. https://doi.org/10.1109/MWC.016.2300600

Jiang XY, Zheng C, Zhuo Y, et al., 2025. Advancing industrial data augmentation in AIGC era: from foundations to frontier applications. *IEEE Trans Instrum Meas*, 74:1-22. https://doi.org/10.1109/TIM.2025.3572162

Li C, Dong SK, Yang SD, et al., 2024. Multi-agent sparse interaction modeling is an anomaly detection problem. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.5890-5894. https://doi.org/10.1109/ICASSP48485.2024.10446644

Liu Y, Chen WX, Bai YJ, et al., 2025. Aligning cyber space with physical world: a comprehensive survey on embodied AI. *IEEE/ASME Trans Mechatron*, early access. https://doi.org/10.1109/TMECH.2025.3574943

Liu YS, Dong XW, Zio E, et al., 2025a. Event-triggered multiple leaders formation tracking for networked swarm system with resilience to noncooperative nodes. *IEEE Trans Cybern*, 55(9):4136-4144. https://doi.org/10.1109/TCYB.2025.3580666

Liu YS, Li WX, Dong XW, et al., 2025b. Resilient formation tracking for networked swarm systems under malicious data deception attacks. *Int J Robust Nonl Contr*, 35(6):2043-2052. https://doi.org/10.1002/rnc.7777

Lu J, Osorio C, 2024. Link transmission model: a formulation with enhanced compute time for large-scale network optimization. *Transp Res Part B Methodol*, 185:102971. https://doi.org/10.1016/j.trb.2024.102971

Mahmoud H, Elbadawy HM, Ismail T, et al., 2024. A comprehensive review of generative AI applications in 6G. Proc 6th Novel Intelligent and Leading Emerging Sciences Conf, p.593-596. https://doi.org/10.1109/NILES63360.2024.10753177

Ray PP, 2025. A review on agent-to-agent protocol: concept, state-of-the-art, challenges and future directions. https://doi.org/10.36227/techrxiv.174612014.42157096/v1

Shahid A, Kliks A, Al-Tahmeesschi A, et al., 2025. Large-scale AI in telecom: charting the roadmap for innovation, scalability, and enhanced digital experiences. https://doi.org/10.48550/arXiv.2503.04184

Silva C, Barraca JP, Aguiar R, 2021. eSIM suitability for 5G and B5G enabled IoT verticals. Proc 8th Int Conf on Future Internet of Things and Cloud, p.210-216. https://doi.org/10.1109/FiCloud49777.2021.00038

Sun G, Xu Z, Yu HF, et al., 2021. Dynamic network function provisioning to enable network in box for industrial applications. *IEEE Trans Ind Inform*, 17(10):7155-7164. https://doi.org/10.1109/TII.2020.3042872

Wang L, Ma C, Feng XY, et al., 2024. A survey on large language model based autonomous agents. *Front Comput Sci*, 18(6):186345. https://doi.org/10.1007/s11704-024-40231-1

Wang XQ, Zhu FH, Yang ZH, et al., 2025. Bridging physical and digital worlds: embodied large AI for future wireless systems. https://doi.org/10.48550/arXiv.2506.24009

Wang YD, Chen K, Tan HS, et al., 2023. Tabi: an efficient multi-level inference system for large language models. Proc 18th European Conf on Computer Systems, p.233-248. https://doi.org/10.1145/3552326.3587438

Xu MR, Du HY, Niyato D, et al., 2024a. Unleashing the power of edge-cloud generative AI in mobile networks: a survey of AIGC services. *IEEE Commun Surv Tut*, 26(2):1127-1170. https://doi.org/10.1109/COMST.2024.3353265

Xu MR, Niyato D, Kang JW, et al., 2024b. When large language model agents meet 6G networks: perception, grounding, and alignment. *IEEE Wirel Commun*, 31(6):63-71. https://doi.org/10.1109/MWC.005.2400019

Xu YF, Chen YN, Zhang XM, et al., 2023. CloudeVal-YAML: a practical benchmark for cloud configuration generation. https://doi.org/10.48550/arXiv.2401.06786

Yang A, Li AF, Yang BS, et al., 2025. Qwen3 technical report. https://doi.org/10.48550/arXiv.2505.09388

Yu HF, Cui XQ, Zhang H, et al., 2025. fMoE: fine-grained expert offloading for large mixture-of-experts serving. https://arxiv.org/html/2502.05370v1

Zhang MJ, Shen XM, Cao JN, et al., 2025. EdgeShard: efficient LLM inference via collaborative edge computing. *IEEE Int Things J*, 12(10):13119-13131. https://doi.org/10.1109/JIOT.2024.3524255

Zhou H, Hu CM, Yuan Y, et al., 2025. Large language model (LLM) for telecommunications: a comprehensive survey on principles, key techniques, and opportunities. *IEEE Commun Surv Tut*, 27(3):1955-2005. https://doi.org/10.1109/COMST.2024.3465447

Zhu FH, Wang XQ, Li XY, et al., 2025. Wireless large AI model: shaping the AI-native future of 6G and beyond. https://doi.org/10.48550/arXiv.2504.14653