SCIENCE CHINA
Information Sciences

• LETTER •

September 2026, Vol. 69, Iss. 9, 190308:1–190308:2
https://doi.org/10.1007/s11432-025-5013-5

Special Topic: User-Centric Intelligent Networking

Service agent for user-centric QoE-aware communication service provision in mobile AR

Conghao ZHOU[1], Lulu SUN[1], Peng YANG[2*], Feng LYU[3], Nan CHENG[1*] & Xuemin (Sherman) SHEN[4]

[1]*School of Telecommunications Engineering, Xidian University, Xi'an 710071, China*
[2]*School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China*
[3]*School of Computer Science and Engineering, Central South University, Changsha 410083, China*
[4]*Department of Electrical and Computer Engineering, University of Waterloo, Waterloo N2L 3G1, Canada*

Received 10 October 2025/Revised 17 March 2026/Accepted 21 June 2026/Published online 14 August 2026

Citation Zhou C H, Sun L L, Yang P, et al. Service agent for user-centric QoE-aware communication service provision in mobile AR. Sci China Inf Sci, 2026, 69(9): 190308, https://doi.org/10.1007/s11432-025-5013-5

Mobile augmented reality (MAR) is a prominent use case of immersive communication in the 6G era [1]. MAR applications must guarantee that virtual contents are precisely aligned with each MAR user's viewpoint according to the real-time estimation of the MAR device's pose. However, existing communication networks encounter two major issues in maintaining satisfactory user experiences for MAR through network-level management. First, most information intrinsic to MAR applications is confined within the over-the-top (OTT) application service domain and is typically beyond the reach of the network controller. This cross-domain isolation constitutes a fundamental bottleneck for formulating QoE-oriented cross-layer scheduling strategies. Second, traditional service-oriented resource orchestration approaches, e.g., network slicing, are insufficient to address fine-grained and user-specific QoS demands. Recent research efforts have shifted toward user quality of experience (QoE) modeling and QoE-aware network management. Despite these advances, most existing studies overlook the influence of MAR-specific operational mechanisms on network resource orchestration [2, 3].

In this study, we propose a large language model (LLM)-driven MAR service agent (MAR-SA), which can operate on behalf of the MAR application service provider to enable user-centric communication service provisioning for MAR within the network domain. The main contributions of this study are as follows.

• We design an agent that employs the LLM as its core engine to enable cross-layer network resource management by bridging the data and functional isolation between the OTT MAR application service and network domains.

• We propose a user-personalized QoE-aware resource management framework that leverages the MAR-SA to model the user-specific relationship between communication resource demand and user QoE, thereby facilitating fine-grained on-demand resource allocation.

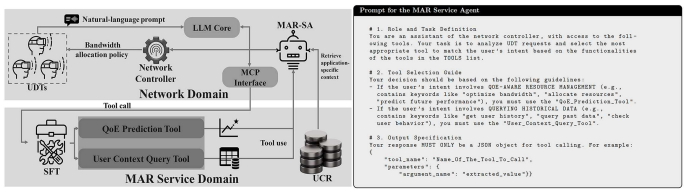
MAR-SA architecture. The proposed MAR-SA framework aims to securely execute cross-layer synergy between the network layer and the MAR service domain. As shown in Figure 1(a), the MAR-SA consists of three principal components: the service function toolkit, user context repository, and LLM core, whose functionalities are introduced as follows.

• **Service function toolkit (SFT)**: Application-side functionalities related to networking within the MAR service domain can be abstracted into modular and callable tools. Each tool is defined and encapsulated in accordance with the model context protocol, which specifies its functional description and standardized parameter schema. Taking MAR annotation rendering as an example, we define two specific tools to enable QoE-aware service provisioning for rendering: (i) the QoE prediction tool, which encapsulates the deterministic computation of user QoE based on the algorithm adopted by the MAR application; and (ii) the user context query tool, which supports the retrieval and processing of historical user data such as user QoE requirements.

• **User context repository (UCR)**: The UCR securely bridges the MAR service domain and the network domain from a data perspective. Serving as a user-centric data management interface, it ensures that the network controller can only obtain the final inference results. Meanwhile, the underlying private user data remains strictly isolated within the MAR service domain, accessed internally and compliantly only by the LLM core when invoking SFT tools. For application-side data, the UCR relies on the LLM core to invoke the corresponding tools in the SFT for access. For network-side data, the network controller can directly retrieve the information through the UCR.

• **LLM core:** The LLM core serves as the reasoning and coordination module of the MAR-SA framework. Instead of directly processing user data within the MAR service domain, e.g., user QoE prediction, the LLM core leverages its intent recognition and tool invocation capabilities to interpret network management intents, expressed in the input prompt, and orchestrates the interaction among the SFT and UCR [4, 5]. As a result, the LLM core enables the integration of application-side and network-side information on behalf of the network controller, without collecting raw user data or analyzing proprietary application mechanisms. This

* Corresponding author (email: yangpeng@hust.edu.cn, dr.nan.cheng@ieee.org)



(a) (b)

Figure 1 (a) Workflow of the proposed agent-driven service provisioning approach; (b) the exemplary structure of the prompt.

design ensures cross-task generalizability, protects user privacy, and prevents latency bottlenecks by separating semantic reasoning from execution.

The MAR-SA can be viewed as a third-party component positioned between the MAR application service provider and the network controller. It can be deployed within the network domain to assist the network controller in making decisions on network resource management. It can also be integrated with our predefined user digital twin (UDT) to profile each MAR user device. In the example of MAR annotation rendering, each UDT characterizes user-personalized service demands, such as QoE requirements, and records user-specific information to provide the LLM core with network-side data, assisting in prompt formulation and reasoning. In addition, each UDT can send the LLM core requests to satisfy its QoE requirement.

MAR-SA workflow. We design a structured workflow for the proposed MAR-SA, ensuring it supports decision-making on multi-user QoE-aware communication service provisioning. To ensure accurate intent recognition and tool selection by the LLM core, we design a structured prompt tailored for the MAR-SA. As shown in Figure 1(b), the primary purpose of this prompt is to establish a well-defined role, task scope, and output format for the LLM core, thereby constraining its general language capabilities within the context of our task framework. Specifically, the template defines the LLM's role as a network management assistant, limits its functionality to intent recognition and tool selection, and requires that the output is formatted as a machine-readable JSON object.

Furthermore, we adopt the model context protocol by abstracting the workflow within the SFT into a standardized representation consisting of a name, functional description, and parameter schema. Prior to processing UDT requests, the MAR-SA serializes the complete list into a JSON-formatted string and dynamically embeds it into the structured prompts. To mitigate the critical bottleneck introduced by LLM inference latency relative to user movement coherence time, the MAR-SA adopts an event-driven decoupled reasoning-execution mechanism. The detailed structures of the prompts and the encapsulated tools are provided in Appendix A.

Semi-distributed service provisioning. Based on the precise QoE inference provided by the MAR-SA, we propose a semi-distributed bandwidth allocation algorithm based on marginal utility. The network controller does not need to acquire the global closed-form QoE models in advance. Instead, in each decision cycle, the controller aggregates the marginal utility returned in real-time by each UDT via the MAR-SA and dynamically allocates bandwidth. Regarding the perceived QoE, bandwidth, and QoE deficiency for all users, the network utility is defined as

$$\Pi(\mathbf{b}) = \sum_{u\in\mathcal{U}} h_u - w \sum_{u\in\mathcal{U}} b_u - v \sum_{u\in\mathcal{U}} \delta_u. \tag{1}$$

The system model, problem formulation, and the proposed algorithm are detailed in Appendix B.

Simulation results. We conduct trace-driven simulations using a real-world 6DoF pose dataset and a virtual scene content dataset to evaluate the performance of MAR-SA in user-centric service provisioning for MAR annotation rendering. By leveraging tool invocation and application-side data access, MAR-SA significantly outperforms conventional AI-based QoE prediction approaches, such as FreTS, PatchTST, and iTransformer, in terms of QoE prediction accuracy. Moreover, the proposed MAR-SA framework avoids the error accumulation inherent in data-driven time-series baselines, thereby effectively improving the immersive experience for multiple users. The results demonstrate the effectiveness of MAR-SA in supporting user-specific network management decision-making without the need to physically collect and process massive volumes of user data. Comparative experiments, detailed simulation results, including scalability and complexity analysis, and latency performance, are presented in Appendix C.

Conclusion and future work. In this study, we have introduced an agent-based framework for user-centric service provisioning in MAR. Utilizing the capabilities of LLMs, the MAR-SA has been developed to represent the MAR service provider and enable cross-domain coordination between the network and MAR application service, eliminating the need for the network controller to access application-specific user data physically. In addition, by abstracting application-specific functions into callable tools, MAR-SA accurately derives personalized resource demands without exposing raw user data. Based on the interaction between the network controller and the MAR-SA, we evaluate the effectiveness of the MAR-SA in QoE-aware service provisioning for MAR annotation rendering by dynamically adjusting bandwidth allocation for individual MAR devices to adapt to their QoE.

The framework is adaptable to various 6G applications beyond MAR via the customizable SFT and can efficiently meet heterogeneous user demands through personalized agent-driven service provisioning. In the future, we will incorporate broader resource dimensions such as computation, whose multidimensional and interdependent effects on personalized QoE modeling in MAR call for more sophisticated designs of the SFT and UCR.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62501449) and Excellent Young Scientists Fund (Overseas).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhou C, Gao J, Hu S, et al. User-centric communication service provision for edge-assisted mobile augmented reality. IEEE Trans Mobile Comput, 2026, 25: 3435–3449
2 Pan G, Xu S, Zhang S, et al. Quality of experience oriented cross-layer optimization for real-time XR video transmission. IEEE Trans Circuits Syst Video Technol, 2024, 34: 7742–7755
3 Feng J, Liu L, Hou X, et al. QoE fairness resource allocation in digital twin-enabled wireless virtual reality systems. IEEE J Sel Areas Commun, 2023, 41: 3355–3368
4 Xi Z H, Chen W X, Guo X, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci, 2025, 68: 121101.
5 Zhou C, Hu S, Gao J, et al. User-centric immersive communications in 6G: a data-oriented framework via digital twin. IEEE Wireless Commun, 2025, 32: 122–129