From Generative AI to Generative Internet of Things: Fundamentals, Framework, and Outlooks

Jinbo Wen, Jiangtian Nie, Jiawen Kang, Dusit Niyato, Hongyang Du, Yang Zhang, and Mohsen Guizani

Abstract

Generative Artificial Intelligence (GAI) possesses the capabilities of generating realistic data and facilitating advanced decision-making. By integrating GAI into modern Internet of Things (IoT), Generative Internet of Things (GIoT) is emerging and holds immense potential to revolutionize various aspects of society, enabling more efficient and intelligent IoT applications, such as smart surveillance and voice assistants. In this article, we present the concept of GIoT and conduct an exploration of its potential prospects. Specifically, we first overview four GAI techniques and investigate promising GIoT applications. Then, we elaborate on the main challenges in enabling GIoT and propose a general GAI-based secure incentive mechanism framework to address them, in which we adopt Generative Diffusion Models (GDMs) for incentive mechanism designs and apply blockchain technologies for secure GIoT management. Moreover, we conduct a case study on modern Internet of Vehicle traffic monitoring, which utilizes GDMs to generate effective contracts for incentivizing users to contribute sensing data with high quality. Numerical results demonstrate the superiority of the proposed scheme. Finally, we suggest several open directions worth investigating for the future popularity of GIoT.

Introduction

The advent of Generative Artificial Intelligence (GAI) represents a significant milestone in the field of AI [1]. In contrast to traditional AI models that primarily classify or analyze existing data, GAI possesses the incredible ability to generate novel content such as digital films, audio, photos, or codes, thus exerting profound impacts across various domains [2]. For instance, in the healthcare domain, GAI can assist physicians in diagnosing conditions based on medical records and images, and generate tailored treatment plans for patients. In the tourism and hospitality domain, GAI can generate hyper-personalized content for tourists, fostering changes in tourism strategies such as destination planning and hotel booking. Simultaneously, the potential of GAI for network optimization has been explored [1], contributing to the optimization of network management and performance, thereby enhancing the efficiency of decision-making in complex networks [1].

Recent advances in cutting-edge technologies, such as Sixth-Generation (6G) wireless communications, Artificial Intelligence (AI), and edge computing, are bringing modern Internet of Things (IoT) technologies to maturity [3]. Modern IoT is considered an intelligent and autonomous ecosystem that revolutionizes device connectivity, data analytics, and intelligent decision-making. With its capabilities of ultra-low latency communications, seamless connectivity, and ubiquitous computing [3], modern IoT has the potential to enable novel and advanced applications across various industries and domains, including intelligent healthcare monitoring and smart homes. For instance, in smart homes, IoT devices can automate and control various aspects of home living, such as smart lighting systems and temperature control, providing enhanced convenience and comfort for users.

Given the remarkable capabilities of GAI in generating realistic data and facilitating advanced decision-making processes [4], we envision that GAI-empowered modern IoT will become more creative and proactive as such the term Generative IoT (GIoT) emerges. By leveraging the advanced data generation and decision-making capabilities of GAI, GIoT has the capacity to drive the progression of IoT-enabled environments. For instance, by processing historical sensing data and real-time sensor readings, GAI can forecast future events, predict system failures, and generate effective resource management to improve the overall system performance [1]. Although GIoT holds the significant potential to revolutionize various domains, there exist several critical challenges that need to be addressed when integrating GAI with modern IoT to enable GIoT, including the impact of IoT resource consumption on the performance of GAI model fine-tuning, the dynamic nature of GIoT networks complicating the identification of optimal decision strategies, and security concerns for GAI models in GIoT networks. *To the best of our knowledge, this is the first work that presents the concept of GIoT and systemically provides foresight research on the integration of GAI with modern IoT for enabling GIoT*. The contributions

The work was supported by NSFC under Grants No. 62102099, U22A2054, and No. 62071343, and Guangzhou Basic Research Program under Grant 2023A04J1699. The work was also supported by the Foundation of State Key Laboratory of Public Big Data under Grant No. PBD2023-12, the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research & Development Programme, DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019), DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) programme, and MOE Tier 1 (RG87/22).

Jinbo Wen and Yang Zhang (corresponding author) are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China; Jiangtian Nie, Dusit Niyato, and Hongyang Du are with the School of Computer Science and Engineering, Nanyang Technological University, China; Jiawen Kang is with the School of Automation, Guangdong University of Technology, China; Mohsen Guizani is with the Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE.

Digital Object Identifier: 10.1109/IOTM.001.2300255

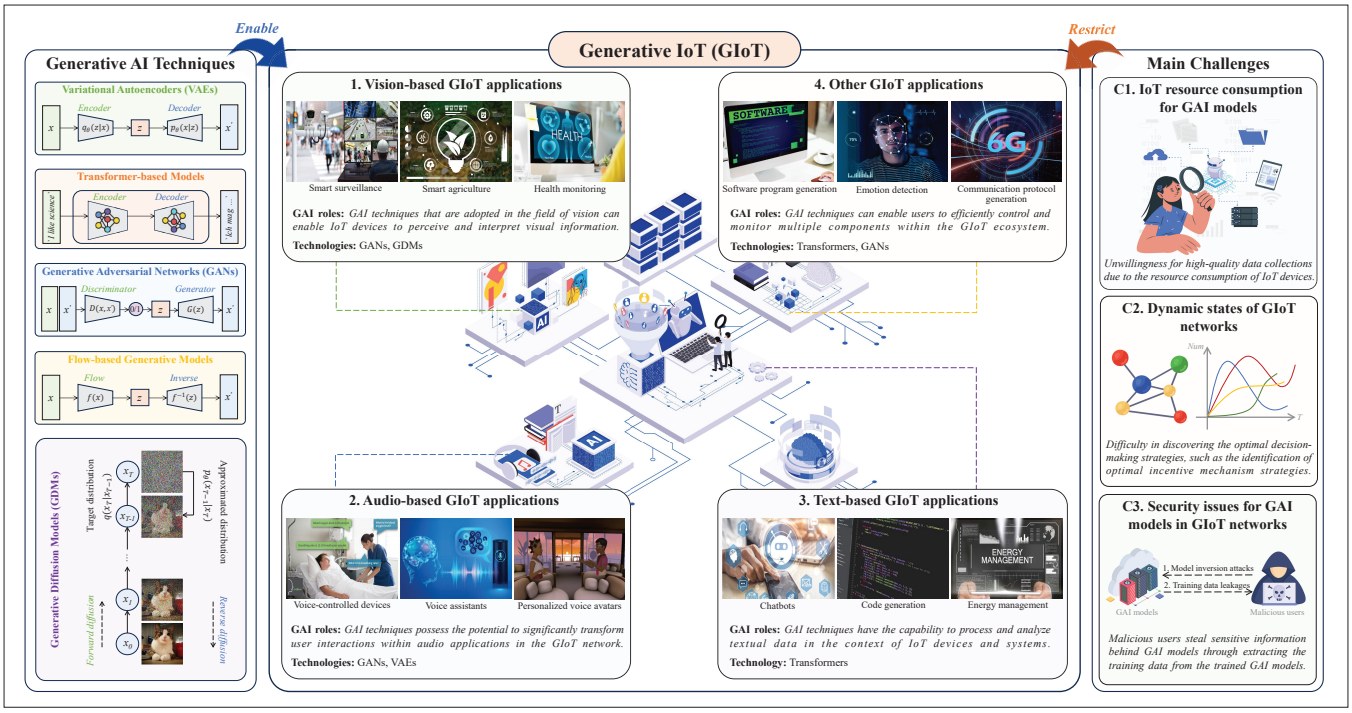

FIGURE 1. The schematic of generative IoT networks. We summarize four potential generative IoT applications, encompassing vision-based, audio-based, and text-based applications. Additionally, we discuss several GAI techniques that enable generative IoT applications and identify the main challenges that restrict the development and widespread adoption of such applications.

of this article can be summarized as follows:

- We first provide a comprehensive overview of GAI techniques that have been widely adopted in the field of computer vision. Then, we systematically discuss the potential GIoT applications and the main challenges of synergy between GAI and modern IoT to enable GIoT.
- We present a general incentive mechanism framework to address the main challenges in enabling GIoT. We utilize blockchain technologies to manage and secure GIoT and adopt Generative Diffusion Models (GDMs) to derive optimal incentive mechanism design.
- We conduct a case study on modern Internet of Vehicles (IoV) traffic monitoring, in which we develop a GDM-based contract theory model for incentivizing users to contribute high-quality sensing traffic data. Numerical results demonstrate that the edge server utility of the proposed scheme is almost 52% higher than that of the Deep Reinforcement Learning (DRL)-based scheme.

Generative Internet of Things

In this section, we introduce several widely adopted GAI techniques, especially in the field of computer vision, involving their basic architectures and applications for modern IoT. Then, we systematically explore the potential GIoT applications. Finally, we discuss the main challenges posed by enabling GIoT, as shown in Fig. 1.

Generative AI Technologies

As a powerful branch of AI, GAI focuses on creating new content in various modalities, such as videos, images, text, and audio [4]. GAI can leverage pre-trained models to generate new content by fine-tuning the model parameters based on user-provided input, i.e., prompts. Moreover, GAI can utilize learning algorithms to automate content generation from existing data. Motivated by recent studies [1, 7], we introduce four widely adopted model-based GAI techniques [7] and explore their potential for IoT applications.

- **Variational Autoencoders (VAEs):** VAEs consist of the encoder and decoder networks [7]. The encoder network compresses the input data to a latent representation. Then, the decoder network learns to reconstruct synthetic data that closely aligns with the original distribution [8]. Due to their ability to effectively represent data in a probabilistic latent space, VAEs can be applied to various IoT applications, such as energy optimization and equipment maintenance. For instance, by training on real-time sensor readings, VAEs can capture complex data distributions and generate more robust predictions on future equipment conditions than traditional AI methods [9].
- **Generative Adversarial Networks (GANs):** GANs have been applied widely in IoT data synthesis, consisting of generator and discriminator networks. The generator network aims to generate new data by learning real data distribution, while the discriminator network aims to distinguish synthetic data from real data [8]. The two networks are trained together in interactive and competitive manners, resulting in continuous enhancement of synthesis performance. With good performance in generating realistic samples [8], GANs can be utilized not only for data augmentation but also for IoT anomaly detection [9]. Notably, unlike traditional AI methods that typically require retraining on labeled data to adapt to changes, GANs can learn the underlying data distribution in an

IoT Applications	Advantages for the Applications	
	Generative AI	**Traditional AI**
Vision-based applications	• *Advanced image recognition*: GAI can attain higher accuracy in image recognition tasks by learning complex visual patterns from extensive datasets [5]. • *Enhanced object detection and tracking*: GAI can accurately detect and track objects without explicit labeling in real-time, enabling more robust surveillance systems [2]. • *Intelligent semantic understanding*: GAI can go beyond basic image processing and understand the semantic context of images, enabling more advanced applications like scene understanding and image captioning [6].	• *Efficient vision processing*: Traditional AI can be computationally less demanding compared to GAI, making them more efficient for specific vision-based tasks, particularly in large-scale image processing tasks. • *Simpler vision architectures*: Traditional AI, such as classical computer vision techniques, often utilize simpler architectures, e.g., convolutional neural networks and support vector machines, which are easier to implement.
Audio-based applications	• *Accurate speech recognition*: GAI can achieve higher accuracy in speech recognition tasks even in noisy acoustic environments, enabling voice-controlled IoT devices [7]. • *Enhanced voice understanding*: GAI can interpret spoken language by enhancing context and semantics, enabling intelligent voice assistants [5]. • *Smart sound anomaly detection*: GAI can detect anomalous audio patterns, enabling IoT applications like acoustic surveillance and predictive maintenance [2].	• *Robust noise handling*: Traditional AI can effectively handle noisy audio signals since they are often designed with specific signal processing techniques, such as filtering and noise cancellation. • *Tailored audio intelligence*: Traditional AI can be tailored to specific audio-based tasks, such as speaker identification and audio classification, demonstrating high performance.
Text-based applications	• *Nuanced text analysis*: GAI can process and analyze text data with more nuanced language understanding, enabling sentiment analysis and language translation [7]. • *Intelligent contextual understanding*: GAI can capture profound meaning from context, enabling more intelligent chatbots and personalized content delivery [2]. • *Precise code assistance*: GAI can generate programming codes from natural language description and provide coding assistance for users [7].	• *Efficient rule-based processing*: Traditional AI, such as rule-based systems, can be effective for text-based applications that require explicit and rule-driven decision-making. • *Transparent decision insight*: Traditional AI can enable users to understand the decision-making process and analyze the logic behind the output results.

TABLE 1. Advantage comparisons of generative AI and traditional AI for significant IoT applications.

unsupervised manner, enabling the adaptation to evolving anomalies without explicit labeling.

- **Flow-based Generative Models (FGMs):** FGMs can transform input data distributions from simple to complex through a series of differentiable and invertible transformations that are implemented as neural networks [7]. Unlike VAEs and GANs, FGMs possess the distinctive capability to learn explicitly the data distribution and directly compute the probability density function during generation [1]. Therefore, FGMs can circumvent resource-intensive computation and directly model complex probability distributions, which can be effectively applied in IoT domains such as traffic flow optimization [10] and anomaly detection in network traffic.
- **Generative Diffusion Models (GDMs):** With the state-of-the-art performance of image synthesis, GDMs are emerging generative models [7], consisting of forward diffusion and denoising processes inspired by non-equilibrium thermodynamics theory[8]. Because of their recent advancements in training and sampling efficiency, GDMs have been used not only for image generation but also for IoT network optimizations [1]. Specifically, GDMs exhibit the ability to capture complex and high-dimensional structures, effectively addressing network optimization problems and decision-making processes [1], while traditional AI methods often converge slowly and stuck in locally optimal solutions.

Generative IoT Applications

Unlike traditional AI, GAI with its capability of generating realistic and context-aware data has the potential to revolutionize various industries, such as healthcare, manufacturing, transportation, and smart cities. Note that the detailed advantage comparison of GAI and traditional AI for significant IoT applications is listed in Table 1. With the incorporation of GAI into modern IoT systems, a new paradigm called GIoT is emerging. GIoT holds the significant potential for transformative applications across various domains. By capitalizing on advanced IoT and GAI technologies, GIoT can enable intelligent systems, optimize resource utilization, enhance decision-making processes, and improve the overall efficiency and sustainability in diverse sectors.

Vision-based GIoT applications: Vision-based GIoT applications leverage the power of GAI technologies, particularly GAI techniques for computer vision like GDMs and GANs, to enable IoT devices to perceive and interpret visual information. Vision-based applications can be applied in various types of GIoT contexts, from real-time monitoring to remote diagnostic and maintenance [5], such as smart surveillance, smart agriculture, and health monitoring. For example, in smart surveillance systems, unlike the limited generalization capability of traditional AI, GAI can leverage data collected by IoT devices equipped with cameras and smart sensors to track objects and detect variable suspicious activities in various domains. Besides, in smart agriculture, GAI-empowered IoT devices can revolutionize farming practices. Specifically, since traditional AI lacks adaptabil-

ity to changing environmental conditions, GAI models can be adopted to predict the growth of crops based on the crop data captured by cameras mounted on drones or ground-based sensors, and the predictions can be shown in the form of images or videos [5], which facilitates data-driven decision-making for farmers.

Audio-based GIoT Applications: GAI technologies can advance our interactions with audio applications in GIoT networks. One of the most notable examples of audio applications based on GAI is voice assistant. Voice assistants, such as Amazon Alexa[1] and Apple Siri[2] that utilize audio-based GAI techniques, can understand and respond to voice commands like adjusting room thermostats and turning on and off lamps. Another important application of GAI in the audio domain is creating personalized audio systems for avatars that are highly accurate digital replicas of users, such as Resemble AI.[3] By analyzing the preferences and characteristics of users, GAI models can generate tailored audio systems for their avatars, seamlessly immersing users, especially in the metaverse.

Text-based GIoT Applications: Text-based GIoT applications involve leveraging GAI models like ChatGPT[4] to process and analyze text data in the context of IoT devices and systems. Specifically, GAI-empowered chatbots can process textual data generated by IoT devices and provide real-time insights. These conversational agents employ natural language processing algorithms to interpret natural language input, generate responses to users, and perform specific tasks like controlling IoT devices. Additionally, text-based GIoT applications also involve automated code generation. Specifically, by analyzing high-level specifications of desired functionality, GAI models, such as Codex[5] as a general-purpose programming model created by OpenAI, can automatically generate corresponding codes for specific IoT applications [5].

Other GIoT applications: There are also other novel GIoT applications in different modalities based on GAI technologies. One potential application is the automated generation of software programs, which can be downloaded to various IoT devices, enabling users to efficiently control and monitor multiple components within the GIoT ecosystem. Another potential application is the generation of secure communication protocols [5], such as the 6G wireless communication protocol. Since wireless communications between IoT devices are vulnerable to being compromised by malicious attackers, GAI techniques can be utilized to develop robust communication protocols that encrypt the data transmitted between devices, making attackers more difficult to access critical data in transit.

Main Challenges in Integrating GAI with Modern IoT

Although GAI technologies hold great potential for transforming the modern IoT ecosystem, the convergence of GAI technologies with modern IoT still suffers from the following challenges, which should be resolved for the future popularization and development of GIoT.

IoT Resource Consumption for GAI Models: In GIoT networks, GAI models require a modest quantity of extra data to perform model fine-tuning and direct inference at the edge, minimizing service latency and enhancing user experiences [10]. The data for model fine-tuning can be generated in the cloud or collected by IoT devices and subsequently uploaded by mobile users [10]. However, if the dataset contains deviations and inaccuracies in information, pre-trained GAI models cannot be accurately fine-tuned to specific tasks or domains, leading to inaccurate and biased inferences. Therefore, the dataset needs to be high-quality to avoid incorrect learning patterns in the GAI model fine-tuning [11]. Since data collection and transmission lead to high costs, users may not be reluctant to contribute high-quality data to the edge due to the resource constraints of IoT devices, affecting the performance of GAI model fine-tuning.

Dynamic States of GIoT Networks: Due to the scale and complexity of interconnected devices as well as the dynamic and real-time nature of the network, GIoT can be considered as a heterogeneous and large-scale system [12]. Consequently, intricate decision-making processes arise, such as the optimal allocation of limited IoT network resources and the identification of optimal incentive mechanism strategies. Generally, the optimal decision-making strategies are determined by employing the traditional optimization principle and tools [1]. However, these approaches often rely on accurate and comprehensive network information, which are not feasible in complex GIoT network scenarios. Additionally, while DRL has shown promise in various network optimization and decision-making tasks, the dynamics of IoT networks can significantly impact the state and action spaces of DRL models. This necessitates the complete retraining of DRL models [8], which may inefficiently discover the optimal decision-making strategies in dynamic GIoT network scenarios.

Security Issues for GAI Models in GIoT Networks: The heterogeneity of GIoT networks, exemplified by the ability of IoT devices to dynamically join or leave the networks as required [12], poses a significant difficulty in secure management for GIoT networks. Ensuring the quality and diversity of collected data by IoT devices is one of the key challenges. Specifically, malicious users equipped with IoT devices would deliberately upload low-quality data to the edge to obtain more benefits. Additionally, malicious users can issue model inversion attacks to steal sensitive information behind GAI models by extracting the training data from trained models [5]. For example, based on the text generated by ChatGPT, malicious users can deduce private information from either the fine-tuning data or the data employed to pre-train the foundation model of ChatGPT, which may lead to serious security threats to other normal users.

Motivated by the above analysis, it is necessary to develop a reliable and secure incentive mechanism framework, thereby enabling more intelligent and autonomous GIoT ecosystems. The proposed framework is discussed.

Generative AI-Based Incentive Mechanism Framework for Generative IoT

In this section, we introduce several representative techniques for designing incentive mechanisms for GIoT. To address the aforementioned challenges, we propose a general GAI-based secure incentive mechanism framework.

Motivated by the above analysis, it is necessary to develop a reliable and secure incentive mechanism framework, thereby enabling more intelligent and autonomous GIoT ecosystems. The proposed framework is discussed.

[1] https://www.aboutamazon.com/news/devices/amazon-alexa-generative-ai

[2] https://appleinsider.com/articles/23/09/06/

[3] https://www.resemble.ai/

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

[5] https://openai.com/blog/openai-codex

In IoT network optimizations, incentive mechanisms play a crucial role in incentivizing network users to actively contribute their resources, share data, or collaborate, thereby improving the performance and reliability of the network [1].

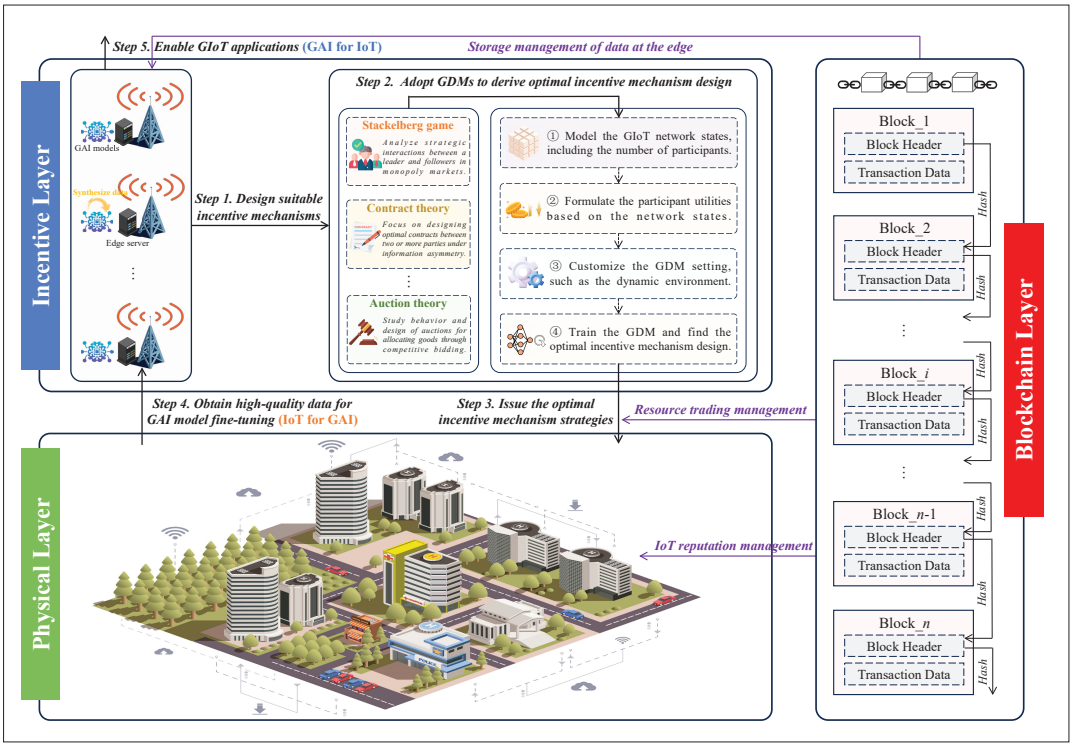

FIGURE 2. Generative AI-based secure incentive mechanism framework for generative IoT. The proposed framework consists of three layers, i.e., a physical layer, an incentive layer, and a blockchain layer, where the incentive layer is used to motivate users with IoT devices to provide high-quality data for generative AI model fine-tuning, and the blockchain layer is used to securely manage generative IoT networks.

Incentive Mechanism Design for Generative IoT

In IoT network optimizations, incentive mechanisms play a crucial role in incentivizing network users to actively contribute their resources, share data, or collaborate, thereby improving the performance and reliability of the network [1]. In the following part, we discuss several representative techniques for developing incentive mechanisms, i.e., Stackelberg game [13], contract theory [11], and auction theory [14], which have been widely adopted in IoT network optimizations [1].

Stackelberg Game: As a non-cooperative game theory, the Stackelberg game focuses on analyzing strategic interactions between a leader and followers, especially in IoT networks, where the leader first determines resource prices, and then the followers determine their resource demands based on the selling prices, until reaching the utility equilibrium [1]. For example, the authors in [13] focused on reliable vehicle twin migrations in vehicular metaverses and proposed a Stackelberg model between vehicular metaverse users and the roadside unit coalition with the highest utility.

Contract Theory: Contract theory is a powerful tool for incentive mechanism design under information asymmetry, which has been effectively applied in IoT network optimizations [1]. Specifically, an employer, typically a service provider, designs contracts for specific tasks, and employees, i.e., the network users, engage in a contractual agreement [1]. For instance, the authors in [11] studied Unmanned Aerial Vehicle (UAV)-enabled AI generative content and proposed a contract model. This model aimed to provide incentives for UAVs to contribute fresh data for GAI model fine-tuning under asymmetric information.

Auction Theory: Auction theory focuses on studying the behavior and design of auctions for allocating resources through competitive bidding [1]. As an interdisciplinary technology, auction theory has been widely adopted for incentivizing resource trading in IoT networks, which can be implemented in asymmetric or incomplete information scenarios [14]. For example, the authors in [14] proposed an auction-based optimization problem for the multichannel cooperative spectrum sharing in hybrid satellite-terrestrial IoT networks.

Framework Design

As shown in Fig. 2, we introduce the GAI-based secure incentive mechanism framework for GIoT, which consists of a physical layer, an incentive layer, and a blockchain layer. We provide more details of the framework as follows:

- **Step 1. Design suitable incentive mechanisms:** To address the reluctance of users to provide high-quality data for GAI model fine-tuning, edge servers as service providers would design suitable incentive mechanisms by considering the current conditions of GIoT networks, including network structures, performance metrics, resource constraints and so on.
- **Step 2. Adopt GDMs to derive optimal incentive mechanism design:** Due to the ability to capture high-dimensional and complex structures of intricate environments, GDMs can be adopted to derive optimal incentive mechanism strategies that can maximize the utilities of edge servers [1]. The motivations and specific process of utilizing GDMs for designing efficient and robust incentive mechanisms are introduced in [1].

- **Step 3. Issue the optimal incentive mechanism strategies:** After finding the optimal incentive mechanism strategies, edge servers issue the strategies to the physical layer. Moreover, the resource trading involved in executing the strategies can be securely recorded and managed in the blockchain layer, ensuring transparency and security in resource trading.
- **Step 4. Obtain high-quality data for GAI model fine-tuning:** Under the role of incentives, IoT devices collect fresh sensing data and provide them to the edge for GAI model fine-tuning. To ensure the quality of collected data, the reputation metric can be utilized to quantify the reliability of IoT devices and discourage malicious behavior, and the reputations of IoT devices would be securely managed in the blockchain layer [13].
- **Step 5. Enable GIoT applications:** Based on the data collected by IoT devices or generated in the cloud, GAI model fine-tuning and inference can be performed on edge servers to efficiently enable GIoT applications, and the data stored on edge servers can be also securely managed in the blockchain layer. Considering that certain types of training data might be idle, GAI techniques have the ability to autonomously synthesize data, enhancing the performance of models [5].

Case Study: GDM-Enabled Modern Internet of Vehicle Traffic Monitoring

In this section, we present a case study on modern IoV traffic monitoring. Specifically, we propose a GDM-based contract theory model, which can incentivize users to contribute high-quality sensing data for GAI model fine-tuning, facilitating substantial advancements in intelligent transportation systems.

System Model

Figure 3 depicts a specific case of the proposed framework. Specifically, with the capabilities of strong generalization and automated content generation [10], GAI can offer personalized and advanced services to users, such as navigation and route optimization [4]. To ensure the quality of services, GAI model fine-tuning at the edge requires high-quality datasets. However, due to the resource constraints of IoT devices, users may be reluctant to contribute fresh sensing data to the edge. Moreover, because of the dynamic and heterogeneous natures of vehicular networks [4], edge servers may lack awareness of the private information of users, such as their ability to collect sensing data, which can lead to users contributing data dishonestly to gain additional benefits [11].

Problem Formulation

We consider that each edge server can support M users. Based on statistical distributions of user types from historical data, we classify M users into N types and the user types are arranged in ascending order as $\theta_1 \leq \cdots \leq \theta_N$. In this definition, the higher type users can provide sensing data with the higher quality. For ease of understanding, the user with type n is called the type-n user.

User Utility: The utility of type-n users is denoted as U_n^C, equaling the difference between its obtained benefit and its cost of participation. As shown in Fig. 3, the obtained benefit of type-n users is defined as $(\theta_n R_n)$ [15], where R_n is the received reward. The cost of type-n users is defined as $\theta_n(L_{\max}/L_n - 1)$ [8], where L_n is the latency spent by type-n users in collecting and transmitting sensing data with a guaranteed amount. Note that L_{max} represents the highest permissible value of the latency.

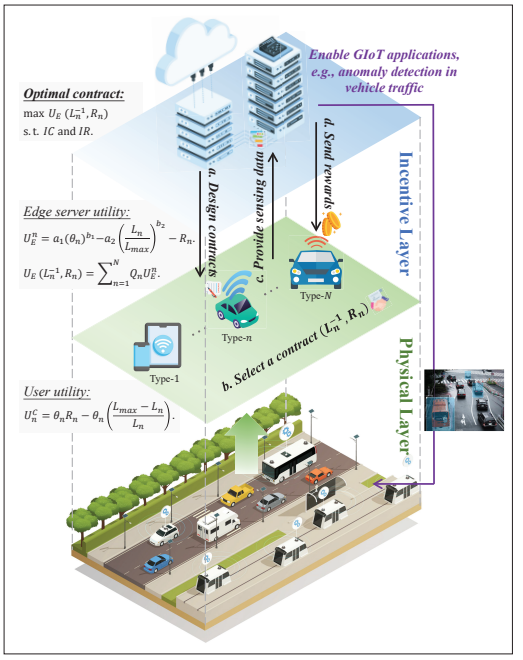

FIGURE 3. The illustration of GDM-based contract theory model for modern Internet of Vehicle traffic monitoring.

Edge Server Utility: The utility obtained by the edge server from type-n users is denoted as U_E^n, equaling the difference between the corresponding revenue for received datasets within L_n and the reward R_n. According to [8, 15], the revenue can be defined as a general quality-latency metric, i.e., $a_1(\theta_n)^{b1} - a_2(L_n/L_{\max})^{b2}$. Here, $a_1 > 0$ and $a_2 > 0$ are pre-defined coefficients about the quality of received data and the latency spent for collecting and transmitting data [15], respectively. Similarly, $b_1 \geq 1$ and $b_2 \geq 1$ are given factors measuring the effects of data quality and the latency [15], respectively. Considering that the probability that a user is of type-n is Q_n, where the sum of probabilities of all types is 1, the expected utility of the edge server U_E is shown in Fig. 3.

Contract Formulation: As an economic tool, contract theory is effective in addressing information asymmetry for incentive mechanism designs [11]. Therefore, the edge server can devise a contract comprising a group of contract items (L_n^{-1}, R_n), where L_n^{-1} is the reciprocal of L_n [15]. To ensure that each user optimally chooses the contract item designed for its type, the contract must satisfy both Individual Rationality (IR) and Incentive Compatibility (IC) constraints [11], where IR constraints indicate that the contract item that a user chooses should ensure a non-negative utility [15], and IC constraints indicate that a user of any type prefers to choose the contract item designed for its type rather than any other contract item [15]. Finally, the optimization problem is to find the optimal

To ensure the quality of services, GAI model fine-tuning at the edge requires high-quality datasets. However, due to the resource constraints of IoT devices, users may be reluctant to contribute fresh sensing data to the edge.

One of the main challenges of GIoT for future development is the computational and storage limitations in training and deploying GAI models.

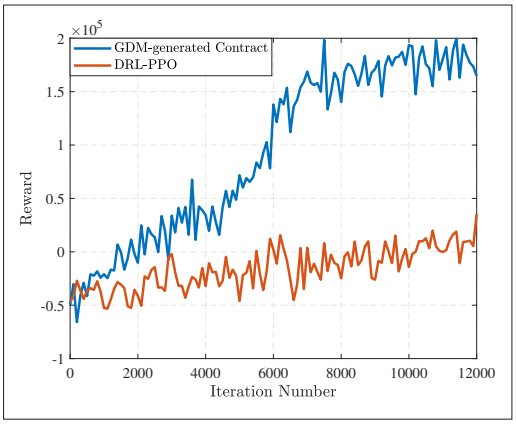

FIGURE 4. Training process of the GDM-based contract generation scheme and DRL-PPO for the optimal contract finding task, where the diffusion step is 100, the batch size is 512, and the contract generation network learning rate and contract quality network learning rate are 2×10^{-7}.

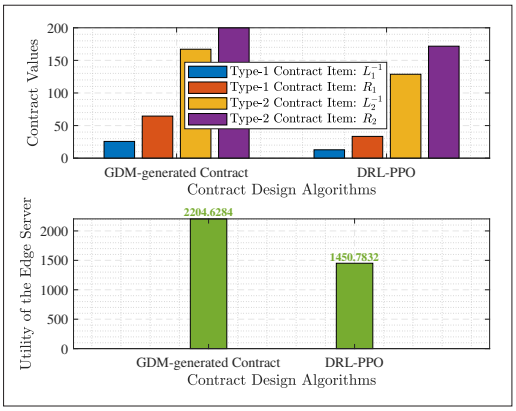

FIGURE 5. The designed contract comparisons between DRL-PPO and the GDM-based contract generation scheme.

contract c^*, i.e., $\{L_n^{-1^*}, \ldots, L_N^{-1^*}\}$ and $\{R_1^*, \ldots, R_N^*\}$, thereby maximizing the expected utility of the edge server U_E while satisfying IR and IC constraints.

GDM-Empowered Contract Generation

In this part, we adopt GDMs to derive optimal contract design [1]. Specifically,

- **Step 1. Model the environment state:** For simplicity, we consider that each edge server supports two types of users in the environment of vehicle traffic monitoring, where θ_1 and θ_2 are randomly sampled within (10,100) and (100,200), respectively [8]. Therefore, the environment state vector is defined as $S \triangleq [M, N, L_{max}, Q_1, Q_2, \theta_1, \theta_2]$. Note that Q_1 and Q_2 are generated randomly [8], following the Dirichlet distribution, and L_{max} is set as 150.
- **Step 2. Formulate the participant utilities:** After determining the environment states, we formulate the utility of type-n users U_n^C and the expected utility of the edge server U_E, where the former is used to guarantee IR and IC constraints, and the latter is the optimization objective that we intend to maximize. Note that the weighting factors a_1, a_2, b_1, and b_2 are set as 15, 10, 1, and 1 [15], respectively.
- **Step 3. Customize the GDM settings:** The action space is the universe of the contract design [1]. Each contract is formed as $\{L_1^{-1}, R_1, L_2^{-1}, R_2\}$. Then, we can customize the model hyperparameters. For instance, in our case, the training epoch of the GDM is set as 120, and the discount factor is set as 0.95.
- **Step 4. Train the GDM and generate the optimal contract:** We train the policy $\pi(c^*|s)$ for generating the optimal contract c^* under the state $s \in \{S\}$ [8]. To evaluate each generated contract, we adopt the action-value function $Q(c^*|s)$ [8], which can guide the diffusion process. Finally, we can obtain the optimal contract c*.

Numerical Results

We carry out the experiment on NVIDIA GeForce RTX 3080 Laptop GPU with CUDA 12.0. Figure 4 shows test reward curves of our proposed GDM-based contract generation scheme and conventional DRL-PPO for the optimal contract finding task. We can observe that our proposed scheme always outperforms DRL-PPO under the same parameter settings. The reason is that the contract generation policy in our scheme is fine-tuned by the diffusion process, which can mitigate the impact of randomness and noise [1]. Figure 5 illustrates the quality of contracts generated by the proposed scheme and DRL-PPO. For a given environment state, we can observe that the proposed GDM-based contract generation scheme can provide a contract design that achieves the edge server utility value of 2204.6284, which is greater than the 1450.7832 achieved by the PPO. Overall, the above numerical results demonstrate that the performance of the proposed GDM-based contract generation scheme is better than that of DRL-PPO.

Future Directions

Distributed and Green Generative AI Models

One of the main challenges of GIoT for future development is the computational and storage limitations in training and deploying GAI models. For instance, GPT-3, as the state-of-the-art language model of OpenAI, consists of 175 billion parameters [10], which is one of the largest language models in existence. Therefore, how to reduce the energy consumption of GAI models during their training and deployment is worth studying. One of the potential solutions is to design lightweight GAI models or adopt federated learning to train GAI models.

Quality Metrics for Reliable Generative AI Outputs

Although GAI techniques have the incredible ability to automate content generation, they can be exploited to generate incorrect or fraudulent content, such as fake videos or wrong texts [5]. To address this issue, future research can explore the Quality of Service (QoS) metric from the user perspective to measure user satisfaction with the generated content. For instance, user preferences and feedback can be incorporated into the model training process. With the help of QoS, the performance of GAI models can be improved, and high-quality content can be generated to meet user satisfaction.

Service Optimization by Prompt Engineering

Formulating technical prompts to effectively instruct GIoT presents a challenge for individuals

lacking adequate training in the relevant domain. Furthermore, the utilization of subpar prompts may diminish the overall generation quality of GAI models. Therefore, the exploration of prompt engineering for achieving the optimization of AI-generated content services is also a topic worthy of investigation. For instance, users can manually formulate diverse prompts and subsequently search for the one that yields the highest quality of generated outputs.

Security and Privacy Protection for Users

Centralized training or fine-tuning of GAI models at the edge may raise user concerns about data privacy and security [10], as IoT data involving sensitive and personal information could potentially be exposed to attackers, leading to threats to user privacy and security. Therefore, future research can develop a user-centric privacy-preserving training approach to protect user security and privacy.

Conclusion

In this article, we presented the concept of Generative IoT (GIoT). First, we reviewed several GAI techniques and explored their potential for IoT applications. Then, we summarized GIoT applications, including vision-based, audio-based, and text-based applications, and discussed the main challenges of integrating GAI with modern IoT to enable GIoT. To address these challenges, a general GAI-based secure incentive mechanism framework was proposed, in which we adopted GDMs for the optimal incentive mechanism design and utilized blockchain technologies for secure GIoT management. Furthermore, we conducted a case study on modern IoV traffic monitoring, leveraging GDMs to generate flexible contracts for motivating users to provide high-quality data for GAI model fine-tuning. The numerical results demonstrated the effectiveness of our proposed GDM-based contract generation scheme compared to DRL-PPO. Finally, we discussed potential research directions that can further facilitate the development of the GIoT ecosystem.

References

[1] H. Du *et al.*, "Beyond Deep Reinforcement Learning: A Tutorial on Generative Diffusion Models in Network Optimization," arXiv preprint arXiv:2308.05384, 2023.

[2] C. Zhang *et al.*, "A Complete Survey on generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need?" arXiv preprint arXiv:2303.11717, 2023.

[3] D. C. Nguyen *et al.*, "6G Internet of Things: A Comprehensive Survey," *IEEE Internet of Things J.*, vol. 9, no. 1, 2022, pp. 359–83.

[4] R. Zhang *et al.*, "Generative AI-Enabled Vehicular Networks: Fundamentals, Framework, and Case Study," arXiv preprint arXiv:2304.11098, 2023.

[5] M. A. Ferrag, M. Debbah, and M. Al-Hawawreh, "Generative AI for cyber threat-hunting in 6G-Enabled IoT Networks," arXiv preprint arXiv:2303.11751, 2023.

[6] L. Xia *et al.*, "Generative AI for Semantic Communication: Architecture, Challenges, and Outlook," arXiv preprint arXiv:2308.15483, 2023.

[7] Y. Cao *et al.*, "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT," arXiv preprint arXiv:2303.04226, 2023.

[8] Y. Liu *et al.*, "Deep Generative Model and Its Applications in Efficient Wireless Network Management: A Tutorial and Case Study," arXiv preprint arXiv:2303.17114, 2023.

[9] A. A. Cook, G. Mısırlı, and Z. Fan, "Anomaly Detection for IoT Timeseries Data: A Survey," *IEEE Internet of Things J.*, vol. 7, no. 7, 2019, pp. 6481–94.

[10] M. Xu *et al.*, "Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services," arXiv preprint arXiv:2303.16129, 2023.

[11] J. Wen *et al.*, "Freshness-Aware Incentive Mechanism for mobile AI-Generated Content (AIGC) networks," *2023 IEEE/CIC Int'l. Conf. Commun. in China (ICCC)*, 2023, pp. 1–6.

[12] N. C. Luong *et al.*, "Data Collection and Wireless Communication in Internet of Things (IoT) Using Economic Analysis and Pricing Models: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 4, 2016, pp. 2546–90.

[13] Y. Zhong *et al.*, "Blockchain-Assisted Twin Migration for Vehicular Metaverses: A Game Theory Approach," *Trans. Emerging Telecommunications Technologies*, 2023, p. e4856.

[14] X. Zhang *et al.*, "Auction-Based Multichannel Cooperative Spectrum Sharing in Hybrid Satellite-Terrestrial IoT Networks," *IEEE Internet of Things J.*, vol. 8, no. 8, 2020, pp. 7009–23.

[15] J. Kang *et al.*, "Toward Secure Blockchain-Enabled Internet of Vehicles: Optimizing consensus Management Using Reputation and Contract Theory," *IEEE Trans. Vehic. Tech.*, vol. 68, no. 3, 2019, pp. 2906–20.

Biographies

Jinbo Wen (jinbo1608@163.com) received the B.Eng. degree from Guangdong University of Technology, China, in 2023. He is currently pursuing an M.S. degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His research interests include AIGC, blockchain, and metaverse.

Jiangtian Nie (jnie001@e.ntu.edu.sg) received her B.Eng degree with honors in Electronics and Information Engineering from Huazhong University of Science and Technology, China. She received her Ph.D. degree with ERI@N in the Interdisciplinary Graduate School, Nanyang Technological University (NTU), Singapore. She was a visiting student at Princeton University and University of Waterloo. Her research interests include network economics, game theory, wireless blockchain, crowd sensing, and learning.

Jiawen Kang (kavinkang@gdut.edu.cn) received the Ph.D. degree from the Guangdong University of Technology, China, in 2018. He was a postdoc at Nanyang Technological University, Singapore from 2018 to 2021. He is currently a professor at Guangdong University of Technology, China. His research interests mainly focus on blockchain, security, and privacy protection in wireless communications and networking.

Dusit Niyato (dniyato@ntu.edu.sg) is a professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. He received the B.Eng. degree from King Mongkuts 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. His research interests are in the areas of sustainability, edge intelligence, decentralized machine learning, and incentive mechanism design.

Hongyang Du (hongyang001@e.ntu.edu.sg) received the B.Sc. degree from Beijing Jiaotong University, Beijing, China, in 2021. He is currently working toward a Ph.D. degree with the School of Computer Science and Engineering, Energy Research Institute @ NTU, Nanyang Technological University, Singapore, under the Interdisciplinary Graduate Program. His research interests include semantic communications, reconfigurable intelligent surfaces, and communication theory. He was the recipient of the IEEE Daniel E. Noble Fellowship Award in 2022. He was recognized as an Exemplary Reviewer of the IEEE Transactions on Communications in 2021.

Yang Zhang (yangzhang@nuaa.edu.cn) is currently an associate professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He received the B. Eng. and M. Eng. from Beihang University in 2008 and 2011, respectively. He obtained a Ph.D. degree in Computer Engineering from Nanyang Technological University, Singapore, in 2015. He is an editor of the IEEE Transactions on Machine Learning in Communications and Networking. His current research topic is edge computing and multi-agent unmanned systems.

Mohsen Guizani (mguizani@ieee.org) received his B.S. (with distinction) and M.S. degrees in electrical engineering, and M.S. and Ph.D. degrees in computer engineering from Syracuse University, New York, in 1984, 1986, 1987, and 1990, respectively. He is currently a Professor with the Machine Learning Department, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE.