Generative AI Empowering Semantic Communication in 6G Industrial Scenarios: A Review of Architecture, Key Technologies and Challenges

Chengbo Yang *

Nanjing University of Posts and Telecommunications, Nanjing, China

* Corresponding author email: clauden2036590039@163.com

Abstract

6G industrial communication places higher demands on real-time performance, reliability and cross-terminal task collaboration. Under conditions of limited bandwidth and low signal-to-noise ratio, traditional bit-layer communication has become difficult to efficiently support distributed tasks. Generative artificial intelligence (GAI) provides a new paradigm for task-oriented semantic communication (GSC) with its capabilities in deep semantic modeling, cross-modal generation, and conditional reconstruction. This paper systematically reviews the architecture and key technologies of GAI in 6G industrial semantic communication and clarifies the implementation ideas of modules such as semantic extraction and compression, cross-modal alignment, robust reconstruction and online adaptation. Three typical scenarios, namely intelligent manufacturing, distributed robot collaboration and cross-enterprise supply chain, were selected to analyze the impact of GSC on bandwidth, latency and task success rate. Finally, operational research and engineering routes are proposed for the evaluation criteria, engineering deployment, credibility and standardization. This article aims to provide systematic references and practical suggestions for researchers and engineering practitioners.

Keywords

Generative AI; 6G; Industrial scenario; Semantic communication.

1. Introduction

With the evolution of the industrial Internet towards intelligence, flexibility, and collaboration, the next generation 6G network in industrial scenarios is faced with not only higher throughput and lower latency but also, more importantly, semantic-level information exchange to support real-time and distributed task execution [1, 2]. In typical 6G industrial scenarios (intelligent manufacturing, distributed robot collaboration, cross-enterprise supply chain, etc.), it is often faced with the challenges of limited bandwidth, mobility, and low Signal-to-Noise Ratio (SNR), which makes the transmission of full raw sensing data inefficient and impractical [3]. Task-oriented semantic-level switching can significantly reduce communication overhead and response delay under the premise of ensuring the success rate and has become the key paradigm of 6G industrial communication.

Generative AI (GAI), including Generative adversarial networks (GAN), Variational Autoencoders (VAE), diffusion models, and large pre-trained models (LLMs), has shown unique advantages in high-order semantic modeling and cross-modal generation [4]. Introducing GAI into semantic communication can bring three key capabilities. The first is to map multimodal observations into compact semantic representations that are highly relevant to downstream tasks to reduce the amount of transmission. The second is robust reconstruction based on semantic priority and knowledge base at the receiver to compensate for channel impairments or incomplete observations [5]. The third is to support end-to-end optimization with task loss as the goal, so as to directly couple communication protocol design and task performance. These capabilities enable GAI-driven Semantic Communication (GSC) to maintain good task-level performance and improve system robustness under resource-constrained or unreliable links.

This paper reviews the application of GAI in 6G industrial semantic communication, sorts out the GSC architecture and key technologies, analyzes its role in improving efficiency, reducing bandwidth, and enhancing robustness through typical scenarios, and points out future challenges and research paths.

2. System Architecture and Key Modules

2.1. Overview of the Overall Architecture

The GSC for 6G industrial scenarios adopts a hierarchical and modular end-to-end system:

The perception layer is responsible for multimodal data acquisition (visual, acoustic, force/vibration, timing sensors, etc.);

The semantic extraction/compression coding layer maps the raw data into compact semantic representations related to the task and performs priority labeling.

The semantic signaling/transport layer schedules semantic code streams with a task-oriented strategy when network conditions are limited.

The receiving end recovers or reconstructs the information required for the task through the semantic decoding/generation module (based on the generation model and knowledge base).

Ultimately, control, prediction or decision-making is accomplished by the task execution layer.

The overall workflow of semantic communication with generative models is shown in **Fig. 1**, reprinted from [3].

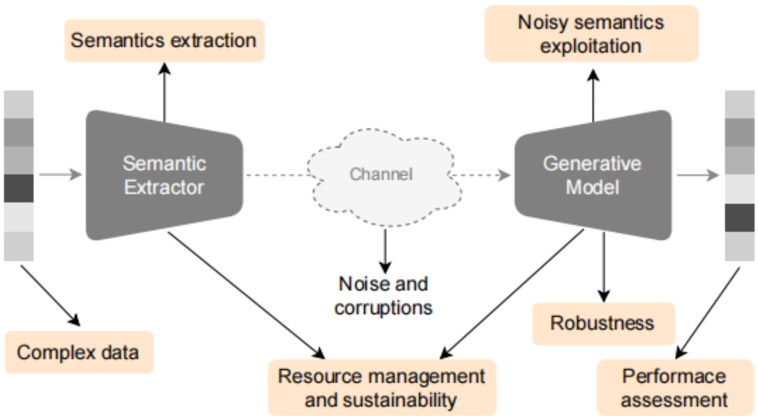

Fig. 1 Semantic communication framework overview with its challenges from the machine learning perspective. Reprinted from [3].

This architecture emphasizes the collaboration among end/edge/cloud: lightweight models are rapidly extracted and preprocessed at the terminal or edge, while complex semantic fusion and global reasoning are completed in the edge cloud, thereby reducing bandwidth and energy consumption while ensuring real-time performance.

2.2. Semantic Extraction and Compression Coding

The core task of the semantic extraction module is to transform high-dimensional and multi-modal raw observations into low-dimensional semantic representations that are strongly related to downstream tasks such as fault diagnosis, target localization, and action decision-making. The design should center on "task effectiveness." Common technical approaches include representation learning based on self-supervised or contrastive learning, end-to-end training aimed at task loss (such as task-oriented autoencoders), and structured semantic modeling using knowledge graphs (for ease of constraints and interpretability). Engineering design should take into account three points simultaneously: First, the task effectiveness of semantic representation (which can directly support the target task); second, the interpretability and auditability of the representation (facilitating fault source tracing and safety verification); third, the deployability of the model (supporting pruning, quantization, or distillation to adapt to edge devices) [3, 5].

Semantic compression coding is to map high-dimensional semantic representations to low-dimensional code streams in order to reduce the transmission burden. Technically, lossy coding, conditional VAE, or generative compressors that take semantics as conditional input (such as conditional diffusion/conditional generator) can be adopted to reconstruct semantics or signals at the receiving end based on the conditional information. Research shows that generative compression can better preserve task-relevant information in scenarios with limited bandwidth or low SNR [5].

2.3. Semantic Transmission and Robustness Design

Semantic layer transmission differs from traditional bit layer transmission. Its design focus shifts to "priority guarantee of key semantics" and "semantic integrity verification." Common engineering approaches include bitstream priority scheduling based on importance ranking, semantic-level forward error correction and hierarchical coding, as well as semantic subset selection strategies under bandwidth or SNR constraints. Generative models (such as conditional Gans, conditional diffusion models, or graph/transformer-based generators) can be used at the receiving end to reconstruct missing or damaged semantic information, treating communication as an inverse problem and using generative priors to compensate for channel deficiencies and noise, thereby ensuring task-level performance without fully restoring all the original data. However, generative refactoring introduces issues such as model synchronization, version control, and credibility evaluation: the receiving end must be able to determine the reliability of the refactoring results (such as confidence estimation) and trigger fallback or request retransmission mechanisms when there is uncertainty in critical tasks [6].

2.4. Cross-modal Fusion and Knowledge Base Support

Industrial scenarios often involve multimodal information such as visual, mechanical, acoustic, and text. The goal of cross-modal semantic fusion is to align these heterogeneous modalities into a unified semantic space. The implementation methods include jointly trained cross-modal Transformers, modal invariant representations based on contrastive learning, and semantic constraints and ambiguity resolution in combination with knowledge graphs [5]. The knowledge base (domain-specific ontology/graph) plays a dual role: on the one hand, it provides structured priors for semantic alignment; on the other hand, it provides constraints for the generative model to suppress hallucinations and improve interpretability. Synchronization strategies, privacy protection, and access control surrounding knowledge bases are also issues that must be addressed in practice [7].

2.5. Online Adaptation and Model Management

The industrial environment is dynamic and changeable. Semantic communication systems need to have rapid adaptive capabilities to cope with equipment aging, changes in working conditions, or the launch of new tasks. It is recommended to adopt a hierarchical model update strategy: perform low-cost local fine-tuning at the terminal/edge (such as few-shot adaptation, LoRA/low-rank fine-tuning, or distillation), and periodically carry out global model retraining and quality verification in the cloud [3]. Model distribution must be accompanied by signature verification and version control. Online evaluation metrics (including semantic fidelity, task success rate, and reconstruction credibility) should be embedded into the closed loop of automated model management. The lack of a sound model management and evaluation mechanism makes it difficult to achieve secure, controllable, and scalable deployment of GAI-driven semantic communication [8].

3. Analysis of Typical Industrial Scenarios

3.1. Intelligent Manufacturing

In the workshop environment, a large number of heterogeneous sensors (visual, vibration, temperature, acoustic) are deployed to continuously collect the status of equipment and products, generating a vast amount of raw data. Traditional full-volume transmission to the cloud for fault diagnosis or quality control can lead to bandwidth bottlenecks and high response delays [8].

GSC extracts and transmits "semantic events" directly related to faults/quality at the edge, significantly reducing the transmission burden. Research shows that it can significantly reduce the transfer load in several tasks (e.g., some IoT directory tasks reduce the communication overhead by about 40%, MU-DeepSC can maintain task accuracy when transferring only 1/3 of the data) [1], and is expected to shorten the response period (depending on the specific task and loop). In cases of low SNR or packet loss, the generative model at the receiving end can reconstruct key information based on the context and knowledge base to maintain diagnostic accuracy [4].

3.2. Distributed Robot Collaboration

Multi-robot/human-machine teams need to frequently exchange intentions, environmental perception (local maps), motion suggestions, etc. to achieve close collaboration (such as assembly, handling, or inspection). The transmission of raw perception data (such as images and point clouds) or detailed state information by traditional methods can lead to high communication load, high decision loop delay, and a sharp drop or even failure of collaborative efficiency when bandwidth is limited or mobile networks are unstable [9].

GSC maps fine-grained perception to compact semantics (such as target position, grasping posture, and feasible path description), significantly reducing the amount of data in a single interaction, making decision iterations more frequent and lowering the communication overhead (some cases show up to 40%) [4, 5]. In scenarios of network interruption or high packet loss, a completion mechanism based on prediction and generation can maintain basic collaborative capabilities, supplemented by semantic priority scheduling to ensure the priority transmission of control information [6].

3.3. Cross-enterprise supply chain collaboration

All nodes in the supply chain (suppliers, manufacturers, and logistics providers) need to efficiently share information such as inventory, production capacity, transportation status, and demand forecasts to optimize global scheduling. Traditional point-to-point data exchange has problems such as heterogeneous formats, high information redundancy, large synchronization delay, and involves the privacy of sensitive business data [10].

GSC maps business data into standardized high-level semantic summaries (such as "Critical Components: Tight/Normal/Sufficient") by sharing domain ontologies/knowledge graphs, which can significantly reduce redundant communication and synchronization delays, and enhance decision consistency within the semantic framework [11]. The generative model can infer missing fields and initiate targeted retransmission, when necessary, thereby balancing privacy and efficiency) [12].

Overall, GSC can maintain or enhance task-level performance in industrial scenarios through task-oriented semantic abstraction and generative reconstruction under resource-constrained or unreliable links. However, its actual effectiveness is limited by model capabilities, deployment costs, and the degree of standardization.

The above analyses are summarized in **Table 1**, which highlights the main challenges, GSC-enabled solutions, and corresponding benefits across typical industrial scenarios.

Table 1. Summary of Typical Industrial Scenarios Enabled by GSC

Scenario	Challenge	GSC Solution	Benefit
Intelligent Manufacturing	Bandwidth bottleneck; high delay	Semantic event extraction; generative reconstruction	Reduce transmission load; Shorten the response cycle
Distributed Robot Collaboration	Heavy data load; unstable networks	Compact semantics (position, posture, path); predictive completion; priority scheduling	Communication overhead ↓ up to 40%; collaboration continuity
Cross-enterprise Supply Chain	Heterogeneous data; redundancy; privacy risk	Semantic summaries (ontology/knowledge graph); generative inference	Less redundancy & delay; better consistency; privacy protection

4. Challenges and Future Directions

Although GAI can enhance efficiency, save bandwidth, and improve robustness, industrial deployment is still constrained by four major factors: the lack of evaluation benchmarks, the energy consumption and latency of sinking large models to the edge, the credibility and compliance risks of generated outputs, and the insufficiency of standardized and open benchmarks. This article summarizes the problems in each dimension and provides research suggestions to promote comparability and engineering implementation.

4.1. The measurement and evaluation framework is missing

Most current studies are based on their own tasks and datasets, lacking comparable and unified benchmarks. It is suggested to construct a three-dimensional evaluation system: task success, semantic fidelity, and system resource consumption (bandwidth/energy consumption/delay), and define clear calculation methods and confidence intervals for each indicator [6, 7].

4.2. Engineering and Energy Consumption Constraints

Large-scale generative models consume huge resources in training and inference and are not suitable for direct deployment to resource-constrained industrial edges. Meanwhile, model compression and edge collaboration will have an impact on semantic fidelity. In the future, a compression process combining pruning, quantification, and distillation can be promoted. Adopt edge-cloud hierarchical inference and parameter-efficient fine-tuning (LoRA/PEFT); and build an energy-performance trade-off baseline on typical hardware to facilitate engineering evaluation and selection [8, 11].

4.3. Credibility, Security and Compliance

Generative models can produce hallucinations or unexplainable outputs, which may lead to serious consequences in safety-critical tasks such as industrial control and fault diagnosis. At the same time, cross-enterprise semantic sharing involves privacy and legal liability issues. In the future, it is necessary to establish output credibility measurement, semantic validators, and fallback strategies (fallback to bit-level verification or manual review when the credibility is low), and design auditable decision logs and responsibility traceability mechanisms. Constraining the generation process with knowledge graphs helps to reduce hallucinations and enhance interpretability. Cross-enterprise deployment also requires clearly defining data permissions, model update responsibilities, and incident handling processes [10, 12].

4.4. Standardization and Open Benchmark Construction

The absence of semantic interfaces, data formats, and evaluation protocols hinders interoperability and industrialization. It is suggested that the "Semantic Communication Benchmark Initiative" be initiated in the academic circle, defining several representative industrial tasks, datasets, and evaluation protocols, and promoting the connection with standard organizations such as oneM2M to promote the standardization of semantic interfaces, version management, and secure exchange specifications [13, 14].

4.5. Future Research Directions

In exploring the potential of GAI in industrial applications, Vajjhala et al. (2025) provide a comprehensive overview of GAI methods that are driving innovation in the industrial sector [15]. They highlight that GAI can handle complex industrial data and offer new solutions for industrial automation and optimization through generative models. These methods range from automating processes in smart factories to managing complex supply chains, demonstrating the significant potential of GAI in enhancing efficiency and flexibility. Vajjhala et al.'s work underscores the crucial role of GAI in Industry 4.0 and 5.0, providing valuable references for future research and practice.

5. Conclusion

Generative artificial intelligence (GAI) brings three core capabilities to 6G industrial semantic communication: compressing multimodal observations into task-related compact semantic representations, robust reconstruction at the receiving end based on prior knowledge and the knowledge base, and end-to-end optimization driven by task loss [1, 4]. Case studies in intelligent manufacturing, distributed robots, and cross-enterprise supply chains show that GSC is expected to bring tangible benefits in reducing bandwidth occupation, shortening response latency, and enhancing anti-interference capabilities [3, 10].

To achieve industrial-scale deployment, priority should be given to promoting unified semantic evaluation standards, edge-oriented model lightweighting and hierarchical reasoning, trusted generation and auditing mechanisms, as well as early integration with industry standard organizations.

The potential of GAI in accelerating Industry 4.0 and 5.0 has been highlighted by Boareto et al. (2025) [16], who discuss how GAI can drive innovation and efficiency in industrial production and logistics. This aligns with the broader trend of integrating AI into the Internet of Things (IoT), as discussed by Wen et al. (2024) [17], who outline the fundamentals and future outlooks of generative IoT. These advancements are indicative of a new era where AI-generated solutions are becoming ubiquitous, as noted by Du et al. (2024) [18].

By advancing technology, engineering, and governance in a coordinated manner, GSC can be safely, controllably, and scalably implemented in real industrial environments. Future research should focus on addressing the identified challenges and exploring innovative solutions to fully realize the potential of GAI in 6G industrial communicatio

References

[1] X. Luo, H.-H. Chen, and Q. Guo, “Semantic Communications: Overview, Open Issues, and Future Research Directions,” IEEE Wireless Commun., vol. 29, no. 1, pp. 210–219, Feb. 2022.

[2] D. Wheeler and B. Natarajan, “Engineering Semantic Communication: A Survey,” IEEE Access, vol. 11, pp. 13965–13995, 2023.

[3] Y. Wang, “Semantic Communication Networks Empowered Artificial Intelligence of Things,” in 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT), Melbourne, Australia: IEEE, July 2024, pp. 189–193.

[4] E. Grassucci, Y. Mitsufuji, P. Zhang, and D. Comminiello, “Enhancing Semantic Communication with Deep Generative Models: An Overview,” in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 13021–13025.

[5] E. Grassucci, J. Park, S. Barbarossa, S.-L. Kim, J. Choi, and D. Comminiello, “Generative AI Meets Semantic Communication: Evolution and Revolution of Communication Tasks,” Jan. 10, 2024, arXiv: arXiv:2401.06803.

[6] B. K. Saha, “Generative Artificial Intelligence for Industry: Opportunities, Challenges, and Impact,” in 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan: IEEE, Feb. 2024, pp. 081–086.

[7] X. Yafei, Y. Wu, J. Song, Y. Gong, and P. Lianga, “Generative AI in Industrial Revolution: A Comprehensive Research on Transformations, Challenges, and Future Directions,” JKLST, vol. 3, no. 2, pp. 11–20, June 2024.

[8] S. Sai, R. Sai, and V. Chamola, “Generative AI for Industry 5.0: Analyzing the Impact of ChatGPT, DALLE, and Other Models,” IEEE Open J. Commun. Soc., vol. 6, pp. 3056–3066, 2025.

[9] A. Lykov et al., “Industry 6.0: New Generation of Industry driven by Generative AI and Swarm of Heterogeneous Robots,” Sept. 16, 2024, arXiv: arXiv:2409.10106.

[10] J. Patalas-Maliszewska, M. Szmołda, and H. Łosyk, “Integrating artificial intelligence into the supply chain in order to enhance sustainable production—a systematic literature review,” Sustainability, vol. 16, no. 16, p. 7110, 2024.

[11] A. Willner, C. Diedrich, R. Ben Younes, et al., “Semantic communication between components for smart factories based on oneM2M,” in 2017 IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE, 2017, pp. 1–8.

[12] A. K. Kar, V. P. S. Varsha, and S. Rajan, “Unravelling the impact of generative artificial intelligence (GAI) in industrial applications: A review of scientific and grey literature,” Global Journal of Flexible Systems Management, vol. 24, no. 4, pp. 659–689, 2023.

[13] T. Ahmed and S. Choudhury, “AI and Semantic Communication for Infrastructure Monitoring in 6G-Driven Drone Swarms,” Feb. 26, 2025, arXiv: arXiv:2503.00053.

[14] N. L. Rane, “ChatGPT and similar generative artificial intelligence (AI) for smart industry: role, challenges, and opportunities for Industry 4.0, Industry 5.0, and Society 5.0,” Innov. Bus. Strateg. Manag., June 2024.

[15] N. R. Vajjhala, S. S. Roy, B. Taşcı, and M. E. Hoque Chowdhury, Eds., Generative Artificial Intelligence (AI) Approaches for Industrial Applications, vol. 24. in Information Systems Engineering and Management, vol. 24. Cham: Springer Nature Switzerland, 2025.

[16] P. A. Boareto, A. L. Szejka, E. F. R. Loures, F. Deschamps, and E. A. P. Santos, “Accelerating Industry 4.0 and 5.0: The Potential of Generative Artificial Intelligence,” in Innovative Intelligent Industrial Production and Logistics, vol. 2372, M. Dassisti, K. Madani, and H. Panetto, Eds., in Communications in Computer and Information Science, vol. 2372. , Cham: Springer Nature Switzerland, 2025, pp. 456–472.

[17] J. Wen et al., “From Generative AI to Generative Internet of Things: Fundamentals, Framework, and Outlooks,” IEEE Internet Things M., vol. 7, no. 3, pp. 30–37, May 2024.

[18] H. Du et al., “The Age of Generative AI and AI-Generated Everything,” IEEE Network, vol. 38, no. 6, pp. 501–512, Nov. 2024.