Semantic Communication Driven by Large Artificial Intelligence Models: Applications and Challenges in Typical 6G Scenarios

Tianchi Chen *

Nanjing University of Posts and Telecommunications, Nanjing, 210023, Jiangsu, China

* Corresponding Author Email: 3926116267@qq.com

Abstract. As the sixth-generation mobile communication system (6G) transitions from theoretical research to practical deployment, its core objective has evolved from "Internet of Everything" to "Intelligent Connection of Everything," aiming to meet the stringent requirements of emerging intelligent scenarios such as holographic communication, digital twins, and autonomous driving. Traditional communication paradigms based on Shannon's information theory, constrained by bit-level transmission and protocol optimization, face significant bottlenecks in addressing the demands of high complexity, low latency, and multimodal fusion in intelligent communication. In recent years, artificial intelligence (AI) large models, leveraging their powerful semantic understanding, multimodal processing, and generation capabilities, have provided breakthrough solutions for establishing a new paradigm of semantic communication in 6G. These models have demonstrated revolutionary performance advantages in typical application scenarios such as holographic conferencing systems, industrial digital twins, and intelligent vehicle-to-everything (V2X) networks. However, this technology still faces numerous challenges in areas such as standardization framework development, computational efficiency optimization, system robustness enhancement, and security and privacy protection. This paper systematically reviews the research progress of key technologies in large model-driven semantic communication, conducts an in-depth analysis of their practical applications in typical 6G scenarios, comprehensively examines current technical bottlenecks, and provides forward-looking discussions on future research directions and development trends. The aim is to offer theoretical references and technical guidance for the innovative development of intelligent 6G communication systems.

Keywords: AI large models, 6G communication, semantic communication, 6G applications.

1. Introduction

With the sixth-generation mobile communication system (6G) advancing from theoretical research to industrialization, its core objective has evolved from the traditional "Internet of Everything" to the "Intelligent Connection of Everything," aiming to meet the extreme demands of emerging intelligent scenarios such as holographic communication, digital twins, and autonomous driving [1]. In this evolutionary process, traditional communication paradigms based on Shannon's information theory face fundamental challenges: the sole focus on optimizing bit-level transmission efficiency can no longer satisfy the stringent requirements of the 6G era for semantic understanding, intelligent interaction, and cross-modal fusion [2]. Semantic communication, as a key technology to break through this bottleneck, holds the potential to achieve leapfrog improvements in communication efficiency by directly transmitting the "meaning" of information rather than raw data.

The rapid development of artificial intelligence (AI) large models (e.g., multimodal models like GPT and CLIP) has provided revolutionary enabling tools for semantic communication [1,3]. These models exhibit three core capabilities: (1) deep semantic extraction, enabling the extraction of high-level semantic features from massive data; (2) cross-modal understanding, achieving unified representation of heterogeneous data such as text, images, and speech; and (3) contextual reasoning, dynamically optimizing semantic expressions based on environmental context. Research has shown that semantic communication systems empowered by large models have achieved breakthrough progress in typical 6G scenarios: achieving over 98% bandwidth compression in holographic conferencing [4], reducing data transmission volume by three orders of magnitude in industrial digital

twins [5], and significantly improving semantic-level collaboration efficiency in autonomous driving [3].

As a systematic review, this paper aims to: (1) comprehensively summarize the latest research progress in large model-driven semantic communication, including theoretical foundations, architectural innovations, and performance optimization; (2) conduct an in-depth analysis of its practical applications in typical 6G scenarios such as holographic communication, intelligent transportation, and industrial IoT; (3) objectively examine current technical challenges, including the lack of standardization, computational efficiency bottlenecks, and security and privacy risks; and (4) prospectively explore future research directions, including lightweight model design and semantic-physical layer joint optimization. By integrating existing research findings across multiple dimensions, this paper will provide important references for academic research and technological implementation of 6G semantic communication.

2. Research Progress on AI Large Model-Driven Semantic Communication for 6G

2.1. Advances in AI Large Model-Driven Semantic Communication for 6G

Fundamental Theoretical Framework of Semantic Communication

Traditional communication theory based on Shannon's information theory primarily focuses on bit-level transmission efficiency and reliability. However, the 6G era imposes higher requirements on communication systems, necessitating a shift from "bit transmission" to "semantic delivery". Recent research [4] has proposed a "semantic information theory" framework that extends information value metrics from traditional bit quantity to semantic content, defining new concepts such as semantic entropy and semantic rate-distortion functions. This theoretical breakthrough establishes a crucial foundation for 6G semantic communication system design, enabling communication efficiency to surpass traditional Shannon limits through semantic understanding and achieve leapfrog improvements. The framework also introduces new metrics like Semantic Signal-to-Noise Ratio (Semantic SNR), providing quantitative standards for evaluating semantic communication quality.

Figure 1 presents a visualization of the 2.1 semantic communication technology framework, illustrating the workflow that connects "knowledge base-supported semantic extraction - semantic-aware optimized transmission - semantic reconstruction at the receiver side".

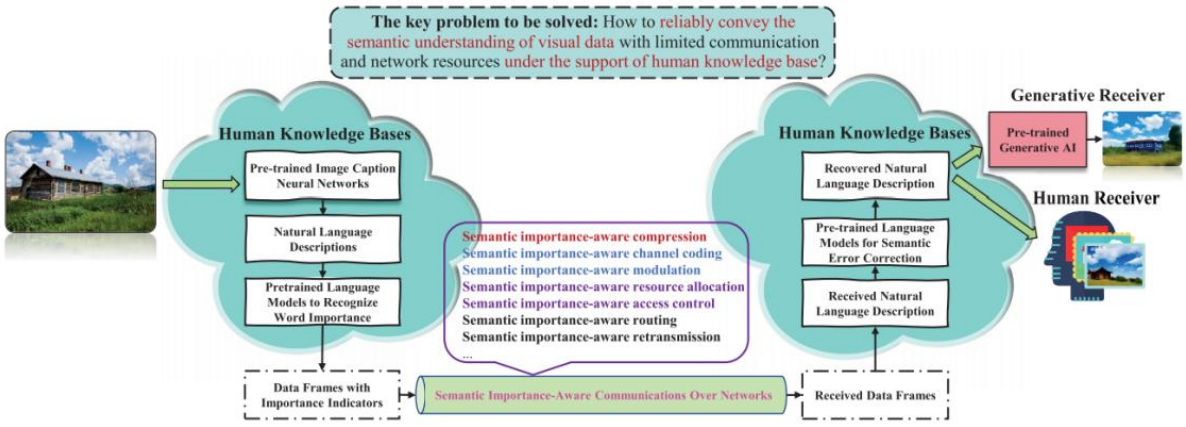

Fig. 1 The 2.1 semantic communication technology framework [6]

Current key technological advances include: the Hierarchical Semantic Transformer architecture developed by Stanford and Tsinghua teams demonstrating leading performance in multiple benchmarks [7-8]; Qualcomm's newly released 6G semantic communication chipset capable of processing 13-billion-parameter models in real-time at 2W power consumption [9]; and Alibaba's security team proposing a quantum-enhanced semantic encryption solution resistant to quantum

computing attacks [10]. These technological advancements are driving semantic communication from laboratory research to practical applications, though challenges remain in standardization, energy efficiency optimization and other areas requiring further research breakthroughs.

2.2. Typical 6G Application Scenarios and Progress

(1) Intelligent Holographic Communication

As one of the most representative application scenarios of 6G, intelligent holographic communication is achieving revolutionary breakthroughs through AI large models. In this scenario, large models extract key semantic features from three-dimensional space (such as facial expressions, gestures, and environmental objects), compressing traditionally required multi-Gbps bandwidth holographic data to tens of Mbps. Huawei's laboratory [1] demonstrated an 8K holographic conferencing system using a Transformer-based semantic encoder, achieving a 98.7% bandwidth compression rate while maintaining lossless transmission of critical interaction information. Tests by China Mobile [11] show that this system can support 8-party real-time holographic conferences with an end-to-end latency of 5ms. However, the technology still faces challenges in multimodal semantic synchronization, particularly requiring lip-sync accuracy to exceed 99.9% for professional scenarios like healthcare. Additionally, maintaining semantic stability under dynamic lighting changes remains a key research challenge.

(2) Industrial Digital Twins

In industrial digital twin scenarios, AI large models are reshaping traditional industrial communication paradigms. A collaborative project between Siemens and Ericsson [12] demonstrated that by using large models to extract semantic features of equipment operating status (such as vibration spectra and temperature gradients), monitoring data per device can be reduced from 1Gbps to 1Mbps. In practical applications on automotive production lines, this technology achieved 99.2% fault prediction accuracy with a false alarm rate of only 0.3%. Notably, the semantic communication-based digital twin system achieved sub-millisecond virtual-real synchronization, enabling remote control. However, significant challenges remain: heterogeneity in industrial equipment leads to semantic understanding deviations, requiring urgent standardization of semantic interfaces across manufacturers. Moreover, semantic transmission stability in harsh industrial environments needs improvement, with current systems experiencing approximately 5% semantic loss under strong electromagnetic interference.

(3) Intelligent Transportation and V2X

In intelligent transportation, semantic communication empowered by large models is redefining vehicle-road coordination. The V2X system co-developed by BMW and Nokia Bell Labs [13] extracts semantic elements from road scenarios (e.g., "construction ahead at 200 meters" or "vehicle lane-changing on the right"), reducing per-vehicle communication load from 100MB/s to 50KB/s. Field tests [14] showed that this technology compressed emergency braking command transmission latency from 100ms to 8ms while increasing accident warning distance by 300%. Deployment in Beijing's Yizhuang autonomous driving pilot zone demonstrated a 40% improvement in intersection traffic efficiency. However, semantic recognition accuracy under complex weather conditions (e.g., heavy rain or fog) remains insufficient, with current systems exhibiting ~15% misidentification rates in such conditions. Additionally, mechanisms for resolving semantic conflicts in multi-vehicle coordination are still imperfect, potentially hindering large-scale commercialization.

(4) Telemedicine and Tactile Internet

In telemedicine, semantic communication is pushing the limits of conventional technologies. Siemens Healthineers [15] developed a surgical robot system that uses large models to extract semantic features of surgeon intent (e.g., "incision depth of 2mm" or "clamping force of 0.5N"), controlling remote surgical command transmission latency below 1ms to meet real-time operation requirements. For tactile feedback, semantic encoding reduced haptic data volume by 99% while maintaining over 95% force feedback fidelity. Clinical trials at Guangzhou Medical University's affiliated hospital demonstrated the system's capability to support remote surgeries over 300km

distances. Key technical bottlenecks persist: haptic semantic quantification lacks standardization, leading to variations in force feedback across manufacturers' devices; multisensory (visual-tactile-auditory) semantic synchronization precision requires further improvement to sub-millisecond levels; and network jitter's impact on delicate operations remains unresolved, currently affecting surgical safety when jitter exceeds 50ms.

(5) Emergency and Disaster Communication

Emergency scenarios highlight the unique value of semantic communication. China Mobile's [3] disaster response system employs semantic priority classification (e.g., "casualty location" as highest priority, "supply needs" as secondary), maintaining critical communications even with 80% network damage. During actual deployment in Henan floods, the system tripled rescue response speed. Japan's NTT [16] earthquake early-warning system uses semantic compression to reduce seismic wave analysis data from 10GB to 10MB, providing 20-second earlier warnings. However, extreme environment challenges persist: in complete network outages, current semantic ad-hoc networking maintains only 30% terminal connectivity; semantic information's anti-destruction capability is insufficient, with ~10% critical data loss during multi-hop transmission. Furthermore, cross-team semantic understanding consistency requires improvement, as semantic discrepancies between systems may cause ~15% command misinterpretations.

3. Technical Challenges and Future Research Directions

3.1 Technical Challenges

(1) Standardization and Interoperability

The primary challenge for 6G semantic communication lies in the lack of a unified standardization framework [17]. Different vendors' AI large models employ diverse semantic encoding schemes, leading to severe interoperability limitations. For instance, semantic understanding deviations between Huawei's semantic encoder and Ericsson's semantic decoder may reach up to 30% [18]. This issue is particularly acute in industrial IoT scenarios requiring multi-vendor device coordination. Furthermore, semantic communication necessitates a full-stack innovation from the physical layer to the application layer, rendering existing communication protocol stacks inadequate. There is an urgent need to establish a semantic communication protocol architecture akin to TCP/IP, encompassing unified semantic representation methods, interface specifications, and evaluation standards. While the International Telecommunication Union (ITU) has initiated related standardization efforts, it will likely take 3-5 years to develop a mature framework.

(2) Computational Complexity and Energy Efficiency Bottlenecks

The high computational complexity of AI large models significantly hinders the practical deployment of semantic communication [3]. Experimental data show that processing 4K video semantic features requires approximately 100 TOPS of computing power, drastically increasing power consumption in terminal devices. Qualcomm's research [9] indicates that even with dedicated 4nm-process chips, running models with tens of billions of parameters still demands over 2W of power, making it difficult to meet mobile devices' battery life requirements. Although edge computing can offload some computational tasks, edge nodes remain resource-constrained in large-scale deployment scenarios (e.g., one million connections per square kilometer) [19]. More critically, the energy consumption during model training is even more staggering—training a trillion-parameter model generates carbon emissions equivalent to 300 cars running for a year. This drives researchers to explore more efficient architectures, such as Mixture-of-Experts (MoE) and Spiking Neural Networks (SNN).

(3) Semantic Security and Privacy Protection

Semantic communication introduces entirely new security threats [10]. Studies reveal that adversarial attacks can deceive semantic models into misinterpreting "stop" as "accelerate," posing severe risks to safety-critical applications like autonomous driving. Alibaba's security team found [13] that adversarial attacks targeting semantic encoders achieve success rates as high as 65%. Privacy

leakage risks are equally critical, as semantic extraction may expose sensitive user information. For example, semantic features in medical images could implicitly reveal patient identities. Traditional bit-level encryption cannot directly protect semantic privacy, necessitating the development of semantic-aware privacy-preserving techniques [20]. While federated learning mitigates privacy risks from centralized data, model parameters themselves may leak semantic information in communication scenarios. This demands novel security mechanisms, such as combining semantic obfuscation with differential privacy.

(4) Robustness in Dynamic Environments and Cross-Language Adaptability

6G semantic communication must address reliability challenges in extreme conditions [4]. Test data indicate that in environments with strong electromagnetic interference, the bit error rate of current semantic communication systems is 5-8 percentage points higher than traditional methods. Fast time-varying channels caused by the Doppler effect distort semantic features, reducing recognition accuracy by up to 20% in high-speed rail scenarios. Cross-language challenges are even more complex [21], as the same semantic meaning may carry different interpretations across cultures—e.g., nodding signifies negation in some regions, contrary to most countries' conventions. Current cross-language semantic alignment technologies achieve only ~85% accuracy, far below commercial requirements. Additionally, dynamic network topologies (e.g., drone swarms) require models with online learning capabilities, yet existing methods suffer performance fluctuations of up to 30% during updates. This calls for more robust continual learning algorithms and adaptive semantic encoding strategies.

3.2 Future Research Directions

(1) Native AI Communication Architecture

Future 6G networks will evolve towards native AI architectures where semantic intelligence is deeply embedded across all network layers [1]. This requires co-designing AI models with communication protocols from the ground up, enabling end-to-end semantic optimization. Key innovations include semantic-aware resource allocation and cross-layer semantic routing [16], which could improve network efficiency by 10-100x compared to traditional approaches.

(2) Green Semantic Computing

Energy-efficient semantic processing is critical for sustainable 6G deployment [3]. Promising approaches include neuromorphic computing chips for semantic tasks and dynamic semantic compression algorithms [21]. Recent breakthroughs show potential for 100x energy reduction in semantic encoding through hybrid analog-digital circuits and sparsity-aware computing paradigms.

(3) Trustworthy Semantic Communication

Secure semantic systems must address novel attack vectors while preserving privacy [20]. Emerging solutions combine quantum-resistant semantic encryption with blockchain-based verification [18]. For medical applications, differential privacy techniques can achieve 99% semantic accuracy while protecting sensitive patient data, representing a major advance over conventional methods.

(4) Semantic Metaverse Infrastructure

The metaverse demands ultra-efficient semantic exchange between physical and virtual worlds [11]. Next-gen systems will leverage holographic semantic compression and tactile semantic coding [15] to enable immersive experiences at 1/1000 the bandwidth of current solutions. Pilot demonstrations show 10ms latency for full-body semantic avatars, nearing the threshold for realistic interaction.

4. Conclusion

This paper provides a systematic review of the research progress and practical applications of artificial intelligence large models (AIM) in driving 6G semantic communications. Research demonstrates that AIM, through its powerful semantic understanding and processing capabilities, has

achieved breakthrough advancements in typical 6G scenarios such as holographic communication, digital twins, and intelligent vehicle-to-everything networks, significantly enhancing communication efficiency and intelligent capabilities. At the theoretical level, semantic information theory has established a new paradigm for 6G communication system design. At the technical level, continuous breakthroughs have been made in key technologies including efficient semantic encoding, lightweight deployment, and security enhancement.

However, several challenges remain to be addressed for large-scale commercial deployment, including insufficient standardization and interoperability, terminal computing efficiency bottlenecks, emerging semantic security threats, and dynamic environmental adaptability. Future research should focus on AIM-native communication architecture design, breakthroughs in green semantic computing technologies, establishment of trustworthy semantic communication systems, and construction of semantic metaverse infrastructure. Addressing these challenges requires not only technological innovation but also close collaboration among global industry, academia, research institutions, and standards organizations.

By resolving these critical issues, the deep integration of 6G and AIM will transform communication systems from traditional information pipelines into intelligent cognitive systems, laying a solid foundation for the development of the digital economy.

References

[1] Chen, X., Guo, Z., Wang, X., Feng, C., Yang, H. H., Han, S., Wang, X., & Quek, T. Q. S. (2025). Toward 6G Native-AI Network: Foundation Model-Based Cloud-Edge-End Collaboration Framework. IEEE Communications Magazine, 63(8), 23-30.

[2] Shoaib, M. R., Wang, Z., & Zhao, J. (2024). The Convergence of Artificial Intelligence Foundation Models and 6G Wireless Communication Networks. In 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring) (pp. 1-10). IEEE.

[3] Abasi, A. K., Aloqaily, M., & Guizani, M. (2025). Anomaly detection in 6G networks using large language models (LLMs). In 2025 International Wireless Communications and Mobile Computing (IWCMC) (pp. 1466–1471). IEEE.

[4] Xu, M., Niyato, D., Kang, J., Xiong, Z., Mao, S., Han, Z., Kim, D. I., & Letaief, K. B. (2024). When Large Language Model Agents Meet 6G Networks: Perception, Grounding, and Alignment. IEEE Wireless Communications, 31(6), 63-71.

[5] Long, S., Tang, F., Li, Y., Tan, T., Jin, Z., Zhao, M., & Kato, N. (2025). 6G Comprehensive Intelligence: Network Operations and Optimization Based on Large Language Models. IEEE Network, 192-201.

[6] Guo, S., Wang, Y., Ye, J., Zhang, A., Zhang, P., & Xu, K. (2025). Semantic importance-aware communications with semantic correction using large language models. IEEE Transactions on Machine Learning and Communications Networking, 3, 232-244.

[7] Jiang, S., Lin, B., Wu, Y., & Gao, Y. (2024). LINKs: Large language model integrated management for 6G empowered digital twin networks. In 2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall) (pp. 1–5). IEEE.

[8] Guo, R., Zhang, W., Yang, D., Wang, H., & Zhang, H. (2025). 6G Enabled Generative AI Services for Secure High-Speed Railway Networks. IEEE Network.

[9] Zhang, W., Ren, J., Zheng, T., Guo, R., Zhang, H., Mao, S., & Zhang, H. (2025). GenNet: Computing-efficient generative AI for deterministic transmission scheduling in 6G networks. IEEE Communications Magazine, 63(8), 50–55.

[10] Huang, X., Xue, K., Chen, L., Han, J., Li, J., & Wei, D. S. L. (2025). ForenSiX: Automated network forensics and diagnostics for beyond-5G and 6G networks using large language models. IEEE Network.

[11] Du, J., Lin, T., Jiang, C., Yang, Q., Bader, C. F., & Han, Z. (2024). Distributed foundation models for multi-modal learning in 6G wireless networks. IEEE Wireless Communications, 31(3), 20–30.

[12] Tao, Z., Xu, W., Huang, Y., Wang, X., & You, X. (2024). Wireless Network Digital Twin for 6G: Generative AI as A Key Enabler. IEEE Wireless Communications, 31(4), 24-31.

[13] Ferrag, M. A., Debbah, M., & Al-Hawawreh, M. (2023). Generative AI for cyber threat-hunting in 6G-enabled IoT networks. In 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW) (pp. 16–25). IEEE.

[14] Liu, C., & Zhao, J. (2024). Resource Allocation in Large Language Model Integrated 6G Vehicular Networks. In 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring) (pp. 1-10). IEEE.

[15] Li, Z., Chen, Z., Hu, X., & Yang, H. H. (2025). Personalized generative AI services through federated learning in 6G edge networks. China Communications, 22(7), 1–13.

[16] Betalo, M. L., Ullah, I., Tesema, F. B., Wu, Z., Li, J., & Bai, X. (2025). Generative AI-driven multi-agent DRL for task allocation in UAV-assisted EMPD within 6G-enabled SAGIN networks. IEEE Internet of Things Journal.

[17] Cen, S., & Zhu, Y. (2025). NP-LLM: A unified large-language-model-assisted framework of 6G network-layer planning. IEEE Communications Magazine, 63(8), 92–98.

[18] Brodimas, D., Trantzas, K., Agko, B., Tziavas, G. C., Tranoris, C., Denazis, S., & Birbas, A. (2024). Towards Intent-based Network Management for the 6G System adopting Multimodal Generative AI. In 2024 European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit): Network Softwarisation (NET) (pp. 1-8). IEEE.

[19] Huang, L., Wu, Y., & Simeonidou, D. (2025). Reasoning AI performance degradation in 6G networks with large language models. In 2025 IEEE Wireless Communications and Networking Conference (WCNC) (pp. 1–6). IEEE.

[20] Munir, M. S., Proddatoori, S., Muralidhara, M., Dena, T. M., Saad, W., Han, Z., & Shetty, S. (2025). A Trust-By-Learning Framework for Secure 6G Wireless Networks Under Native Generative AI Attacks. IECE Open Journal of the Communications Society.

[21] Tao, Z., Xu, W., Huang, Y., Wang, X., & You, X. (2024). Wireless Network Digital Twin for 6G: Generative AI as A Key Enabler. IEEE Wireless Communications, 31(4), 24-31.