The Age of Generative AI and AI-Generated Everything

Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Ping Zhang, Shuguang Cui, Xuemin Shen, Shiwen Mao, Zhu Han, Abbas Jamalipour, H. Vincent Poor, and Dong In Kim

Abstract

Generative AI (GAI) has emerged as a significant advancement in artificial intelligence, renowned for its language and image generation capabilities. This paper presents "AI-Generated Everything" (AIGX), a concept that extends GAI beyond mere content creation to real-time adaptation and control across diverse technological domains. In networking, AIGX collaborates closely with physical, data link, network, and application layers to enhance real-time network management that responds to various system and service settings as well as application and user requirements. Networks, in return, serve as crucial components in further AIGX capability optimization through the AIGX lifecycle, i.e., data collection, distributed pre-training, and rapid decision-making, thereby establishing a mutually enhancing interplay. Moreover, we offer an in-depth case study focused on power allocation to illustrate the interdependence between AIGX and networking systems. Through this exploration, the article analyzes the significant role of GAI for networking, clarifies the ways networks augment AIGX functionalities, and underscores the virtuous interactive cycle they form. It is hoped that this article will pave the way for subsequent future research aimed at fully unlocking the potential of GAI and networks.

Introduction

Within the evolving field of artificial intelligence (AI), the shift is evident from merely analyzing expansive datasets to actively generating innovative content. Milestones like AlphaGo's 2016 victory over a Go world champion laid the foundation, but Generative AI (GAI) represents a significant turn in AI's progression [1]. ChatGPT exemplifies this trend with its advanced conversational capabilities, resulting in more context-aware and detailed user interactions [2]. Similarly, DALL-E 3's ability to produce images from text descriptions blends linguistic comprehension with visual creation. Such GAI advancements underscore the expanding role of machines in domains once believed to be unique to human creativity. This growth in GAI promises to reshape our view of AI across academic, industrial, and societal spheres [3].

AI-driven networks have conventionally employed Discriminative AI (DAI) models, adept at tasks like data classification and prediction. While DAI excels at detecting existing patterns, GAI extends capabilities by creating new data samples, for instance, entirely fresh images or audio not present in original datasets. This introduces broadened applications in content creation, data augmentation, and even crafting network optimization strategies [1]. This capability transition positions GAI as a pivotal tool in network functions:

- **Data Synthesis and Augmentation:** Beyond AI's traditional data interpretation capabilities, GAI creates synthetic data vital for humans and networks. For example, GAI enhances anomaly detection systems by producing realistic simulations of infrequent network irregularities and cybersecurity threats, enriching training datasets, and improving system robustness [3].
- **Predictive Analysis and Management:** GAI extends beyond pattern recognition to forecast network behaviors and preemptively generate actions based on network conditions and anticipated future events [4]. This forward-looking functionality facilitates optimal allocation of network resources and aids in averting potential network bottlenecks [1].

Hongyang Du and Dusit Niyato are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798; Jiawen Kang (corresponding author) is with the School of Automation, the Key Laboratory of Intelligent Information Processing and System Integration of IoT, Ministry of Education, and the Guangdong-Hong Kong-Macao Joint Laboratory for Smart Discrete Manufacturing, Guangdong University of Technology, Guangzhou 510006, China; Zehui Xiong is with the Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372; Ping Zhang is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; Shuguang Cui is with the School of Science and Engineering (SSE), the Shenzhen Future Network of Intelligence Institute (FNii-Shenzhen), and the Guangdong Provincial Key Laboratory of Future Networks of Intelligence, The Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China; Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada; Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA; Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea; Abbas Jamalipour is with the School of Electrical and Computer Engineering, University of Sydney, Sydney, NSW 2006, Australia; H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA; Dong In Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea.

Digital Object Identifier: 10.1109/MNET.2024.3422241
Date of Current Version: 18 November 2024
Date of Publication: 2 July 2024

0890-8044/24©2024IEEE

- **Personalized User Interaction:** By leveraging GAI, networks can offer highly personalized services that adapt to individual user preferences [5]. This includes customizing content delivery, providing tailored recommendations, and developing specific user interfaces, offering a level of personalization that traditional DAI cannot match [6].

As we embrace these merits, merging GAI into networking and technical areas presents unique challenges and opportunities. Within this confluence of GAI and networks emerges the innovative notion of AI-Generated Everything (AIGX).

AIGX, progressing beyond AIGC's user-centric content creation, represents a paradigm that utilizes GAI models to optimize, refine, and devise applications and systems, enabling them to interact and adapt to instantaneous environmental shifts dynamically. Spanning from reshaping transportation and healthcare to innovating in urban planning and optimizing power grids, AIGX promises widespread transformation. Particularly in networking, a field ripe for AIGX-driven evolution, GAI's influence on every network component—from content delivery to architectural configurations—is pivotal [3]. AIGX enables dynamic adaptations to real-time conditions [1], deploys predictive insights for improved decision-making [4], and introduces resource allocation schemes that ensure optimal performance [7].

As shown in Fig. 1, while AIGX aims to revolutionize networks, the networks reciprocally play a significant role in optimizing AIGX. This symbiosis is evident throughout the AIGX lifecycle, including data collection, pre-training, fine-tuning, and inference. Specifically, the Internet-of-Things (IoT) is pivotal for efficiently collecting and preprocessing massive data streams from interconnected devices [8]. In the training phase, federated learning (FL) becomes useful, providing a decentralized AIGX model training paradigm that avoids centralized data hub limitations [9]. This approach facilitates localized data processing at the edge and enhances training and inference. Meanwhile, specialized offloading strategies ensure computational tasks are intelligently distributed from resource-constrained devices to more resourceful nodes or cloud infrastructures [5]. As shown in Fig. 2, recent advances in AIGX highlight its symbiotic evolution with intelligent networks. This paper

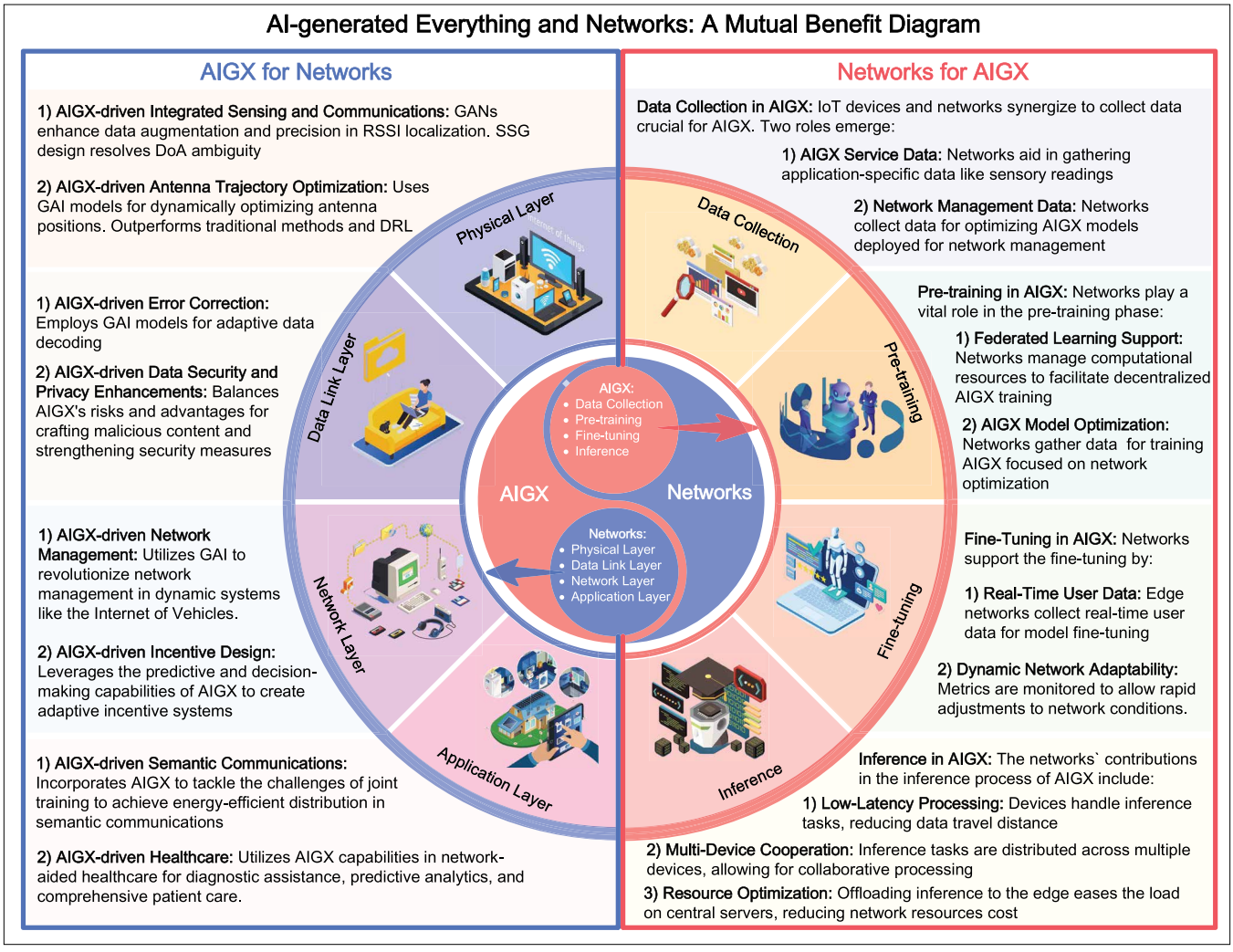

FIGURE 1. Symbiotic interaction between AIGX and data communications and networking. On the left, the diagram delineates the functionality of AIGX across various network layers: physical, data link, network, and application layers. On the right, it illustrates how networks contribute at distinct stages of AIGX's lifecycle, including data collection, pre-training, fine-tuning, and inference.

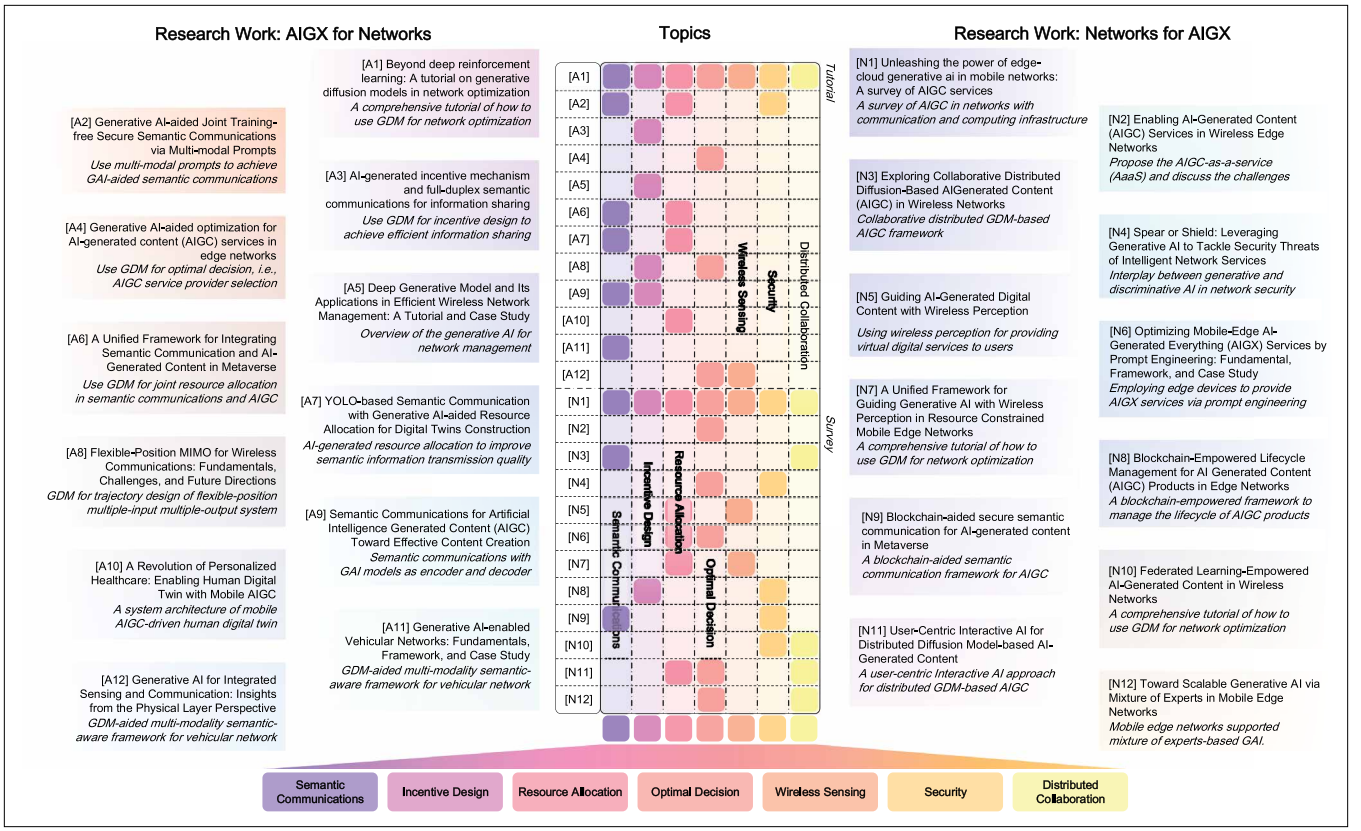

FIGURE 2. Summary of research topics and key contributions in *AIGC for networks and networks for AIGX*.

elucidates this collaboration and its consequences for computing and communication frameworks. Our main contributions are:

- *AIGX Enhancing Networks:* We discuss the impacts of AIGX across network layers and show how AIGX facilitates real-time modulation adjustments, elevates data security, and tailors adaptive resource management in various network settings.
- *Networks Enabling AIGX:* We spotlight the role of networks throughout the AIGX lifecycle. From enabling efficient data collection and distributed learning during pre-training stages to contributing to AIGX model optimization and ensuring low latency during the inference process, networks emerge as pivotal to deploying AIGX capabilities.
- *Virtuous Interactive Cycle:* Through a case study focused on power allocation, we examine the "virtuous interactive cycle" between AIGX and networks. This exploration exemplifies the symbiotic relationship between AIGX and intelligent networks in real-world scenarios and offers tangible guidelines for integrating AIGX techniques into network design.

Generative AI and AI-Generated Everything

In this section, we delve into GAI's foundational techniques and introduce the AIGX paradigm.

Generative AI: Core Techniques

GAI and DAI hold distinct methodologies and proficiencies. While DAI predominantly focuses on distinguishing between inputs by outlining class boundaries, GAI emphasizes generating content reminiscent of its training data.

1) Generative Adversarial Networks (GANs): GANs are pivotal models within GAI, expertly integrating data generation and discrimination capabilities. The model consists of two key components: a generator for data production and a discriminator to distinguish between original and generated data. Engaging in iterative training and competition, the generator crafts increasingly accurate data. In networking, GANs enable the simulation of complex network traffic patterns for cybersecurity, enriching datasets to train sophisticated intrusion detection systems.

2) Transformers: Initially developed for natural language processing, Transformers have broadened their impact, notably influencing the GAI. Their core advantage lies in attention mechanisms, which assign varied importance to different input sections, enabling effective parallel processing. In networking, Transformers provide real-time bottleneck prediction and optimize packet routing based on historical traffic trends. Models like BERT and ChatGPT, which excel in linguistic applications, find use in designing encoders and decoders for semantic communications (SemCom) [3].

3) Generative Diffusion Models (GDMs): GDMs represent a novel facet of GAI, gradually converting an initial random sample into a targeted output through multiple iterative denoising

Their core advantage lies in attention mechanisms, which assign varied importance to different input sections, enabling effective parallel processing.

Instead of just generating content, AIGX leverages GAI to adaptively design, fine-tune, and optimize applications and systems as they interact with real-time environmental changes.

steps. These models offer a departure from traditional neural architectures, presenting a new approach to content generation. In networking, GDMs have been applied in various network optimization tasks such as resource allocation, error correction coding, network economics, and SemCom [1].

4) Other GAI Techniques: *Autoregressive Models (ARMs)* leverage sequences of previous values to predict upcoming data points and are notably effective in generating sequential content such as text, audio, or video. ARMs can forecast network loads, optimize bandwidth allocation, and identify potential failure points using historical data in networks. *Variational Autoencoders (VAEs)* capture compressed data representations, generating new samples from this latent space, and are vital in creating synthetic network traffic patterns or enhancing multimedia content to boost user Quality of Experience (QoE) in content delivery frameworks. *Flow-based Models (FBMs)*, on the other hand, facilitate data generation by converting basic distributions into complex target ones through reversible transformations, ensuring content aligns with specified distributions, which is crucial for dynamic content stream adjustments and network security tasks where accurate traffic pattern replication is vital.

AI-Generated Everything (AIGX): The New Network Paradigm

1) Defining AIGX, An Evolution from AIGC: AIGC refers to the application of AI techniques to facilitate and automate the creation of content that is specifically tailored to user preferences and requirements. The 'C' in AIGC emphasizes the content-centric nature of this application, wherein the generated outputs are primarily forms of media or information such as text, images, videos, 3D models, and audio [5].

AIGX builds upon the foundational concepts of AIGC but extends its reach. 'X' in AIGX denotes 'Everything', representing its influence across all technological aspects. Instead of just generating content, AIGX leverages GAI to adaptively design, fine-tune, and optimize applications and systems as they interact with real-time environmental changes. For instance, in telecommunications, AIGX can dynamically provide network resource allocation schemes to maximize user QoE [7]. In manufacturing, AIGX could redefine automation by enabling machines to detect and proactively manage production anomalies.

2) Virtuous Interactive Cycle of AIGX and Networks: Embedding AIGX into network systems symbolizes transitioning from traditional static architectures to a dynamic, GAI-driven framework. Specifically, AIGX and networks have the following reciprocal benefits:

- **AIGX for Networks.** AIGX facilitates enhancements across network layers. In the *Physical Layer*, it enables real-time modulation adjustments, optimizing data rates and minimizing errors based on channel conditions [1]. In the *Data Link Layer*, AIGX dynamically augments error correction algorithms adapting to interference levels and enhances data security [3], [10]. In the *Network Layer*, it improves management in specialized systems, such as the Internet of Vehicles (IoV), and generates adaptive incentive mechanisms [7]. At the *Application Layer*, AIGX benefits the design of SemCom and healthcare systems [11]. By leveraging generative models and GAI-aided data processing, AIGX improves the efficacy and user-centricity of applications.
- **Networks for AIGX.** Network infrastructure is pivotal across all AIGX lifecycle stages. During data collection, networks enable the collection of both application-specific and network management data, fostering a feedback loop that boosts AIGX capabilities and network efficiency. In the pre-training phase, FL, supported by the network, facilitates decentralized model training, further optimizing AIGX models [9]. The fine-tuning stage leverages networks for real-time data collection and dynamic adaptability [9]. Lastly, networks employ edge offloading and multi-device cooperation during inference, reducing latency and optimizing resource allocation to enhance system throughput [5].

This interplay between AIGX and networks marks a paradigm shift, transforming networks into dynamic, evolving ecosystems that continuously adapt to immediate needs. Next, we delve into two aspects of the mutually beneficial relationship between AIGX and networks, followed by a case study to illustrate this virtuous interactive cycle and gains.

AIGX for Networks

This section studies AIGX methodologies that permeate and influence the network architecture's layers.

Physical Layer

1) AIGX-driven Integrated Sensing and Communications (ISAC): ISAC merges wireless sensing and communication to efficiently use constrained resources, with applications ranging from autonomous driving to gesture identification [12]. Integrating GAI with ISAC leads to novel applications and enhancements. In ISAC data processing, GAI models such as GANs significantly enhance system accuracy even with scarce real-world data, particularly evident in Received Signal Strength Indicator (RSSI) fingerprint localization. GAI adoption enhances dataset diversity, improving the precision of human activity detection. Furthermore, GAI addresses more intricate challenges, such as estimating the Direction of Arrival (DoA) of signals in ISAC systems in near-field or far-field scenarios. For instance, a GDM-developed signal spectrum generator (SSG) is proposed in *[A12]* of Fig. 2, as shown in Part B of Fig. 3. In a test with four antennas, the SSG used 10,000 paired signal spectrums, designated 80% for training and 20% for validation, and introduced noise into expert-generated solutions, followed by a stepwise denoising process.

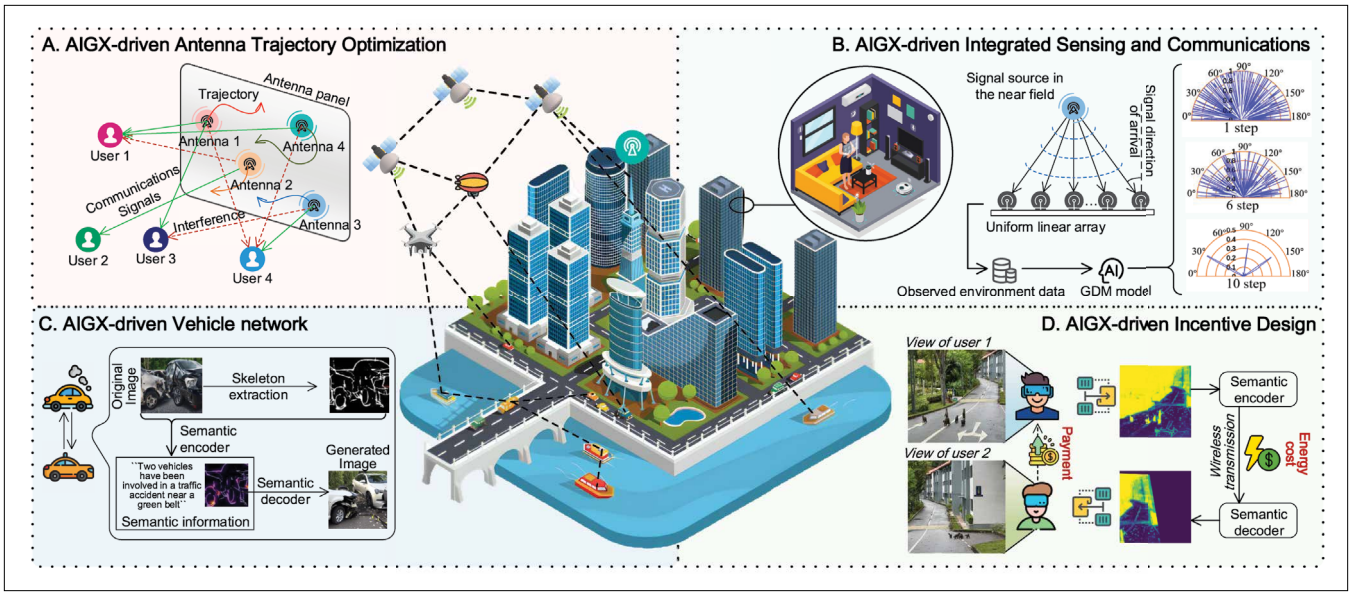

FIGURE 3. Typical system models of AIGX technologies in networks. *Part A* depicts the generation of optimal antenna trajectories using AIGX. *Part B* demonstrates the localization results produced by AIGX, leveraging perceived wireless environments. *Part C*, the application of AIGX technology for designing encoders and decoders in vehicular network semantic communication systems is presented. *Part D* shows the incentive mechanism generated by AIGX in mixed-reality user information-sharing systems.

The evaluation revealed that the SSG output converges with expert solutions during training and achieves a test loss of −10, outperforming the −80 loss in deep reinforcement learning (DRL)-based methods.

2) AIGX-driven Antenna Trajectory Optimization: With their repositionable antennas, flexible-position MIMO systems improve wireless communications by optimizing channel conditions and boosting spectral and energy efficiencies. The key challenge is trajectory optimization for maximizing spectral (SE) or energy efficiency (EE). Traditional optimization techniques can get stuck in complex scenarios, often settling in local optima. AIGX employs GAI-driven optimization under dynamic network conditions [1]. In the case study in *[A8]* of Fig. 2, a GDM aimed to enhance SE is trained to generate antenna trajectories, as shown in Part A of Fig. 3. Compared to DRL methods, which quickly peak and then level out, AIGX optimization steadily increases rewards, raising the sum SE from an average of 11.7 (using DRL) to 13.3. Results revealed that the generated solution enabled antennas to either adjust positions to enhance user coverage or move to the system's periphery to mitigate multi-user interference to optimize SE, while EE-prioritized scenarios prompted antennas to shift towards areas with higher user density.

Data Link Layer

1) AIGX-driven Error Correction: Error correction is significant for ensuring data integrity across interference-prone channels in the data link layer. Identifying the codeword most suitable with the received signal is traditionally considered an NP-hard challenge, implying a potential exponential search for optimal decoding. Fortunately, GAI models in the AIGX framework, like Transformer-based decoders, have brought efficiency to this task, enhancing accuracy and computational speed [10]. However, the invariant computational load is a persistent challenge, independent of the codeword corruption degree. The introduction of the GDM within the AIGX framework presents a solution, enabling iterative decoding that adjusts according to the varying degrees of codeword corruption, significantly reducing computational demands [10]. A case study in [10] highlights AIGX's effectiveness. It presents data transmission as an iterative forward diffusion process that requires inversion at the receiving end. The Bit Error Rate (BER) of the GDM method is notably lower than traditional schemes, with it being only 11% of a specific Transformer scheme when the signal-to-noise ratio is at 4 dB.

2) AIGX-driven Data Security and Privacy Enhancements: The surge in digital communication has heightened the importance of effective security and privacy measures. While AIGX introduces novel ways to create content, it also opens doors to potential risks. We discuss the attack scenarios and defense mechanisms as:

- **Attack Scenarios:** Foundation models, such as OpenAI's ChatGPT, present emerging threats by potentially generating harmful content that bypasses safeguards of network service providers, as discussed in *[N4]* of Fig. 2. Additionally, AIGX might exploit subtle similarities between encrypted and plain images in the embedding space, compromising the trustworthiness of prevailing encryption methods.
- **Defense Mechanisms:** LLMs can enhance the training dataset of a DAI model by creating adversarial examples. Take a network intrusion detection system (NIDS) trained to distinguish between 'normal' and 'malicious' network traffic. If a malicious pattern, like several failed logins followed by a successful one, is detectable, the LLM might generate an adversarial sample

that spreads the login attempts over various user accounts or IPs. This disrupts the recognizable pattern and might go unnoticed.

An illustrative example of the synergy between AIGX's advantages and associated risks is evident in wireless image transmission. In this setting, discriminative AI attempts to disrupt communication by initiating data poisoning attacks on an image dataset hosted on a server. Counteracting this, AIGX-driven defenses utilize a GDM to authenticate each image before transmission. As shown in *[N4]* of Fig. 2, utilizing this AIGX-driven defense led to an 8.7% reduction in energy consumption.

Network Layer

1) AIGX-driven Network Management: Network management is crucial for modern systems, ensuring efficient communication and performance, especially in extensive networks with data flow across many devices [13]. AIGX, with GAI's capability to discern and emulate intricate data patterns, enhances troubleshooting, predictive maintenance, and data synthesis [1]. Consider the IoV, where vehicles continuously exchange data. The dynamism of this system requires adaptable network management. AIGX, given its data representation and generation prowess, can predict network congestion, allocate bandwidth adaptively, and even fill in data gaps where actual data is lacking. As shown in Part C of Fig. 3, a specialized study, i.e., *[A11]* of Fig. 2, addresses V2V resource allocation, formulating a QoE metric grounded in transmission rate and received image fidelity. Comparative analyses reveal that AIGX-based approaches outperform DRL counterparts, recording an 18.5% increase in average QoE.

2) AIGX-driven Incentive Design: Incentives play a key role in prompting network participants to share resources, boosting network efficiency. Without the right rewards, users might hold back, considering costs like battery usage or bandwidth. Unlike the conventional DAI-based method, AIGX leverages the learning capabilities of GAI models to dynamically adapt to evolving network conditions and user behaviors. This adaptability achieves more intelligent and responsive network management, surpassing the performance of conventional DRL methods. As shown in Part D of Fig. 3. a noteworthy application involves deploying AIGC within Mixed-Reality (MR) technologies [7]. MR headset-mounted devices (HMDs) often face constraints in computational power, impacting user experience. An effective information-sharing strategy, leveraging full-duplex device-to-device (D2D) SemCom, emerges as a solution [7]. Instead of each user doing repetitive computational tasks like generating AIGC, the system facilitates sharing this content with relevant semantic information to nearby users. Such a framework calls for an effective incentive mechanism to motivate users. A contract theory-based incentive model is proposed, utilizing an AIGX-based approach, notably the diffusion model, to refine the design of contracts, which boosts user QoE by 11.7% over the traditional DRL approach [7].

Application Layer

1) AIGX-driven Semantic Communications: Semantic Communications (SemCom) addresses the challenge of exponential data volume growth in wireless communication networks by converting messages into semantic information for transmission using a semantic encoder and decoder. However, challenges exist in joint training and energy-efficient distribution of these AI-based encoders and decoders. Fortunately, GAI models, such as advanced language and image generation models, can reconstruct complex messages from simpler semantic representations, alleviating the need for joint training [14]. For example, using multi-modal prompts, i.e., visual and textual prompts, can lead to accurate semantic decodings [11] to solve the problem of instability brought by the diverse generative capabilities of GAI models, which can be used in scenarios that require accurate information transfer such as human face images. Another example is integrating SemCom and AIGC (ISGC) to enhance user immersion as discussed in *[A6]* of Fig. 2. ISGC balances computing and communication resources for semantic extraction, AIGC inference, and graphic rendering. An effective resource allocation mechanism can be achieved with the help of AIGX to obtain near-optimal strategies. Numerical results show that the GDM-based method can improve the QoE by 8.3% compared with the Proximal Policy Optimization (PPO) method.

2) AIGX-driven Healthcare: Network technologies, aided by IoT, are transforming healthcare, allowing real-time patient monitoring and efficient data sharing. AIGX further elevates network-based healthcare by simulating human thought and analyzing large datasets, improving diagnostics, predictions, and overall care, as discussed in *[A10]* of Fig. 2. It can foresee health concerns by reviewing vast data and devising personalized treatment plans using a patient's history and present health data. In AIGX-driven health applications like virtual physical (VR) therapy, maintaining high-quality VR video streams is essential. Although SemCom helps reduce data, preserving VR quality is a challenge, especially with potential inaccuracies. AIGX can recreate these streams closely to the original. Yet, managing the network's efficiency, VR video quality, and genuineness can be challenging. An optimization design is presented in *[A10]* of Fig. 2 considering constraints like bandwidth, computational capacity, and QoE benchmarks, aiming to enhance user QoE by focusing on aspects including resolution ratio and the diffusion step that is crucial for GDMs. Compared to the Soft Actor-Critic (SAC) algorithm, a conditional GDM-based approach has increased the total users' QoE by 32.4%.

3) AIGX-driven User-centric Networks: User-centric networks harness the power of AIGX to revolutionize how network services cater to individual user needs. Central to this transformation are the foundation models, e.g., LLMs, enabling a deeper understanding of user intents, preferences, and behaviors. These advanced foundation models, through interactive AI (IAI) mechanisms, facilitate a dynamic, personalized interaction layer

between the network and users. This level of personalization is achieved by integrating user-centric design principles, where feedback loops, powered by LLMs, continuously adapt to user interactions. For example, for distributed AIGC studied in *[N11]* of Fig. 2, the authors address the challenges of enhancing both the subjective QoE and energy efficiency in services utilizing GDM for image creation. Crucially, the approach incorporates a Reinforcement Learning With Large Language Models Interaction (RLLI) strategy, leveraging the power of LLMs to employ generative agents to mimic user interactions. This enables the provision of instantaneous and personalized QoE feedback, capturing the nuances of individual user personalities. Through LLMs, the system dynamically adapts to user feedback, continually refining the content generation process to align with user preferences and enhance the overall experience.

Lesson Learned

In leveraging AIGX methodologies for network systems, several key lessons emerge. The primary motivation behind leveraging GAI is its ability to amplify accuracy, optimize network management schemes, and predict intricate data patterns, especially in dynamic systems like the IoV. Conventional DRL-based methods, prone to getting stuck in local optima, can struggle with real-time changes. Using tools such as GANs and GDMs, GAI provides flexible solutions that adapt to evolving conditions. This flexibility is crucial when navigating vast, dynamic systems or addressing challenges like error correction. Furthermore, while AIGX's introduction brings forth data security opportunities, it demands careful implementation. In practical applications, areas like healthcare are transformed by AIGX-driven networks, facilitating real-time tracking, individualized treatments, and enhanced diagnostic capabilities.

Network for AIGX

Networks play a pivotal role in the AIGX lifecycle, from data collection and training to fine-tuning and inference.

Data Collection

Data collection is fundamental to the AIGX ecosystem, impacting the efficiency and reliability of AIGX applications. Utilizing varied techniques enriched by IoT capabilities, integrated devices, and sensors are crucial channels for detailed environmental data collection. In smart cities, for instance, IoT devices acquire data on air quality, vehicular movement, and energy consumption, and use these data for AIGX applications. Networks play two synergistic roles in AIGX data collection:

- **AIGX Service Data:** As shown in Part A. 1 of Fig. 4, networks enable the acquisition of application-specific data, such as sensory and visual information from IoT cameras, which form the empirical foundation for developing and refining AIGX service models.
- **Network Management Data:** Beyond service-specific needs, networks collect data useful for fine-tuning and optimizing AIGX models deployed for network management. As shown in Part A. 2 of Fig. 4, expert decisions under various channel conditions can be collected for further AIGX model training.

This dynamic fosters a virtuous interactive cycle: AIGX models optimized for network performance enhance data rates and network efficiency,

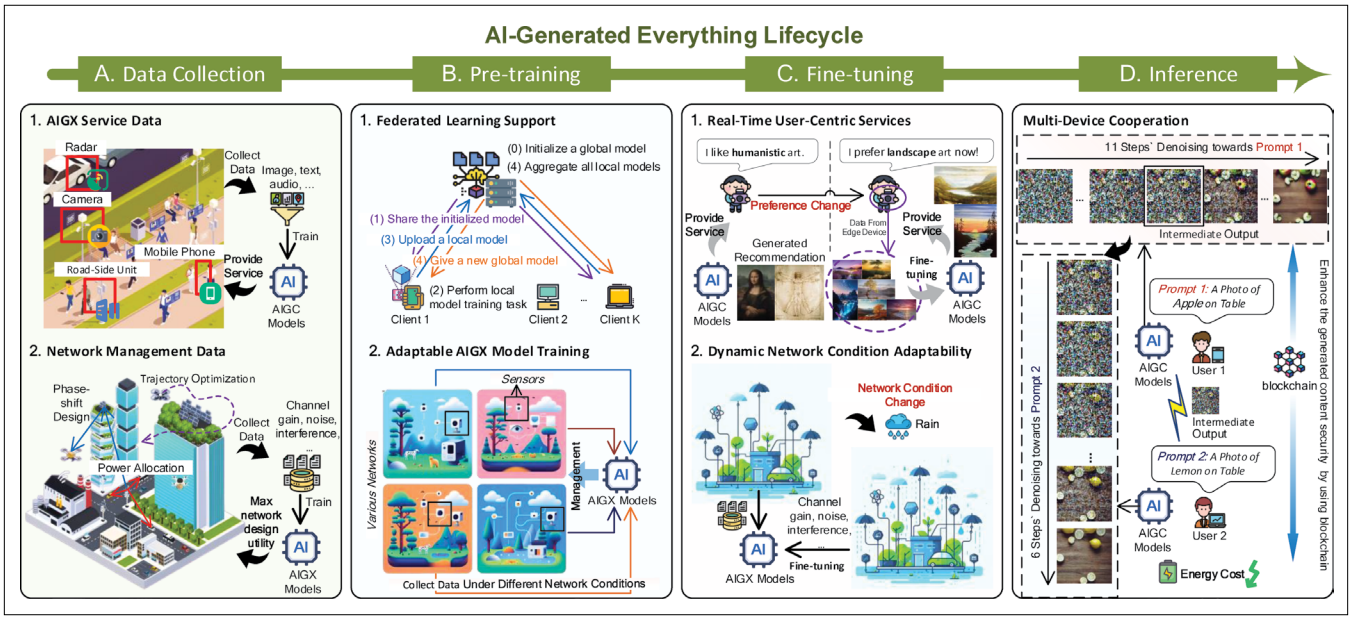

FIGURE 4. The role of networks across various stages of the AIGX lifecycle. In the *Data Collection* phase, networks enable efficient data gathering essential for training effective AIGX models that cater to user needs or manage the network. During the *Pre-training* stage, networks support AIGX models in flexible training, such as FL supporting distributed training. In the *Fine-tuning* stage, network devices update pre-trained models with new user data for personalized services and adapt decision-making models to emerging network conditions. Collaboration among network devices at the *Inference* stage promotes more adaptable and energy-efficient inference models.

which in turn improve data collection for AIGX model refinement, perpetually fine-tuning both AIGX capabilities and network efficacy.

Pre-Training

The pre-training phase in the AIGX lifecycle, crucial for developing foundational models across various applications, significantly leverages network infrastructure capabilities:

- **Federated Learning Support:** Networks employ FL, allowing decentralized model training while keeping data at the edge [9] as shown in Part B. 1 of Fig. 4 While FL enhances data privacy, effective network resource management must address computational challenges. Networks manage resources, including dynamic bandwidth allocation and low-latency communication, ensuring efficient localized AIGX model training across devices without overloading the network [9]. An example of this approach in action is FATE-LLM,[1] a federated learning framework tailored for large language models, designed for industrial applications.
- **Adaptable AIGX Model Training:** Networks, by collecting diverse environmental and network-related data such as latency and bandwidth utilization, enrich AIGX model training input as shown in Part B. 2 of Fig. 4, focusing on network optimization. For instance, if a pre-trained AIGX model incorporates high-bandwidth environment data, its performance may be suboptimal under bandwidth constraints.

This symbiotic relationship ultimately achieves improvements in both AIGX functionalities and network performance.

Fine-Tuning

In contrast to the generalizations of the pre-training stage, fine-tuning zeroes in on swift optimization for user preferences and fluctuating network conditions. The network bolsters AIGX fine-tuning as follows:

- **Real-Time User-Centric Services:** Leveraging the capabilities of edge networks, as shown in Part C. 1 of Fig. 4 real-time data reflecting user preferences is collected for model fine-tuning [9]. This makes AIGX models agile in adapting to shifts in user behavior or network dynamics.
- **Dynamic Network Condition Adaptability:** For network management-focused AIGX models, the network continuously monitors metrics like fadings and packet loss, feeding this data into the fine-tuning process to swiftly adapt to network changes as shown in Part C. 2 of Fig. 4

Consequently, the network functions as an effective platform, fostering the rapid fine-tuning of AIGX models, and making them practical for real-world scenarios.

Inference

The inference stage deploys trained AIGX models for specific needs. Rather than depending on traditional centralized servers, which often result in bottlenecks and delays, AIGX can leverage edge-based offloading to enhance the efficiency [5]. The network's role in AIGX inference includes:

- **Low-Latency Processing:** Offloading inference tasks to edge devices significantly minimizes latency. This is achieved by reducing the round-trip data travel distance between the user and the processing unit. For instance, Qualcomm reported a nine-fold increase in speed compared to the baseline Stable Diffusion model when executed on a phone equipped with Snapdragon 8 Gen 3.[2]
- **Multi-Device Cooperation:** AIGX inference can be partitioned and executed across multiple cooperating edge devices as shown in Part D of Fig. 4. This distributed method removes individual device limits, including computation and energy resources, and boosts the overall system effectiveness through collaborative processing.
- **Optimized Resource Allocation:** Distributing the inference process across edge devices alleviates central server workloads, optimizing the utilization of network resources and averting potential congestion points.

With edge offloading and multi-device cooperation, the network sharply reduces latency, enhances system throughput, and optimizes resource allocation.

Lesson Learned

Networks are integral to the AIGX ecosystem, providing the infrastructure and data pathways that support AIGX models. By enabling decentralized training, networks ensure data privacy while enhancing model adaptability, essential in ever-changing real-world scenarios. Networks are pivotal in gathering diverse data, which strengthens AIGX models, and are indispensable for ensuring real-time adaptability to user preferences and shifting network conditions. Furthermore, with the trend towards edge-based offloading, networks optimize resource allocation and reduce latency, ensuring that AIGX models are both efficient and practical in real-world applications.

Case Study: The Virtuous Interactive Cycle of AIGX and Networks in Power Allocation

Managing efficient communication between a base station and a user across multiple channels, as depicted in Part B of Fig. 5, is a representative challenge in modern wireless systems. While conventional techniques like water filling are precise, they are resource-intensive and need constant adjustments for each set of channel gains. On the other hand, AIGX offers an adaptive, real-time solution suitable for fluctuating channel conditions. The case study,[3] with different environmental stages T_1, T_2, and T_3 switching between AIGX model training and channel data collection, shows AIGX's adaptability in optimizing power allocation based on real-time channel feedback. This adaptability suggests AIGX's broader applicability, which could be the key to enhancing a range of wireless communication problems where environmental conditions and data patterns change rapidly, making this case

[1] https://github.com/FederatedAI/FATE-LLM

[2] https://www.qualcomm.com/news/onq/2023/11/accelerating-generative-aiat-the-edge

[3] The code is available on https://github.com/HongyangDu/VirtuousAIGX.

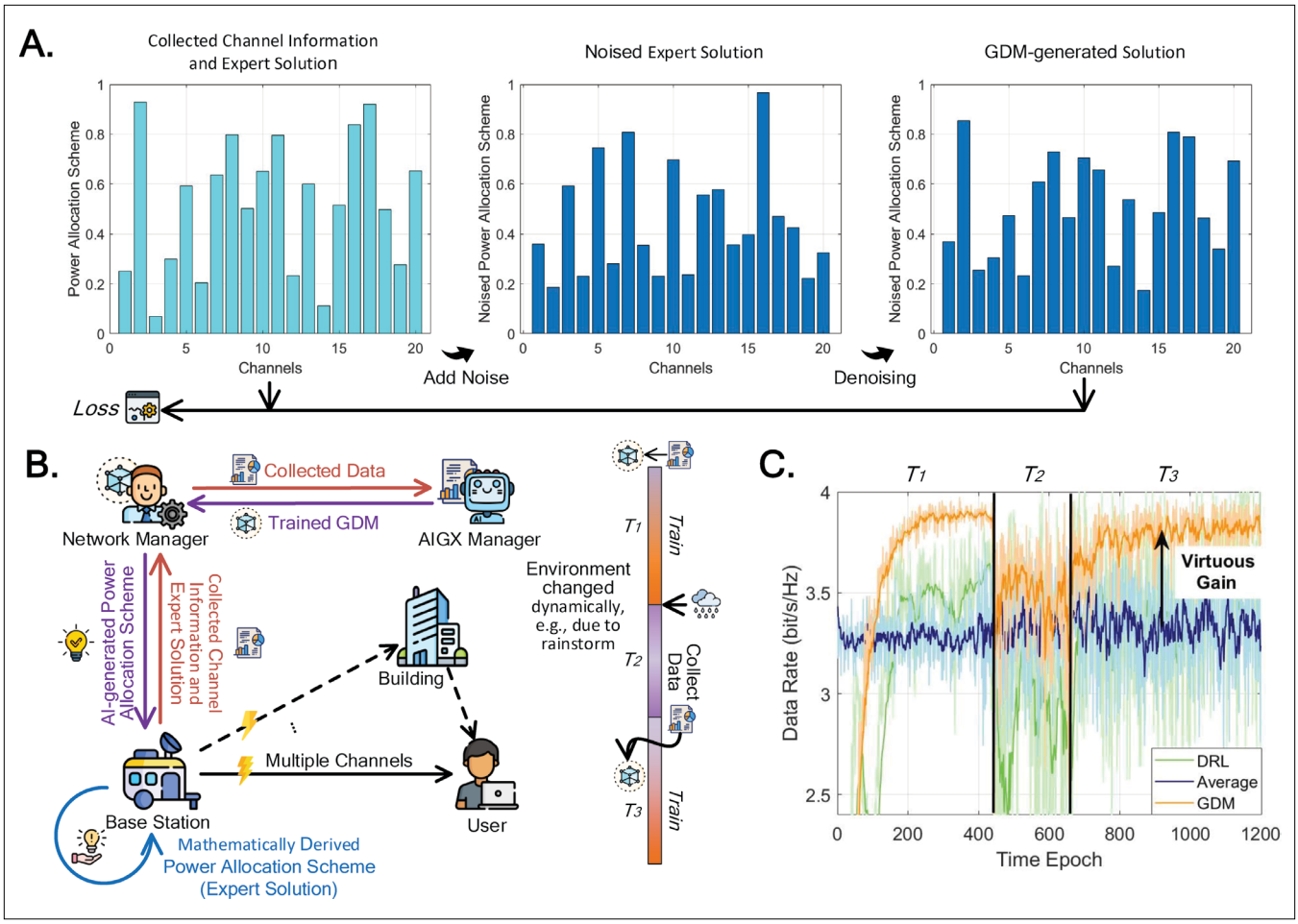

FIGURE 5. Case study of the AIGX-network virtuous interactive cycle: *Part B* shows the system model where a base station and user communicate through multiple channels, demanding power allocation to optimize the user's sum rate. Networks feed training data to the AIGX Manager, which generates a powe allocation model. *Part A* shows the training process of the GDM, generating optimal decisions under given channel conditions by adding noise to the expert solution and denoising it. *Part C* reveals the AIGX-network cycle's adaptation to network changes, illustrating how improved sum rates benefit the network and provide varied data. The AIGX model performance is tested by randomly sampling environmental conditions during training

study pivotal in demonstrating the impact of AIGX on network optimization.

Initial Conditions and AIGX Training (T_1)

During T_1, the network operates under stable conditions, analogous to good weather conditions. With $M = 20$, channel gains are set between 5 and 8 for the initial 10 channels and 3 to 6 for the subsequent 10 to simulate a range of channel conditions. These gains, serving as conditions, are collected in training the GDM model to allocate power by identifying and adapting to channel-specific states to maximize the data rate. Specifically, the GDM model is trained through a process where optimal decision generation is achieved by introducing noise to an expert solution and denoising it, as demonstrated in *Part A*. As shown in Part C of Fig. 5. by the close of T_1 model training, the AIGX model enhances the network data rate by 18.8% compared with the average allocation scheme.

Environmental Variations and Their Impact (T_2)

In phase T_2 channel data collection, the network grapples with varied environmental changes, for instance, a rainstorm, which is particularly challenging for wireless communications due to the scattering and absorption of radio signals at high-frequency bands, e.g., mmWave. Specifically, we consider that channel gains fluctuate capriciously between 1 and 7 for all 20 channels, reflecting real-world scenarios where meteorological shifts evoke unpredictable network behavior. Even though the AIGX model trained under T_1 conditions manifests the robustness, it becomes suboptimal in T_2, evidenced by a conspicuous decline in the data rate. Such a situation motivates a virtuous interactive cycle to gain additional expert solutions to enhance the AIGX model training under diverse conditions.

The Virtuous Interactive Cycle in Action (T_3)

In phase T_3, the synergistic virtuous interactive cycle between AIGX and the network becomes evident. The network, under T_2's rain-impacted conditions, actively acquires new channel gains and corresponding expert solutions, serving to retrain the AIGX model. This adaptation of the model's power allocation strategy to the new channel conditions triggers a significant

These gains, serving as conditions, are collected in training the GDM model to allocate power by identifying and adapting to channel-specific states to maximize the data rate.

improvement in the data rate. Specifically, the virtuous gain, defined as the enhancement in data rate achieved by the retrained AIGX model compared to a conventional average method, is 15.1%. Conversely, a deep reinforcement learning-based method. i.e., SAC [1], cannot achieve performance analogous to AIGX, attaining merely an 8.9% virtuous gain even under the virtuous interactive cycle.

Lesson Learned

The virtuous gain illuminates the importance of the AIGX-network virtuous interactive cycle, emphasizing the essential role of the network in the AIGX-network cycle. This stresses the importance of ongoing data acquisition and feedback mechanisms in maintaining the relevance and adaptability of AIGX algorithms under different network conditions. Rather than just benefiting from AIGX, networks actively contribute to AIGX adaptability and fortifying decision-making efficacy. In addressing future challenges, we identify two major limitations of our case study: the generalizability of the AIGX model to new network conditions and its dependency on expert-derived training solutions. These factors highlight the complexity of real-world network environments, pointing out the crucial need for AIGX solutions that are resilient and adaptable. In this context, RL offers a promising way to enhance the flexibility and strength of AIGX systems [1].

Future Directions

AIGX Enhancing Networks

AIGX could serve as a foundation for automating complex network management tasks, streamlining data flow, and enhancing security. We discuss some representative future research directions:

- **AIGX-Driven Near-Field Communications:** The unique challenges in near-field MIMO communications, especially complex antenna dependencies [15], are mitigated by AIGX by leveraging its capability to model and adapt to the multifaceted channel dynamics. AIGX's learning algorithms adapt to the complex channel states, facilitating optimal signal processing and resource allocation.
- **Aigx-Driven Space-Air-Ground Integrated Network (SAGIN):** AIGX addresses the need for robust communication across different atmospheric layers in SAGIN. By crafting adaptive pathways for data flow among satellite, aerial, and terrestrial layers, AIGX's adaptability meets the network's diverse demands, resulting in stable data transmission.
- **AIGX-Driven Multimodal Communications:** Networks managing a spectrum of data types, from text and images to video, benefit from the integration of AIGX agents. These agents are the key to synchronizing and prioritizing multimodal data flows efficiently. For example, platforms like SORA[4] by OpenAI demonstrate the effectiveness of AIGX in orchestrating cross-modal content, substantially enhancing user interaction and experience.

Networks Supporting AIGX

Networks empower AIGX's functionalities by providing robust data transportation, enabling rapid model deployment, and facilitating edge computing capabilities. We identify the following avenues for future research:

- **Swarm Intelligence for Collaborative AIGX Service:** Networks enabled with swarm intelligence can distribute AIGX services across nodes more efficiently. For instance, swarm algorithms could distribute GAI tasks across network nodes and exchange information efficiently, thus accelerating model training and inference.
- **Transfer Learning Across Network Hubs:** Network-supported transfer learning can enhance AIGX efficiency by distributing pretrained models across multiple hubs. This allows each node to leverage shared learning and insights, reducing the need for AIGX services to train new models from scratch.
- **Energy-Efficient Networking for Sustainable AIGX:** To address the high energy demands of AIGX models like ChatGPT, networks can implement energy-aware routing algorithms and task distribution strategies. Data-intensive AIGX tasks may be routed through network nodes utilizing renewable energy sources, and low-energy hardware accelerators can be engaged for specific computations. In this context, the development and management of AI Data Centers (AIDCs) play a pivotal role. AIDCs, designed to support AIGX operations, must prioritize energy efficiency through advanced cooling systems, optimized server utilization, and the adoption of green energy solutions to align with sustainable AIGX practices.

Conclusion

In this article, we have explored AIGX's role in intelligent networks. Incorporated across different network layers, AIGX promotes adaptive responses and enhances network management. Networks, in return, are instrumental in the AIGX lifecycle, from data collection to reducing inference latency. A case study on AIGX-driven power allocation highlighted the "virtuous interactive cycle" between AIGX and networks, emphasizing their reciprocal benefits. While our findings illuminate the promising relationship between AIGX and networks, continued research is essential to develop new methodologies and address emerging challenges, aiming for joint advancement of AIGX and intelligent networks.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62102099, Grant U22A2054, and Grant 62293482; in part by the Guangzhou Basic Research Program under Grant 2023A04J1699; in part by the National Research Foundation (NRF), Singapore and Infocomm Media Development Authority under the Future Communications

[4] https://openai.com/sora

Research Development Programme (FCP), and DSO Nationa Laboratories under the AI Singapore Programme (AISG), under Energy Research Test-Bed and Industry Partnership Funding Initiative, part of the Energy Grid (EG) 2.0 programme, under DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) programme, Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), under Award AISG2RP-2020-019; in part by the Ministry of Education, Singapore, under its SMU-SUTD Joint under Grant 22-SIS-SMU-048; in part by the SUTD Kickstarter Initiative under Grant SKI 20210204; in part by the MSIT (Ministry of Science and ICT), South Korea, under the ICT Creative Consilience Program Supervised by the IITP (Institute for ICT Planning & Evaluation) under Grant IITP-2020-0-01821; in part by the NSF under Grant CNS-2148382, Grant CNS-2107216, Grant CNS-2128368, Grant CMMI-2222810, and Grant ECCS-2302469; in part by the Basic Research Project of Hetao Shenzhen-HK S&T Cooperation Zone under Grant HZQB-KCZYZ2021067; in part by the Shenzhen Outstanding Talents Training Fund under Grant 202002; in part by the Guangdong Research under Project 2017ZT07X152 and Project 2019CX01X104; in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001; in part by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence under Grant ZDSYS201707251409055; in part by the U.S. National Science Foundation under Grant ECCS-2335876; in part by the U.S. Department of Transportation, Toyota; and in part by the Amazon and Japan Science and Technology Agency (JST) Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) under Grant JPMJAP2326.

References

[1] H. Du et al., "Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization," *IEEE Commun. Surveys Tuts.*, early access, May 10, 2024, doi: 10.1109/COMST.2024.3400011.
[2] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," *J. AI*, vol. 7, no. 1, pp. 52–62, Dec. 2023.
[3] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1127–1170, 2nd Quart., 2024.
[4] H. Du et al., "Diffusion-based reinforcement learning for edge-enabled AI-generated content services," *IEEE Trans. Mobile Comput.*, early access, Jan. 19, 2024, doi: 10.1109/TMC.2024.3356178.
[5] H. Du et al., "Exploring collaborative distributed diffusion-based AI-generated content (AIGC) in wireless networks," *IEEE Netw.*, vol. 38, no. 3, pp. 178–186, May 2024.
[6] J. Wang et al., "Generative AI for integrated sensing and communication: Insights from the physical layer perspective," 2023, *arXiv:2310.01036*.
[7] 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, pp. 2981–2997, Sep. 2023.
[8] H. Xiao, C. Xu, Y. Ma, S. Yang, L. Zhong, and G.-M. Muntean, "Edge intelligence: A computational task offloading scheme for dependent IoT application," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7222–7237, Sep. 2022.
[9] X. Huang et al., "Federated learning-empowered AI-generated content in wireless networks," *IEEE Netw.*, early access, Jan. 12, 2024, doi: 10.1109/MNET.2024.3353377.
[10] Y. Choukroun and L. Wolf, "Denoising diffusion error correction codes," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–18.
[11] H. Du et al., "Generative AI-aided joint training-free secure semantic communications via multi-modal prompts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 12896–12900.
[12] A. Liu et al., "A survey on fundamental limits of integrated sensing and communication," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 994–1034, 2nd Quart., 2022.
[13] L. Song et al., *Aerial Access Networks. Integration of UAVs, HAPs, and Satellites*. Cambridge, U.K.: Cambridge Univ. Press, 2023.
[14] T. Han et al., "Generative model based highly efficient semantic communication approach for image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[15] Y. Liu et al., "Nearfield communications: A tutorial review," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1999–2049, 2023.

Biographies

Hongyang Du (hongyang001@e.ntu.edu.sg) received the B.Eng. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, in 2021, and the Ph.D. degree from the Interdisciplinary Graduate Program, College of Computing and Data Science, Energy Research Institute, Nanyang Technological University (NTU), Singapore, in 2024. His research interests include edge intelligence, generative AI, semantic communications, and network management.

Dusit Niyato (Fellow, IEEE) (dniyato@ntu.edu.sg) received the B.Eng. degree from the King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada, in 2008. He is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include the Internet of Things (IoT), machine learning, and incentive mechanism design.

Jiawen Kang (kavinkang@gdut.edu.cn) received the Ph.D. degree from the Guangdong University of Technology, China, in 2018. He was a Post-Doctoral Researcher at Nanyang Technological University, Singapore, from 2018 to 2021. He currently is a Full Professor with the Guangdong University of Technology, China. His research interests include blockchain, security, and privacy protection in wireless communications and networking.

Zehui Xiong (zehui_xiong@sutd.edu.sg) received the B.Eng. degree in telecommunications engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, and the Ph.D. degree in computer science and engineering from Nanyang Technological University (NTU), Singapore. He is currently an Assistant Professor with the Singapore University of Technology and Design (SUTD), and also an Honorary Adjunct Senior Research Scientist with the Alibaba-NTU Singapore Joint Research Institute, Singapore. His research interests include wireless communications, the Internet of Things, blockchain, edge intelligence, and metaverse.

Ping Zhang (Fellow, IEEE) (pzhang@bupt.edu.cn) received the M.S. degree in electrical engineering from Northwestern Polytechnical University, Xi'an, China, in 1986, and the Ph.D. degree in electric circuits and systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1990. He is currently a Professor with BUPT. His research interests include cognitive wireless networks, fifth-generation mobile networks, communications factory test instruments, universal wireless signal detection instruments, and mobile Internet. He was a recipient of the First and Second Prizes from the National Technology Invention and Technological Progress Awards, and the First Prize Outstanding Achievement Award of Scientific Research in College.

Shuguang Cui (Fellow, IEEE) (shuguangcui@cuhk.edu.cn) received the Ph.D. degree in electrical engineering from Stanford University, CA, USA, in 2005. Afterwards, he has been working as an Assistant Professor, an Associate Professor, a Full Professor, and a Chair Professor in electrical and computer engineering with the University of Arizona, Texas A&M University, UC Davis, and CUHK, Shenzhen, respectively. He worked as the Executive Dean of the School of Science and Engineering, CUHK, the Executive Vice Director of the Shenzhen Research Institute of Big Data, and the Director of the Future Network of Intelligence Institute (FNii). His current

research interests include merging between AI and communication networks.

Xuemin (Sherman) Shen (Fellow, IEEE) (sshen@uwaterloo.ca) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include network resource management, wireless network security, the Internet of Things, AI for networks, and vehicular networks. He is a Registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, an International Fellow of the Engineering Academy of Japan, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Shiwen Mao (Fellow, IEEE) (smao@ieee.org) is currently a Professor and Earle C. Williams Eminent Scholar Chair, and Director of the Wireless Engineering Research and Education Center (WEREC), Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grid. He is a Distinguished Lecturer of IEEE Communications Society and the IEEE Council of RFID. He is the Editor-in-Chief of IEEE Transactions on Cognitive Communications and Networking.

Zhu Han (Fellow, IEEE) (hanzhu22@gmail.com) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. Currently, he is a John and Rebecca Moores Professor with the Department of Electrical and Computer Engineering and the Department of Computer Science, University of Houston, Texas. His main research interests include novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing data science, smart grid, carbon neutralization, and security and privacy.

Abbas Jamalipour (Fellow, IEEE) (a.jamalipour@ieee.org) received the Ph.D. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1996. He holds the position of a Professor of ubiquitous mobile networking with The University of Sydney. He is a Fellow of the Institute of Electrical, Information, and Communication Engineers (IEICE) and the Institution of Engineers Australia, an ACM Professional Member, and an IEEE Distinguished Speaker. Since January 2022, he has been the Editor-in-Chief of IEEE Transactions on Vehicular Technology.

H. Vincent Poor (Life Fellow, IEEE) (poor@princeton.edu) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests include information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields.

Dong In Kim (Fellow, IEEE) (dongin@skku.edu) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a Tenured Professor at the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a Distinguished Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is a fellow of the Korean Academy of Science and Technology and a member of the National Academy of Engineering of Korea.