



# Generative AI for 6G IoT- Using Digital Twin as an Example

Fangyu Liu

College of Engineering, Yanbian University, Changchun, Jilin, China  
1224024641@ybu.edu.cn

**Abstract.** As 6G evolves towards high-density intelligent networks, integrating digital twins and generative AI is vital for 6G IoT. This paper explores generative AI in 6G Intelligent Radio digital twin. It first clarifies the synergy between intelligent radio and digital twin in 6G, and reveals how GAN, Diffusion Model, and RAG-LLM enhance data, optimize transmission, and generate policies. Generative AI outperforms traditional methods by 8.3% - 50% in resource allocation and noise resistance. The study also identifies challenges like physical-digital asynchrony and heavy edge-computing load, suggesting solutions. Finally, it anticipates generative AI's applications in the metaverse and smart grid, guiding 6G IoT's intelligent development. The significance of this research lies in its pioneering exploration of generative AI's transformative potential for 6G IoT, bridging critical gaps between theoretical frameworks and practical implementations. By systematically addressing key challenges and demonstrating measurable performance improvements, the study establishes a foundational roadmap for future intelligent network evolution. Furthermore, it provides actionable insights for industry stakeholders to accelerate 6G standardization while fostering sustainable development of next-generation IoT ecosystems. The findings also highlight the strategic importance of AI-native network design in achieving global connectivity goals for emerging smart cities and Industry 5.0 applications.

**Keywords:** Generative AI, 6G IoT, Digital Twin, Intelligent Radio, Network Optimization.

## 1 Introduction

With the goal of interconnection and intelligent ubiquity, 6G networks need to realize intelligent resource allocation, low-latency communication and cross-layer collaboration in a high-density radio environment, and the traditional network architecture is difficult to meet the optimization requirements in dynamic scenarios. Digital twin technology can simulate network state, fault prediction and optimize strategies in real time by constructing a virtual mapping of the physical system, while generative AI, with its data generation, pattern learning and decision reasoning capabilities, has become the core of improving digital twin accuracy and efficiency. For example, intelligent radio, as a key technology for 6G, needs to dynamically adjust transmission parameters in complex electromagnetic environments, and generative AI can optimize beamforming

© The Author(s) 2026

S. Zhang (ed.), *Proceedings of the 2025 International Conference on Electronics, Electrical and Grid Technology (ICEEGT 2025)*, Advances in Engineering Research 292,

[https://doi.org/10.2991/978-94-6463-986-5\\_50](https://doi.org/10.2991/978-94-6463-986-5_50)

by simulating the signal propagation law to reduce channel estimation errors. Theoretical Level This study reveals the fusion mechanism of generative AI and digital twins to improve the theoretical system of 6G network intelligence. Technical Level This study breaks through the performance bottleneck of traditional models in data-scarce and noise-sensitive scenarios and improves the utilization of network resources. At the application level, this research provides 6G network optimization solutions for industrial manufacturing, smart healthcare and other vertical fields, and promotes the realization of the digital twin as a service model.

## **2 Technology convergence of generative AI and 6G digital twins**

### **2.1 Synergistic architecture for smart radio and digital twins**

The 6G smart radio uses a distributed radio access network (RAN) architecture, which enables full domain modelling of the physical, link and application layers through digital twins. Generative AI plays a triple role in which the Data Enhancer utilizes GAN models to generate synthetic wireless signal data to alleviate model training challenges in small sample scenarios [1-4]. For example, in channel estimation, GAN generates multipath fading samples by learning the real signal distribution to improve the generalization ability of neural networks for complex channels [3]. The transmission optimizer uses the diffusion model to optimize the anti-jamming strategy in wireless transmission through the process of forward diffusion and backward denoising. The policy generator utilizes the RAG-LLM framework to combine knowledge base retrieval with large language model reasoning to provide dynamic policy recommendations to network designers [5]. For example, when network congestion is detected, RAG-LLM can quickly retrieve historical solutions and generate a combined policy of switching service nodes + dynamic bandwidth adjustment.

## **3 Typical technology comparison analysis**

### **3.1 RAG-supported LLMs**

The RAG architecture consists of database, retrieval, and decision-making modules; the database splits the stored academic text into knowledge blocks, and the embedded model generates vectors to be stored into the database; The retrieval module converts user requests into vectors and matches the knowledge block extension context; The decision-making module generates optimization problems through LLMs (e.g. ChatGPT) and combines them with historical strategies to improve response accuracy [6]. At the same time, RAG works in synergy with LLM to reduce the error of LLM in generating optimization problems, reduce the number of token calls, and improve efficiency by utilizing the external retrieval knowledge base of RAG. The generated problem is solved by the generative diffusion model GDM, which is utilized to deal with

high-dimensional, noise-containing complex scenarios as a way to dynamically output the optimal policy. This technique can address the limitations of traditional AI, where discriminative AI relies on fixed data patterns, is difficult to cope with some dynamic environments, and requires frequent repetitive training leading to high energy consumption; It is also possible to optimize the problem generation inefficiency, manual design of carbon reduction strategies requires a lot of domain knowledge, and it is easy to miss the conditions of constraints leading to time-consuming and long-time; It can improve the data and computation bottlenecks, and the performance of traditional methods in data-scarce and noise-sensitive scenarios is insufficient (e.g. Energy optimization of edge devices) [7], while this RAG+LLM architecture improves the robustness through data augmentation and semantic understanding.

This technique designs pluggable LLM and RAG modules and introduces a GDM solver, which integrates the problem of optimizing the dynamic generation of knowledge bases (e.g., the bandwidth of the meta-universe AIGC task), replaces the traditional reinforcement learning algorithms (e.g., PPO), and improves the efficiency of the policy generation and the noise immunity (in the case study, the performance of the GDM has a 17.97% higher accuracy than that of the PPO[7]). The RAG-based LLM framework can help LLM agents to perform accurate reasoning, and the improved accuracy enables RAG's cue engineering to better perform fine-grained information retrieval and generation based on finely tuned cues, efficiency is optimized by the fact that the framework reduces the risk of manual errors and improves the accuracy of the problem formulation compared to manual design. Currently, RAG-LLM has been applied to scenarios such as energy internet, data center network and mobile edge network, for example, optimizing energy routing from vehicles to the grid in energy internet, balancing power loss and carbon reduction targets; and managing ICT and cooling system energy consumption in data center network. This method can optimize the deployment of intelligent reflective surfaces, the lack of dynamic adaptation, and significantly improve the accuracy and stability of the strategy compared with the traditional method. However, the performance is limited by the database coverage, and the lack of domain-specific data will affect the effectiveness of the strategy[8]; RAG retrieval and LLM require a certain amount of arithmetic power, and the deployment of edge devices faces resource constraints; for extreme emergency scenarios, the dynamic response capability of this technology needs to be improved, and needs to be optimized in combination with the online learning mechanism.

### 3.2 Digital twins for hierarchical GAI wireless networks

The hierarchical digital twin consists of two levels: a message-level digital twin used externally to exchange information and a policy-level digital twin that utilizes the internal actions and states of the digital network. The messaging layer is based on the Transformer model that transforms messages between network entities into sequential text for training to model the core network control plane functions, thus enabling the modeling of messages external to the network; the policy layer then utilizes a multi-layer neural network, combined with the data augmentation capabilities of Adversarial Generative Networks (GANs), to model the internal states and actions of the network,

and finally optimize the network policy through Deep Reinforcement Learning (DRL) [8, 9]. The two work together to improve the simulation accuracy of digital twins for complex network behavior. This technology addresses three core challenges in 6G networks: physical-digital modeling, synchronization, and slicing capabilities. For physical-digital modeling, traditional model-driven approaches require extensive programming and are difficult to adapt to heterogeneous scenarios, data-driven approaches, on the other hand, face the problem of data scarcity, whereas GAI can alleviate the data shortage through data augmentation, while using Transformer and diffusion models to improve modeling efficiency; for the synchronization problem, a generative transmission scheme based on the diffusion model can be used to compress the key information and reduce the noise reduction, so that it can be adapted to the requirements of 6G air-air-heaven integrated network with wide-area coverage; for the slicing capability, generative transfer learning can enable digital twins to quickly adapt to the customization needs of different slices, solving the problem of insufficient slicing adaptation of traditional methods. Building digital twins through programming, mathematical modeling, and artificial intelligence approaches for performance evaluation of core networks, resource management of radio access networks, and network topology digital twins to improve latency prediction. In data-driven approaches, neural networks, Bayesian neural networks, and recurrent neural networks are used for modeling different network components.

Transformer's parallel architecture and attention mechanism make its message prediction accuracy in highly concurrent scenarios significantly better than traditional models such as LSTM; the synthetic data generated by GAN extends the diversity of training samples and mitigates the problem of high cost and privacy sensitivity of data collection in 6G networks; the noise reduction capability of the diffusion model enhances the reliability of data transmission in the wireless channel and reduces the synchronization delay and bandwidth consumption; the migration learning and fine-tuning mechanism digital twins can quickly adapt to the resource allocation strategies of different slices. However, the technology still suffers from the challenges of cross-layer collaboration, where multilayer architectures may introduce synchronization delays and the mechanisms for digital twin collaboration across different network elements are not yet fully standardized; the large-scale signaling of generative AI still faces the risk of eavesdropping, and the encryption mechanism needs to be further enhanced; the deployment complexity is high, and the computational power of edge devices may not be able to support large-scale models such as Transformer, while centralized deployment suffers from a single point of failure and high latency; there is a lack of a unified Pingu metrics system, and performance benchmarks for different slicing scenarios are not yet clear.

### **3.3 PPO methods with MoE in satellite communications**

The technique combines a hybrid expert (MoE) architecture with proximal policy optimization (PPO) to build a framework for hierarchical decision making. MoE handles specific optimization variables separately through multiple expert networks, and each expert optimizes the sub-tasks it is good at through an independent neural network and

then uses a gating network to dynamically aggregate expert outputs based on the input states to form a joint optimization strategy. The PPO updates the policy parameters through an Actor-Critic structure, where the Actor-network integrates the MoE's expert decisions and the Critic network evaluates the state values to guide the policy optimization. To deal with the high-dimensional state space and complex action space in satellite communications, MoE-PPO employs hyperbolic tangent functions to constrain the power and rate variables and reduces the modeling complexity by state decomposition [5]. This technology mainly solves two major core problems in 6G communication, one is the resource optimization problem in complex scenarios, such as the dynamic resource allocation of multi-users and multi-slices in heterogeneous satellite networks, and the traditional single model is difficult to take into account the efficiency and accuracy of the multi-variable joint optimization; second, transmission interference and heterogeneous constraints management, such as beam interference between satellites, user quality of service guarantee, MoE-PPO can dynamically select expert combinations according to user distribution and channel state, optimize beam formation to reduce the same-frequency interference between LEO and GEO satellites, and at the same time, meet the differentiated resource requirements of different slices. Satellite communication resource optimization mostly adopts traditional optimization algorithms or a single reinforcement learning model, but this technique combines the gating mechanism to achieve dynamic aggregation of strategies and breaks through the performance bottleneck of the traditional single model.

This technique optimizes efficiency gains, with MoE-PPO improving about 5% in sum-rate optimization compared to PPO, converging faster, and keeping the policy effective especially when there is an increase in the number of households; It supports different access protocols and optimization objectives, and can be adapted to diversified scenarios through expert extension; the gating network dynamically selects the optimal expert combination, and the anti-interference ability in time-varying channels is better than the traditional method, and the power constraint satisfaction rate reaches more than 98%. However, the computational complexity of this technique is high, and as the number of experts increases, the parameter update of the gated network and the aggregation of the experts' outputs impose additional computational overhead; the expert division of labor granularity and gating mechanism weight initialization affect convergence stability and need to be tuned by grid search; the state space is highly dependent, and the existing state coding may not be able to capture all the key features in a sudden increase in the number of users or high-dimensional interference scenarios; decision latency of centralized MoE-PPOs in distributed satellite networks may affect real-time services and requires optimized deployment architecture in conjunction with edge computing.

### 3.4 Diffusion modeling with AIGC

The denoising mechanism of the diffusion model is utilized to generate high-quality AIGC content, where inputs such as text and images are transformed into potentially noisy sequences, which are progressively denoised by a multilayer U-Net network to generate the target content. The technical process includes forward diffusion and backward denoising, and is guided by the introduction of classifiers to enhance the semantic

alignment of the generated content [2]. In the AIGC scenario, the diffusion model supports the tasks of text-to-image generation and image restoration, and balances the generation speed and quality by adjusting the hyperparameters such as the number of denoising steps and the noise variance [1]. This technique mainly solves the problems of mode collapse and detail blurring of traditional AIGC models in generating complex contents, especially in multimodal generation, the diffusion model can better capture the subtle differences of data distribution and generate contents with higher diversity. For example, in medical image generation, the diffusion model can generate high-fidelity synthetic data based on a small number of real samples; in the field of creative design, it supports the generation of multi-style images from natural language descriptions, which breaks through the limitations of traditional regularized generation. GAN-dominated AIGC suffers from the defects of insufficient diversity of generation samples and unstable training, while VAE generates lower accuracy, generates dynamic contracts through the diffusion process, solves the problem of information asymmetry in resource allocation, and improves the incentive efficiency by 15%-20% compared with the traditional DRL method [7-9].

This technique generates high quality and supports 1024×1024 resolution image generation with better detail texture than GAN; It can be trained using a small amount of labeled data, which is suitable for data-scarce scenarios such as medical and remote sensing, strong flexibility, you can control the style and theme of the generated content by adjusting the conditional inputs to achieve personalized customization; support for continuous optimization; supports sequential optimization for smooth content transitions by interpolating the latent space. However, the technique is computationally expensive, requiring dozens of iterations for a single generation and a large video memory footprint; training is time consuming and requires thousands of epochs to converge; controllability limitations, weak ability to generate local details of content, needs to be combined with prompt engineering or additional optimization modules[10], difficulty in reverse generation, difficulty in reverse generation and inference of accurate text descriptions from images, limiting the application of interactive authoring scenarios.

## 4 Challenges and developments

In terms of model training and optimization, generative AI models are large in scale, and the training and inference process requires strong arithmetic support [2]; the communication environment is complex and variable, and the communication data in different scenarios have different characteristics, and the performance of some of the current models may decline in cross-scenario applications; promoting the innovation of AI chip technology can increase the chip's computational capacity, unify the standard of arithmetic and the software-hardware interfaces, and integrate a variety of heterogeneous arithmetic; in terms of data processing, data privacy and security, the communication field involves a large amount of sensitive information of users, when using generative AI, it is crucial to ensure the privacy and security of data in the process of collection [1], transmission, storage and processing, using encryption technology to encrypt the transmitted and stored data, to ensure the confidentiality and integrity of the data;

in terms of real-time requirements, the real-time response of communication systems to complex AI models may result in longer processing delays, network transmission delays [3], and data transmission can be limited by network bandwidth and latency; in edge device applications, it is necessary to achieve dynamic allocation of resources with the help of generative AI [5], to solve the problem of high power consumption of complex models such as RAG-LLM [7, 8], and to reach a balance between model efficiency and load, and it is possible to study the strategy of updating the parameters of the funny model compression amount, to reduce the amount of data transmitted during the communication process, to optimize the energy management of the AI system, and to reduce the energy consumption by adopting the strategies of intelligent dormancy and dynamic adjustment of the arithmetic power; in order to achieve real-time synchronization between the physical network and the virtual model in hierarchical digital twins, it is necessary to focus on optimizing the RAN digital twin modeling accuracy. Adaptive AI agents and intelligent resource scheduling algorithms can be developed to cope with the dynamic topology and resource allocation challenges of satellite networks to cope with their communication complexity. At the same time, encryption and synthetic data technologies are utilized to protect privacy, avoiding the leakage of sensitive information due to randomness in generative AI and ensuring data security. And to enhance the model generalization ability, to solve the problem of diffusion models and other performance degradation when there is insufficient training data or drastic changes in the network state, so as to make it better adapt to the dynamic environment.

## 5 Conclusion

Generative AI significantly improves the intelligence of 6G IoT digital twins through data augmentation, policy optimization and cross-layer modeling. It has significant advantages over traditional methods in terms of resource allocation efficiency and noise robustness, but challenges such as synchronization and computational load still need to be broken through. Future research should focus on the integration of multiple technologies, privacy protection mechanisms and industry standardization, to promote generative AI from the laboratory to 6G commercial deployment and provide the underlying technical support for the digital economy and intelligent society.

## References

1. H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. A. Heng, S. Z. Li, A survey on generative diffusion models. *IEEE Trans. Knowl. Data Eng.* (2024)
2. V. Chamola, G. Bansal, T. K. Das, V. Hassija, S. Sai, J. Wang, D. Niyato, Beyond reality: The pivotal role of generative AI in the metaverse. *IEEE Internet Things Mag.* 7(4), 126-135 (2024)
3. Y. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, A. Jamalipour, Deep generative model and its applications in efficient wireless network management: A tutorial and case study. *IEEE Wirel. Commun.* (2024)

4. S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, M. Liyanage, Explainable AI for 6G use cases: Technical aspects and research challenges. *IEEE Open J. Commun. Soc.* (2024)
5. R. Zhang, H. Du, Y. Liu, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, Generative AI agents with large language model for satellite networks via a mixture of experts transmission. *IEEE J. Sel. Areas Commun.* (2024)
6. J. Wen, R. Zhang, D. Niyato, J. Kang, H. Du, Y. Zhang, Z. Han, Generative AI for low-carbon artificial intelligence of things with large language models. *IEEE Internet Things Mag.* **8**(1), 82-91 (2024)
7. F. Khoramnejad, E. Hossain, Generative AI for the optimization of next-generation wireless networks: Basics, state-of-the-art, and open challenges. *IEEE Commun. Surv. Tutor.* (2025)
8. H. Du, D. Niyato, J. Kang, Z. Xiong, P. Zhang, S. Cui, D. I. Kim, The age of generative AI and AI-generated everything. *IEEE Netw.* (2024)
9. Z. Tao, W. Xu, Y. Huang, X. Wang, X. You, Wireless network digital twin for 6G: Generative AI as a key enabler. *IEEE Wirel. Commun.* **31**(4), 24-31 (2024)
10. X. Xu, R. Zhong, X. Mu, Y. Liu, K. Huang, Mobile edge generation-based digital twins: Architecture design and research opportunities. *IEEE Commun. Mag.* (2024)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

