

GenNet: Computing-Efficient Generative AI for Deterministic Transmission Scheduling in 6G Networks

Weiting Zhang, Jiadong Ren, Tao Zheng, Ruibin Guo, Han Zhang, Shiwen Mao, and Hongke Zhang

ABSTRACT

With the development of the sixth generation (6G) networks, ubiquitous intelligence, computing, and networking integration, low energy consumption, and low delay have become the key characteristics. These characteristics pose significant challenges for traffic scheduling, particularly in the context of intelligent computing services. In recent years, generative artificial intelligence (AI) has demonstrated remarkable performance not only in natural language processing and image generation but also in network optimization. This article presents a generative AI-endogenous three-layer network architecture, named GenNet, to support more efficient scheduling of intelligent computing services in 6G networks by the dynamic adaptation of heterogeneous computing resources and diversified service requirements. Moreover, we propose a dueling double deep Q-network (D3QN)-based transmission scheduling algorithm that leverages diffusion models to achieve cross-domain end-to-end deterministic transmission. The proposed algorithm facilitates efficient interaction between edge devices and intelligent computing centers while reducing system cost and delay, and utilizes the denoising capability of the diffusion model to adaptively configure networks and optimize resource allocation. Finally, we present a case study, followed by a discussion of open research issues that are essential for generative AI and future networks.

INTRODUCTION

International industrial, academic, and standard groups have started researching sixth-generation (6G) communication systems as the requirements of society for networks increase. The Third Generation Partnership Project (3GPP) has planned to finalize the 6G base version standard, also known as the Rel-21 version standard, by 2029 [1]. Compared with fifth-generation (5G) networks, 6G will not only improve performance indicators such as bandwidth and delay, but also realize *ubiquitous intelligence* and computing and networking integration. In such a case, artificial intelligence (AI) technologies will be deeply embedded into the 6G network architecture rather than as an external enhancement [2]. For example, traffic

scheduling in 6G networks can be accomplished by deploying AI agents inside switches and routers, thereby facilitating efficient transmission for multimodal data. As a result, AI is expected to be the core driver of 6G networks, bringing great benefits to society and the economy.

6G networks support larger-scale intelligent access, such as Internet of vehicle (IoV) scenarios, which may generate a large amount of high-dimensional multimodal information [3]. This information needs to be transmitted deterministically between end nodes and computing centers. However, existing optimization algorithms have difficulty in efficient transmission scheduling. In such a case, generative AI provides a new solution for solving the efficient deterministic transmission scheduling problem in 6G scenarios [4]. Particularly, due to the unique denoising capability, diffusion models are expected to be applied to traffic scheduling and routing mechanisms [5].

Developing deterministic transmission scheduling algorithms faces many challenges in 6G networks due to the emerging requirements [6]. First, at the information processing level, the multimodal data generated by a large number of diversified applications needs to be extracted and unified into features, which is critical for AI agents to efficiently perceive network states. In addition, users' semantic requirements need to be accurately represented to support deterministic transmission for multimodal data. Second, the interaction between end nodes and computing centers is frequent, which may cross multiple heterogeneous networks, for example, time-sensitive networking (TSN) and deterministic networking (DetNet), which further complicates the deterministic scheduling for multimodal data. Therefore, it is necessary to develop generative AI-assisted solutions to facilitate cross-domain end-to-end deterministic transmission scheduling in 6G networks.

In this article, we study the cross-domain deterministic transmission scheduling problem in 6G networks and present a novel generative AI-assisted network architecture named GenNet to support deterministic transmission scheduling for intelligent computing services. Specifically, the architecture is composed of three layers: the intent abstraction layer, the mapping adaptation

Weiting Zhang, Jiadong Ren, Tao Zheng (corresponding author) Ruibin Guo, and Han Zhang are with Beijing Jiaotong University, China; Han Zhang is with Tsinghua University, China; Shiwen Mao is with Auburn University, Auburn, USA.

Digital Object Identifier: 10.1109/MCOM.001.2400588

Ref	Research Problem	Contribution	Solution	Future Network Architecture	Deterministic Transmission	Generative AI
[6]	Improve the adaptability of LLMs in the network to achieve “one model for all”	NetLLM is proposed, which can effectively adapt LLM to solve network problems	Design multimodal encoder, networking head, data-driven low-rank networking adaptation to adapt LLM	✗	✗	✓
[7]	Empower LLMs with knowledge and expertise in the wireless domain	A WirelessLLM framework is proposed to equip LLMs with knowledge and expertise in the wireless domain	Build the WirelessLLM by prompt engineering, retrieval augmented generation (RAG), tool usage, and the processes of pre-training and fine-tuning	✓	✗	✓
[8]	Improve the generalization ability of traffic scheduling algorithms in TSN	A traffic-aware TSN scheduler, namely DeepScheduler, based on DRL is proposed to learn effective scheduling strategies	Scalable state information encoder and path-based flow-aware encoder are designed to model the network topology	✓	✓	✗
[9]	Overcome the barriers of inter-domain microbursts not considered by TSN and DetNet	An efficient heuristic algorithm is proposed to solve the scheduling problem and reduce the search complexity in multi-domain networks	Establish I2DCM relationships to solve bandwidth, cycle, and queue mismatch issues between different queuing mechanisms	✗	✓	✗
[10]	Reduce the need for human effort and achieve expert-free problem optimization	AI-Generated Network is proposed to achieve expert-free resource scheduling	The original data distribution is first obtained via online RL, and then the expected solution strategy is learned using offline RL with diffusion models	✓	✗	✓
[11]	Integrate generative AI with trust management for reliable real-time communication	A GAN-based trust decision-making model is proposed to determine whether a network device is trustworthy	The trust vector of network devices is used as the input of the GAN-based autoencoder to test whether it satisfies the distribution of the training data	✓	✗	✓
[12]	Solve the routing and scheduling problem of mixed-critical flows	A DRL-based flow scheduler is proposed to support efficient routing and scheduling for deterministic flows	Enhance the efficiency and schedulability by dividing a complete flow schedule and action masking	✗	✓	✗
Ours	Integrate diffusion models and DRL to achieve efficient cross-domain end-to-end traffic deterministic transmission scheduling in 6G networks	A three-layer network architecture and a diffusion model-enhanced scheduling algorithm are proposed to support deterministic transmission scheduling	Utilize the denoising capability of the diffusion model-enhanced algorithm to generate traffic scheduling decisions	✓	✓	✓

TABLE 1. Generative AI-assisted deterministic transmission scheduling for future networks.

layer, and the network execution layer. Moreover, we propose a deterministic transmission scheduling algorithm based on dueling double deep Q-network (D3QN), in which the diffusion model is utilized to make traffic scheduling decisions through a denoising module. Simulation results show that the proposed algorithm outperforms the benchmark algorithm in terms of convergence speed, average system cost, and delay.

The remainder of this article is organized as follows: Cases of generative AI for 6G networks are discussed in the next section. The three-layer architecture and the generative AI-assisted deterministic scheduling algorithm are presented following that. A case study is then provided. Open research issues are discussed, followed by the conclusion.

GENERATIVE AI FOR 6G NETWORKS

Table 1 illustrates that generative AI has been extensively employed for network optimization

and deterministic transmission scheduling in 6G, particularly with diffusion models and large language models (LLMs). Taking the diffusion model as an example, a denoising module that efficiently reverses the multi-step noise process is employed to produce high-quality samples for deterministic transmission scheduling. In addition, LLMs also have extensive applications, such as LLM wireless knowledge bases [7], intent-driven network [10], and trust decision models [11].

APPLICATION OF GENERATIVE AI IN NETWORK SCENARIOS

The advent of generative AI has revolutionized numerous fields, especially in network technologies, with its content generation capabilities in resource allocation and transmission scheduling [13–15]. For resource allocation, generative AI models can analyze vast amounts of network data, including user behavior, traffic patterns, and network topology, to predict future requirements

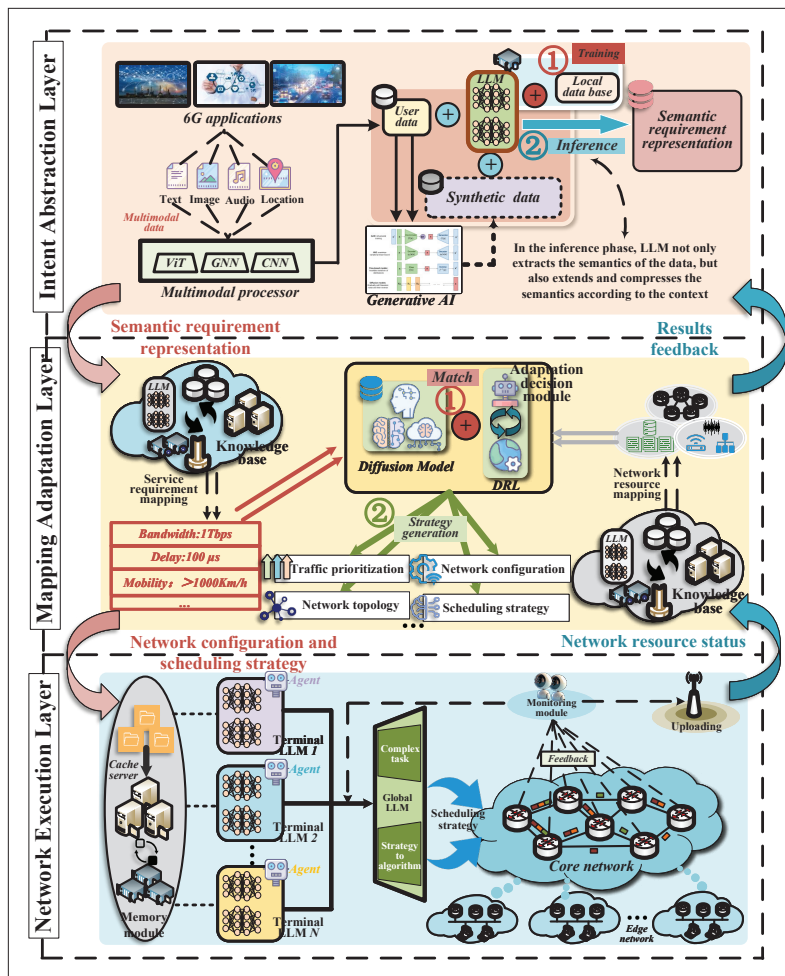


FIGURE 1. Generative AI-assisted three-layer architecture for 6G networks.

and optimize resource distribution. Through simulating various network scenarios, these models can ensure that multi-dimensional resources, such as bandwidth, computing, and storage, are allocated efficiently. This is particularly crucial in network scenarios with limited resources, such as IoV. For transmission scheduling, traditional approaches generally rely on fixed rules or heuristic algorithms, which may not adapt well to dynamic network conditions. Generative AI models, however, can learn from historical data and adapt to changing network conditions in real time. This capability allows for more intelligent scheduling decisions, ensuring that traffic is transmitted efficiently and deterministically. For example, generative AI models can prioritize high-priority traffic, such as control signaling or real-time gaming, over lower-priority traffic, such as data acquisition or file downloads.

GENERATIVE AI TOWARD 6G NETWORKS

In the 6G era, deterministic transmission has become a representative feature of networks, enabling characteristics such as bounded delay, low jitter, high reliability, and highly accurate time synchronization. Generative AI can significantly enhance the performance of deterministic scheduling algorithms through intelligent prediction and adaptive optimization [8]. In TSN, generative AI can predict the network load and optimize the scheduling strategy for time-sensitive traffic. Det-

Net can dynamically adjust the routing strategy to ensure cross-domain end-to-end deterministic transmission. Furthermore, generative AI is crucial in the integration scheduling of deterministic mechanisms across network domains [9]. In terms of heterogeneous deterministic mechanisms, generative AI intelligently recognizes their types and dynamically generates optimal deterministic transmission parameters, for example, gate control list (GCL), for each mechanism, ensuring that uniform quality of service (QoS) is maintained in complex heterogeneous network environments. In summary, the combination of generative AI and 6G can facilitate heterogeneous deterministic mechanism integration, thereby supporting mission-critical applications in 6G networks.

INTERPLAY BETWEEN GENERATIVE AI AND 6G NETWORKS

Generative AI is not a unilateral gain for 6G networks; both of them are complementary. On one hand, generative AI can optimize 6G networks in terms of resource allocation and transmission scheduling. First, 6G users have more diverse requirements on computing and networking that need to be accurately perceived and analyzed. Generative AI understands user requirements through its natural language processing capabilities and translates them into executable machine parameters. Second, 6G applications consume a lot of network resources to collect and transmit model training data. Generative AI is capable of saving network resources by distributed learning and predicting data for model training. On the other hand, as an infrastructure that connects various devices and services, 6G networks have the ability to provide massive and diverse real-time data for the self-optimization of generative AI, which is crucial to improving the generation ability and computing efficiency of generative AI models. Moreover, generative AI can also take into account real-time feedback from the 6G network, allowing for dynamic adjustments and further optimization. Meanwhile, the widespread deployment of generative AI in 6G networks also results in synergistic management by distributed AI agents, which in turn facilitates developing advanced 6G architectures to support emerging intelligent computing services.

COMPUTING-EFFICIENT GENERATIVE AI FOR 6G NETWORKS

In this section, we present a novel three-layer network architecture and design a deterministic scheduling mechanism to support cross-domain end-to-end transmission for intelligent computing services.

GENERATIVE AI-ASSISTED 6G NETWORK ARCHITECTURE

As shown in Fig. 1, the architecture is composed of three layers, that is, intent abstraction layer, mapping adaptation layer, and network execution layer. The specific functions of each layer are as follows.

Intent Abstraction Layer: This layer is directly connected to the application and user ends, deploying a multimodal processor and a generative AI module. By processing multimodal data and combining user data with synthetic data, the system efficiently extracts user semantic requirement representation to pass to the next layer for service requirement mapping.

Multimodal Data Processing: Emerging scenarios such as telemedicine and IoV generate a large amount of data in different modalities, such as text, image, audio, and video. These data are first translated into a unified representation through a multimodal processor. For example, a vision transformer (ViT) model is utilized to divide an image into patches of fixed size, and each patch is treated as a serialized token, but for text, each word can be treated as a token.

Generative AI Module: After unifying the multimodal data into serialized tokens, a generative AI module analyzes the distribution of the serialized tokens, and then generates new samples that match this distribution, enabling the prediction of users' behavior data. Based on user data and synthetic data, user semantic requirement representation can be finally extracted.

Mapping Adaptation Layer: In this layer, two knowledge bases and an adaptation decision module are deployed, which can realize the mapping of service requirements and network resources and strategy generation. The adaptation decision module adapts the mapped network resource status according to the service requirement and generates the optimal adaptation decision and scheduling strategy. Network resource status is uploaded from the network execution layer to this layer.

Service Requirement Knowledge Base: This base is established to map the semantic requirement representation extracted by the intent abstraction layer into specific QoS indicators, such as bandwidth and delay requirements. The knowledge base defines an effective search space through an LLM that specifies the path to generate QoS indicators from the semantic requirement representation.

Network Resource Knowledge Base: Similarly, the network resource knowledge base is established to map the real-time network status uploaded by the network execution layer into quantitative network indicators, such as link utilization and queue length. Both knowledge bases are sustainable due to the dynamic variation of user requirements and network conditions.

Adaptation Decision Module: A diffusion model-enhanced DRL algorithm is proposed and deployed in an adaptation decision module to adaptively make the optimal scheduling decisions for deterministic transmission. Since a denoising structure is embedded into the algorithm, computing-efficient decision-making can be achieved with the denoising generation capability. In addition, quantitative network indicators and QoS indicators are matched in the module by the proposed algorithm, where the former is viewed as the network state information and the latter is viewed as the optimization objective of the algorithm. The state information is denoised under the guidance of QoS indicators to generate the scheduling strategy.

Network Execution Layer: This layer contains multiple terminal LLMs, a global LLM, and network hardware devices. This layer configures and manages the network according to the strategy passed by the upper layers. Meanwhile, its monitoring module continuously feeds real-time network resource status to upper layers.

Terminal LLM and Global LLM: End users download terminal LLMs from cache servers that

are deployed at the network edge and perform personalized initialization on their devices, such as personal or vehicle information. In addition, terminal LLMs can also obtain historical state data from the memory module. Global LLMs can store long-term memory aggregated by terminal LLMs through an aggregate memory layer, and refine the scheduling algorithm into a specific scheduling strategy by in-context learning and fine-tuning.

Scheduling Strategy Deployment: Through global LLM, we deploy scheduling strategies and specific configurations to the edge network (i.e., TSN) and core network (i.e., DetNet). Global LLM configures the GCL, idle slope, and send slope generated by the algorithm to the router to control the queue.

DETERMINISTIC TRANSMISSION SCHEDULING IN 6G

In 6G networks, the diffusion model is adopted to implement cross-domain deterministic transmission scheduling. The deployment and scheduling process of deterministic mechanisms is as follows.

As shown in Fig. 2, we consider a 6G network scenario that is composed of two access networks and a core network, in which different deterministic mechanisms are deployed. In the access networks, TSN technology is adopted for local-area networking, and the time-aware shaper (TAS) mechanism is deployed in edge switches. The TAS reduces network congestion, delay, and jitter by precisely controlling the sending time of data packets and calculates an accurate GCL based on the routing path and time slots to ensure deterministic transmission. In the core network, DetNet technology is adopted for wide-area networking. Based on TAS, the credit-based shaping (CBS) mechanism is introduced to flexibly control the transmission priority and rate by dynamically adjusting the credit value of the data flow, which not only ensures deterministic transmission but also improves the efficiency of link resource utilization. Taking into account cross-domains between TSN and DetNet, learning agents based on diffusion models are deployed in the edge switches, named diffusion-agent, to dynamically generate the GCL, idle slope, and send slope parameters based on real-time network conditions. The cross-domain interaction process involves three key steps:

- During cross-domain transmission, the diffusion agent first analyzes historical network status data to predict potential congestion points.
- Based on this analysis, it dynamically adjusts the GCL parameters and optimizes transmission scheduling.
- Additionally, the diffusion agent implements a feedback mechanism that continuously monitors data transmission performance and network resource utilization across domains, enabling real-time adjustment of scheduling strategies.

This adaptive cross-domain coordination ensures efficient traffic transmission between heterogeneous networks while maintaining deterministic performance guarantees.

GENNET: GENERATIVE AI-ENHANCED DRL FOR DETERMINISTIC TRANSMISSION SCHEDULING

As shown in Fig. 3, we propose a deterministic transmission scheduling algorithm based on diffusion model-enhanced D3QN. We deploy

A diffusion model-enhanced DRL algorithm is proposed and deployed in an adaptation decision module to adaptively make the optimal scheduling decisions for deterministic transmission. Since a denoising structure is embedded into the algorithm, computing-efficient decision-making can be achieved with the denoising generation capability.

The IIoT scenario consists of two TSN domains (each with eight industrial sensors and two controllers) and one DetNet domain (with six edge servers and one cloud server), all configured in ring topologies.

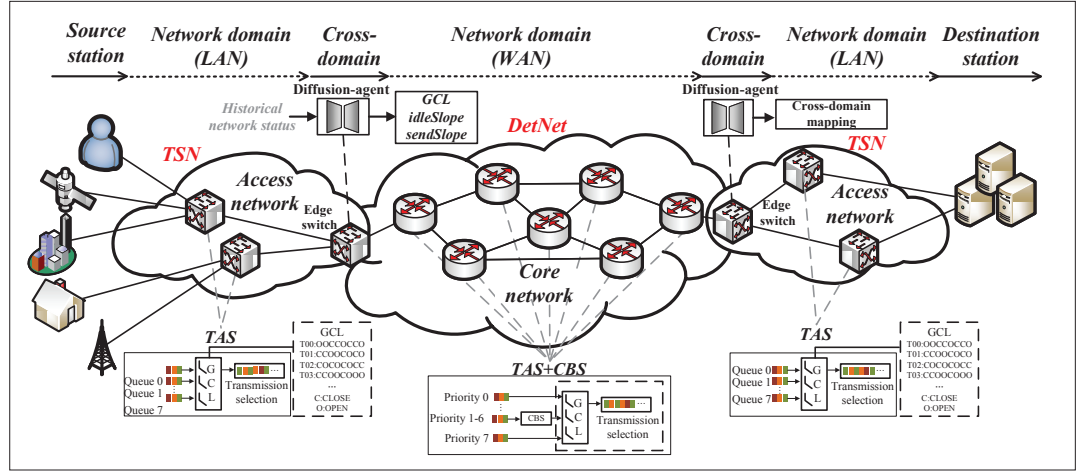


FIGURE 2. Deterministic transmission scheduling in 6G.

diffusion agents in each switch and router to make decisions through their denoising generation capability. The operation procedure of the proposed algorithm is as follows. First, a data abstraction module observes the link status and topology structure of the data plane that consists of user ends and deterministic network devices, and abstracts them into a data structure. Second, it is input into a requirement description module, which integrates the flow information and link load into state vector S and inputs it into the diffusion agents. Third, the diffusion agents generate action vectors (i.e., deterministic scheduling decisions), regard the current network resource status as noise, perform iterative denoising according to service requirements, generate preliminary configuration actions, and input them into the action translator. An action translator then converts the action value into a vector A , which contains GCL, idleslope, and sendslope. Finally, the data abstraction module configures scheduling decisions to the data plane, updates the network status, and obtains corresponding rewards based on delay, jitter, and throughput. The core elements of the algorithm are defined as follows.

Network State: We define the observed state space as $S = \{o_1, o_2, \dots, o_N\}$, where $o_n = (G, L, Q, F, T)$ represents the state at time slot n , $\forall n \in \{1, 2, \dots, N\}$. Here, L , Q , F , A and T denote the network topology, link load, queue state, flow feature, and delay requirement, respectively. In addition, queue state matrix $Q = \{q_{ij}\}_{i \in \mathcal{I}, j \in \mathcal{J}}$ denotes the state (i.e., length and priority) of the j -th queue of node i , and flow feature matrix $F = \{f_k\}_{k \in \mathcal{K}}$, where $f_k = (s_k, d_k, qos_k)$ denotes the source node, the destination node, and the QoS requirements of the k -th flow.

Agent Action: We define action space as $A = \{a_1, a_2, \dots, a_M\}$, where $a_m = (R, B, C_{tsn}, C_{det})$, $\forall m \in \{1, 2, \dots, M\}$. Here, $R = \{r_i^k\}_{i \in \mathcal{I}, k \in \mathcal{K}}$ is the routing decision matrix, where r_i^k represents the next hop selection of flow k at node i . $B = \{b_i^k\}_{i \in \mathcal{I}, k \in \mathcal{K}}$ is the bandwidth allocation matrix, where b_i^k represents the bandwidth allocated to flow k at node i . $C_{tsn} = \{GCL_i\}_{i \in \mathcal{I}}$ is the TAS configuration of the TSN domain, where GCL_i is the gating list of node i . $C_{det} = \{GCL_i, ID_i, SD_i\}_{i \in \mathcal{I}}$ is the configuration of the DetNet domain, where ID_i and SD_i represent the idle slope and send slope of node i .

System Reward: The reward function is used to evaluate the selected action a_m , $\forall m \in \{1, 2, \dots, M\}$. We define the reward function as $R(o, a) = w_1 \cdot R_{delay} + w_2 \cdot R_{jit} + w_3 \cdot R_{thr} - w_4 \cdot P_{vio}$, which represents the optimization objective. The reward function takes into account the most critical performance indicators of the deterministic scheduling mechanism, including delay, jitter, and throughput, while ensuring that QoS is met through constraint violation penalties. In addition, w_i , $\forall i \in \{1, 2, 3, 4\}$, is the weight of each item. The reward function effectively guides the proposed algorithm to learn a scheduling strategy that satisfies deterministic transmission requirements.

CASE STUDY

This section provides a case study on diffusion model-enhanced traffic scheduling with the objective of reducing the average system cost and delay.

SIMULATION SCENARIO

We evaluate our algorithm in two representative scenarios, both supporting end-to-end deterministic transmission for cross-domain task scheduling in 6G networks.

IoV Scenario: There are two TSN and one DetNet domains in the considered IoV scenario, where vehicle location sharing and real-time navigation are considered as primary scheduling tasks. The TSN domain adopts a star topology containing five sensor nodes and one controller, while the DetNet domain adopts a mesh topology with ten edge servers and two cloud servers.

Industrial Internet of Things (IIoT) Scenario: A smart manufacturing environment requiring precise control and real-time monitoring, where machine synchronization and production line coordination are considered as primary scheduling tasks. The IIoT scenario consists of two TSN domains (each with eight industrial sensors and two controllers) and one DetNet domain (with six edge servers and one cloud server), all configured in ring topologies. Diffusion agents are deployed on these servers, which are responsible for optimizing the resource allocation and traffic scheduling. There are numerous CPUs and GPUs with 1,200 cores and 1,500 TFLOPS of computing resources available in both the TSN and DetNet domains. Seven spectrum resource blocks are considered, with subcarrier spacing (SCS) ranging from 500 to 3,500 kbit/ms.

In recent years, network intelligence has received extensive attention, and the deep integration of generative AI and 6G will become a key research direction.

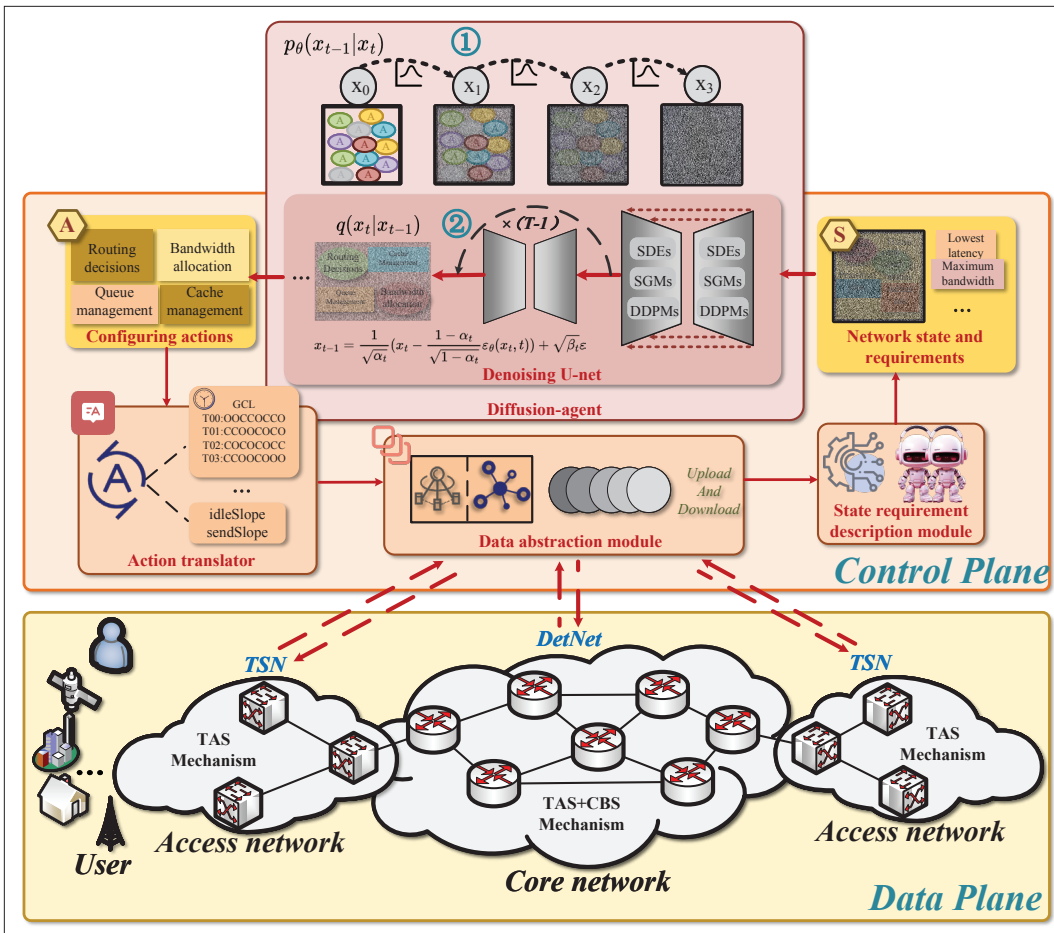


FIGURE 3. Diffusion model-enhanced DRL for deterministic transmission scheduling.

- We define three types of traffic:
- Periodic control flow with strict delay and jitter requirements
 - Video surveillance flow with high bandwidth and delay-sensitive requirements
 - Data acquisition flow with low priority and delay-tolerant requirements.

The total system cost of all K flows is defined as $V = \sum_{k=1}^K (\beta_s V_s^k + \beta_c V_c^k + \beta_q V_q^k)$, where V_s^k , V_c^k , and V_q^k represent the spectrum resource consumption, computing resource consumption, and QoS violation penalty for the k -th flow transmission, respectively.

Subject to the constraints of available computing resources, spectrum resource blocks, and the QoS requirements of different traffic types, the weights of β_s , β_c , and β_q are set to be 0.77, 12.43, and 2.61, respectively. A larger β_c reflects the priority of computing resources, while β_s and β_q are balanced to maintain spectrum efficiency and QoS requirements. These values were empirically optimized through grid search; we can practically adapt them to the importance of different constraints. For the diffusion model-enhanced algorithm, the number of neurons in the hidden layers is set to 128 and 64, respectively, and the learning rate is set to 0.0001. The considered scenario employs GPT-2 for terminal LLMs and GPT-3 for the global LLM, and all simulations run on NVIDIA GeForce RTX 4090 GPUs with the PyTorch framework. In addition, we adopt a D3QN-based transmission scheduling algorithm, a conventional

DQN-based scheduling algorithm, and a genetic algorithm (GA) for performance comparison.

SIMULATION RESULTS

We evaluate the convergence speed in the IIoT scenario by comparing the diffusion model-enhanced algorithm with the D3QN-based algorithm, the DQN-based algorithm, and genetic algorithms, while comparing the system cost and delay performance with the D3QN-based algorithm in the IoV scenario. As shown in Fig. 4a, it can be seen that the proposed algorithm has converged after 400 learning episodes, and the convergence performance is better than that of other benchmark algorithms. As shown in Fig. 4b, the results indicate that the proposed algorithm can obtain a lower average system cost and delay than that of the D3QN-based algorithm. Particularly, when the bandwidth is set to 3,500 kbit/ms, the average system cost and delay of the proposed algorithm can be reduced by 12.8% and 26%, respectively. This is because the diffusion model can adapt to different network conditions and generate the optimal bandwidth allocation solution, thus promoting efficient deterministic transmission scheduling via a denoising module.

OPEN RESEARCH ISSUES

In recent years, network intelligence has received extensive attention, and the deep integration of generative AI and 6G will become a key research direction. In the following, we discuss some open

This integration will significantly enhance future network systems by enabling advanced cross-layer optimization and resource allocation, and the strategies will reduce operational costs while improving QoS, particularly for delay-sensitive applications such as augmented reality (AR) and autonomous systems.

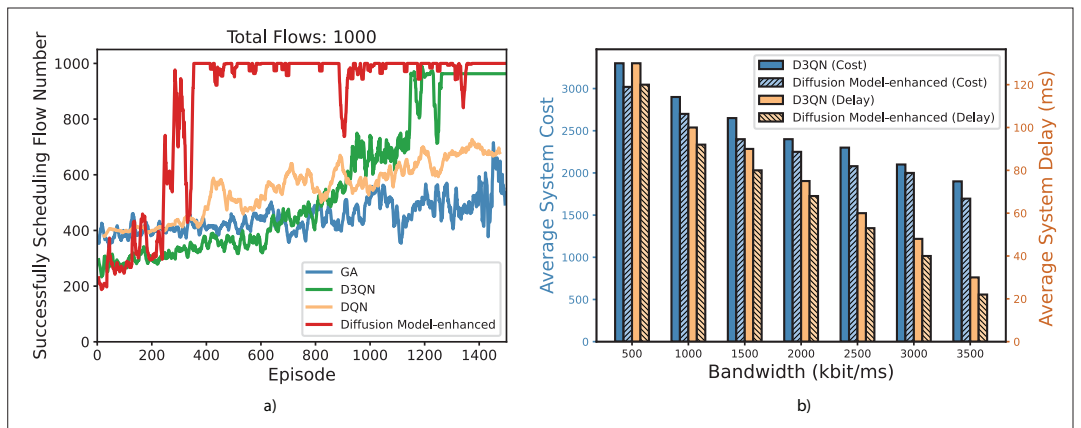


FIGURE 4. Performance evaluation of the proposed diffusion model-enhanced deterministic transmission scheduling algorithm: a) Convergence speed comparison in IIoT scenario; b) Cost and delay comparison in IoV scenario.

research issues on the synergy of generative AI and future networks (Fig. 5).

NOVEL NETWORK ARCHITECTURE FOR GENERATIVE AI

Future network architecture needs to be continuously optimized in terms of adaptive capabilities to better support generative AI technologies. On one hand, generative AI requires multi-dimensional resource management, including networking, computing, and storage, to facilitate efficient operation for data synthesis. On the other hand, the adaptive adjustment and feedback mechanism of the future network becomes especially important, as it puts forward high requirements to respond flexibly to rapidly changing network conditions and service requirements, thereby realizing highly intelligent network management.

Moreover, the evolution of this architecture faces critical implementation challenges in practice. The standardization of interfaces across diverse network domains and AI platforms is particularly vital, as it ensures seamless integration and interoperability. This requires coordinated industry efforts to develop unified protocols while maintaining compatibility with existing infrastructure investments.

COMPUTING AND NETWORKING INTEGRATION FOR GENERATIVE AI

As an emerging technological paradigm, generative AI stands poised to revolutionize contemporary society, posing significant challenges for resource collaboration. First, the widespread deployment of generative AI results in an exponential surge in resource requirements for diverse applications, such as content generation and the metaverse. Second, the generative AI-based applications require not only the efficient allocation of network resources but also the dynamic provision of heterogeneous computing resources. Consequently, the integration of computing and networking is an inevitable research direction for empowering future networks with generative AI.

This integration will significantly enhance future network systems by enabling advanced cross-layer optimization and resource allocation, and the strategies will reduce operational costs while improving QoS, particularly for delay-sensitive applications such as augmented reality (AR) and autonomous systems. This integration provides robust support for generative AI, enabling more efficient and responsive generative AI-driven services.

FEDERATED REINFORCEMENT LEARNING-BASED GENERATIVE AI

Federated reinforcement learning (FRL) offers a pioneering framework that enables multiple entities to collaborate on model training within a dynamic network environment while preserving data privacy. This solution effectively enhances the scale of training data for the generative AI by capturing a broader spectrum of information from the participating entities, ultimately improving the quality and diversity of the generated content. Furthermore, FRL facilitates the generative AI's adaptability across various fields and scenarios, enabling efficient knowledge transfer. Thus, FRL-based generative AI stands out as a promising research area that holds the potential to broaden the application of generative AI in future networks.

The key of FRL-based generative AI is data quality assurance and security enhancement. Encryption protocols and secure aggregation methods are necessary to develop to protect sensitive data during model training and parameter exchange. These considerations are crucial for building reliable and secure FRL-based generative AI systems that can be deployed in practice.

CONCLUSION

For 6G, network intelligence has been emerging as a pivotal development trend, with generative AI demonstrating remarkable advantages. In this article, we have introduced a novel three-layer network architecture leveraging computing-efficient generative AI to facilitate intelligent resource management and support emerging intelligent computing services in 6G networks. Furthermore, we have proposed a diffusion model-enhanced traffic scheduling algorithm to achieve cross-domain end-to-end deterministic transmission. By synergizing the DRL with the diffusion model, the proposed algorithm enables automatic configuration of network parameters, which can not only optimize resource management but also significantly improve network responsiveness for intelligent computing services. To foster the advancement of future networks, we have delved into prospective directions and trends of generative AI, exploring its potential applications in computing and networking integration, thereby contributing to constructing more robust, efficient, and adaptable 6G network infrastructures.

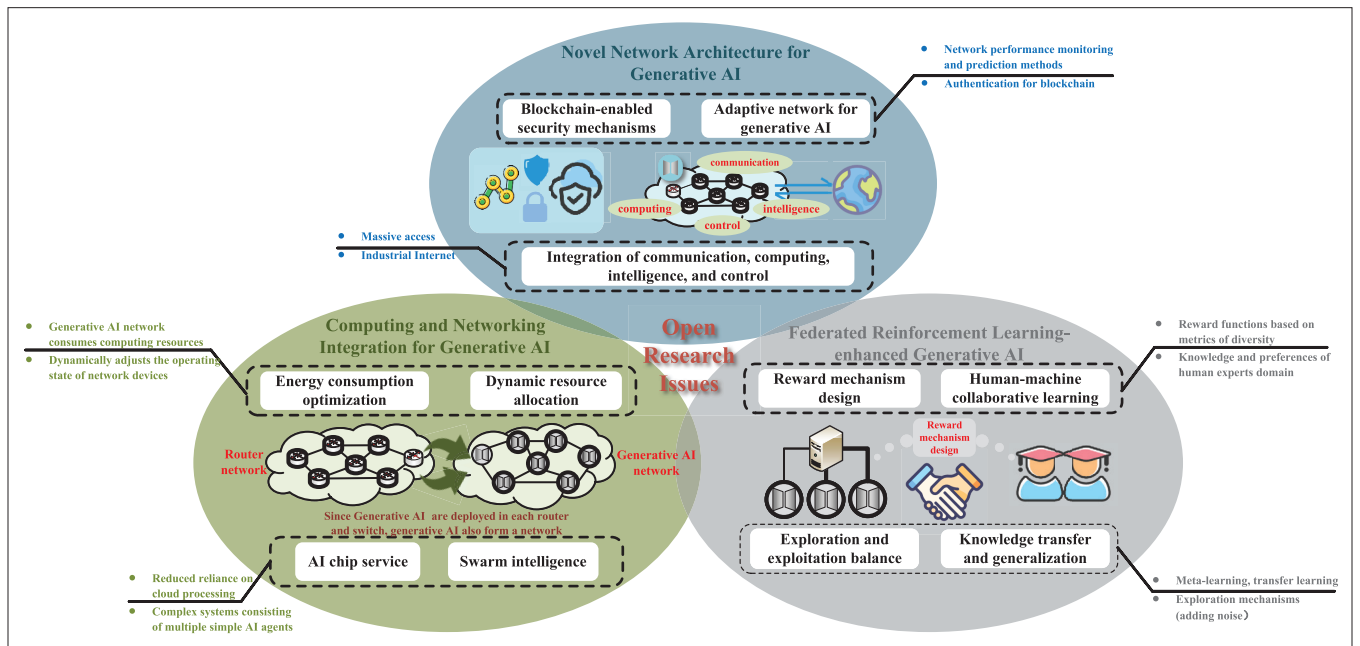


FIGURE 5. Open research issues of computing-efficient generative AI.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 62394321.

REFERENCES

- [1] N.-N. Dao et al., "A Review on New Technologies in 3GPP Standards for 5G Access and Beyond," *Comput. Networks*, vol. 245, 2024, p. 110370.
- [2] H. Du et al., "Diffusion-Based Reinforcement Learning for Edge-Enabled AI-Generated Content Services," *IEEE Trans. Mob. Comput.*, vol. 23, no. 9, 2024, pp. 8902–18.
- [3] H. Du et al., "AI-Generated Incentive Mechanism and Full-Duplex Semantic Communications for Information Sharing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 9, 2023, pp. 2981–97.
- [4] B. Du et al., "YOLO-Based Semantic Communication With Generative AI-Aided Resource Allocation for Digital Twins Construction," *IEEE Internet Things J.*, vol. 11, no. 5, 2024, pp. 7664–78.
- [5] W. Zhang et al., "Det(Com)2: Deterministic Communication and Computation Integration Toward AIGC Services," *IEEE Wireless Commun.*, vol. 31, no. 3, 2024, pp. 32–41.
- [6] D. Wu et al., "NetLLM: Adapting Large Language Models for Networking," *Proc. ACM SIGCOMM 2024 Conf.*, 2024, pp. 661–78.
- [7] J. Shao et al., "WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence," *arXiv preprint arXiv:2405.17053*, 2024.
- [8] X. He et al., "DeepScheduler: Enabling Flow-Aware Scheduling in Time-Sensitive Networking," *Proc. IEEE 2023 Conf. Comput. Commun.*, IEEE, 2023, pp. 1–10.
- [9] G. Peng et al., "Deterministic Cognition: Cross-Domain Flow Scheduling for Time-Sensitive Networks," *IEEE Trans. Cognit. Commun. Networking*, vol. 10, no. 4, 2024, pp. 1481–95.
- [10] Y. Huang et al., "AI-Generated Network Design: A Diffusion Model-Based Learning Approach," *IEEE Network*, vol. 38, no. 3, 2024, pp. 202–09.
- [11] L. Yang et al., "Generative Adversarial Learning for Intelligent Trust Management in 6G Wireless Networks," *IEEE Network*, vol. 36, no. 4, 2022, pp. 134–40.
- [12] H. Yu et al., "Deep Reinforcement Learning-Based Deterministic Routing and Scheduling for Mixed-Criticality Flows," *IEEE Trans. Ind. Inf.*, vol. 19, no. 8, 2022, pp. 8806–16.
- [13] Y.-C. Wang et al., "An Overview on Generative AI at Scale With Edge-Cloud Computing," *IEEE Open J. Commun. Soc.*, vol. 4, 2023, pp. 2952–71.
- [14] C. Liang et al., "Generative AI-Driven Semantic Communication Networks: Architecture, Technologies and Applications," *IEEE Trans. Cognit. Commun. Networking*, 2024, pp. 1–1.
- [15] Y. Fu et al., "Scalable Extraction Based Semantic Communication for 6G Wireless Networks," *IEEE Commun. Mag.*, vol. 62, no. 7, 2024, pp. 96–102.

BIOGRAPHIES

WEITING ZHANG [S'20, M'21] (wtzhang@bjtu.edu.cn) earned the Ph.D. degree in Communication and Information Systems from Beijing Jiaotong University, Beijing, China, in 2021. From November 2019 to November 2020, he was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Starting in December 2021, he worked as an associate professor at the School of Electronic and Information Engineering, Beijing Jiaotong University. His research interests include industrial Internet of Things, deterministic networks, edge intelligence, and machine learning for network optimization.

JIADONG REN (23125050@bjtu.edu.cn) is currently pursuing an M.S. degree at the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China. His research interests include industrial Internet of Things, deterministic networks, and machine learning for network optimization.

TAO ZHENG [S'09, M'15] (zhengtao@bjtu.edu.cn) received the B.S. degree in communications engineering and the Ph.D. degree in communication and information systems from Beijing Jiaotong University, Beijing, China, in 2006 and 2014, respectively. He is currently an associate professor at Beijing Jiaotong University. His specific areas of research interest mainly focus on wireless communication technology in industrial networks, the Internet of Things, and intelligent transportation systems.

RUIBIN GUO [S'21] (20111022@bjtu.edu.cn) received the B.E. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include deterministic networks, machine learning for resource allocation, and computing the first network.

HAN ZHANG (zhhan@tsinghua.edu.cn) received the B.E. degree in computer science and technology from Jilin University and the Ph.D. degree from Tsinghua University. He is an Assistant Professor in the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include computer networks, network security, and network system.

SHIWEN MAO [S'99, M'04, SM'09, F'19] (smao@ieee.org) received his Ph.D. in electrical and computer engineering from Polytechnic University, Brooklyn, NY. He is a Professor and Earle C. Williams Eminent Scholar, and Director of the Wireless Engineering Research and Education Center at Auburn University. His research interests include wireless networks and multimedia communications.

HONGKE ZHANG [M'13, SM'16, F'20] (hkzhang@bjtu.edu.cn) received the M.S. and Ph.D. degrees in Electrical and Communication Systems from the University of Electronic Science and Technology of China, Chengdu, China, in 1988 and 1992, respectively. From 1992 to 1994, he was a Postdoctoral Researcher with Beijing Jiaotong University, Beijing, China, where he is currently a Professor with the School of Electronic and Information Engineering and the Director of the National Engineering Research Center on Advanced Network Technologies. His research has resulted in many papers, books, patents, systems, and equipment in the areas of communications and computer networks.