Edge Intelligence in the Generative Artificial Intelligence Era

Xinyuan Zhang, Gaochang Xie, Yudong Huang, Zehui Xiong, Jiang Liu, Shuguang Cui, *Fellow, IEEE*, Sumei Sun, *Fellow, IEEE*, and Xuemin Sherman Shen, *Fellow, IEEE*

Abstract

Edge intelligence (EI), by leveraging abundant edge resources and positioning AI algorithms closer to end-users, has long been considered a fundamental catalyst for the AI industry. As the AI realm shifts towards new Generative AI (GAI), EI offers a broader data source, reduced latency, and enhanced privacy protections, making it a more conducive environment for GAI advancements than cloud-based approaches. However, compared to traditional AI models, GAI challenges existing EI with its significantly larger model size, markedly intricate operations, and substantially heightened resource demands. This article delves deeply into the evolution of EI in the upcoming GAI era. Particularly, we first provide a thorough overview of challenges introduced by GAI, including escalated communication costs, greater computational demands, and intensified security and privacy concerns. We then extend the EI scope to encompass the entire lifecycle of GAI within EI, while jointly considering sensing, communication, and computation against these emerging challenges. Additionally, we spotlight key techniques designed to pave the way for the future of EI, elaborating on each of these in detail. To provide concrete insights into how EI adapts for GAI, we present two illustrative case studies: one focusing on diffusion model-based GAI fine-tuning in vehicular networks and the other highlighting large language model-based real-time inference offloading in wireless edge networks. Lastly, we outline three future research directions for EI, guided by the latest advancements in GAI.

I. Introduction

Edge intelligence (EI) is widely recognized as a leading paradigm that consistently empowers artificial intelligence (AI) applications [1]. By deploying AI algorithms on pervasive edge resources near end users, EI bridges AI's last mile gap with users, enhancing its accessibility. The vast data produced at the edge daily also catalyzes the development of new AI applications, propelling continuous growth in AI.

Given EI's widespread acceptance and its proven efficacy over the years, it is expected to continue supporting the rapidly evolving AI industry. In the past year, new Generative AI (GAI) has been a significant milestone in AI's evolution [2], as evidenced by the phenomenal successes of Stable Diffusion and ChatGPT. While traditional AI utilizes recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to classify data, new GAI leverages diffusion and transformer models to grasp the real-world data distribution. Whereas traditional AI yields simple prediction, classification, and clustering results, GAI facilitates the creation of highly realistic content across images, text, sound, and videos [2]. We foresee that GAI will completely reshape our lives and EI needs to be positioned to support the growing GAI prominence.

EI provides GAI with numerous benefits, including reduced latency, augmented service capacity, a sustainable data supply, and enhanced privacy protection [3]. However, EI encounters specific challenges due to GAI's distinct characteristics compared to conventional AI models. GAI's huge model size and complex computations enhance content creation capabilities, but impose a significantly greater communication, computation, and energy burden on EI compared to traditional AI. Moreover, GAI amplifies data security and privacy issues as it regularly requires training data, often sourced from sensitive domains, to adapt to downstream applications. As these challenges call for EI's adaptation and evolution, it is urgent to revisit EI in a fundamental and holistic manner.

In this article, we delve into the evolution of EI in the new era of GAI. We begin by demonstrating the new challenges and opportunities that GAI has introduced to EI. Next, we revisit the EI framework from a new GAI perspective, presenting and analyzing several key techniques within this framework. Finally, we

This work was supported in part by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme, in part by SUTD-ZJU IDEA under Grant SUTD-ZJU (VP) 202102, in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 under Grant MOE-T2EP20221-0017, in part by SMU-SUTD Joint Grant under Grant 22-SIS-SMU-048, in part by SUTD Kickstarter Initiative under Grant SKI 20210204, in part by Shenzhen Outstanding Talents Training Fund under Grant 202002, in part by Guangdong Research Projects under Grant 2017ZT07X152 and Grant 2019CX01X104, in part by Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001, and in part by Shenzhen Key Laboratory of Big Data and Artificial Intelligence under Grant ZDSYS201707251409055.

Digital Object Identifier: 10.1109/MWC.2025.3599652
Date of Current Version: 26 December 2025
Date of Publication: 24 October 2025

Xinyuan Zhang and Yudong Huang are with the State Key Laboratory of Networking and Switching Technology, BUPT, Beijing 100876, P.R. China; Gaochang Xie was with the State Key Laboratory of Networking and Switching Technology, BUPT, Beijing 100876, P.R. China. He is now with Purple Mountain Laboratories, Nanjing 211111, China; Zehui Xiong (corresponding author) is with the Information Systems Technology and Design (ISTD) Pillar, Singapore University of Technology and Design, Singapore 487372; Jiang Liu is with the State Key Laboratory of Networking and Switching Technology, BUPT, Beijing 100876, P.R. China, and also with Purple Mountain Laboratories, Nanjing 211111, P.R. China; Shuguang Cui is with the School of Science and Engineering (SSE), Shenzhen Future Network of Intelligence Institute (FNii-Shenzhen), and Guangdong Provincial Key Laboratory of Future Networks of Intelligence, The Chinese University of Hong Kong, Shenzhen 518066, China; Sumei Sun is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632; Sumei Sun is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Feature		Traditional AI	GAI	
Representative model		DNN/CNN/RNN	GPT-4.0	Stable Diffusion
Number of parameters		Thousands to millions	~1.8 Trillion	~9.8 Million
Model size		KBs to MBs	TBs	GBs
Training/ Fine-tuning	Data volume	MBs to GBs	PBs /TBs to PBs	12 million images /Tens of images
	Time costs	Seconds to hours	Weeks to months /Weeks	Weeks /Seconds to hours
	GPU consumption	~Tens of GPUs	Tens of thousand GPUs / Thousands of GPUs	256 GPUs /Few GPUs
	Money expenses	Thousands of dollars +	63 million dollars / Thousands to millions of dollars	Millions of dollars / Most under tens of dollars
Inference	Time costs	Milliseconds to seconds	Seconds to minutes	Seconds to minutes
	GPU consumption	~One to tens of GPUs	128 GPUs	~One to tens of GPUs
	Money expenses	Low	~10 dollars for a hundred thousand words	~One dollar for five thousand images

TABLE I. Comparison between traditional AI and GAI across different phases.

present two case studies to illustrate the application of some novel techniques in GAI fine-tuning and inference processes and highlight three promising future directions for exploration within the framework. *To the authors' knowledge, this article is among the first works that illuminates the evolution of EI oriented towards the new GAI age*. The primary contributions of this article are:

- We outline the benefits, challenges, and state-of-the-art efforts of bringing GAI services into EI with examples. From these insights, we introduce three foundational principles for the EI framework tailored for GAI.
- Guided by these principles, we undertake a comprehensive revisit of the EI framework, promoting a resource-effective, energy-conserving, and sustainable GAI lifecycle that seamlessly integrates sensing, communication, and computation.
- We present and analyze several key techniques within the framework that EI can harness to achieve the feasibility and practicality of fostering GAI.
- We illustrate two case studies: diffusion model-based GAI fine-tuning for image generation in vehicular networks and large language model-based GAI offloading for real-time text generation in wireless edge networks. Drawing from these case studies, we discuss the prospective approaches to implementing GAI within future EI.

The remainder of the article is organized as follows. Section II presents the benefits, challenges, and design principles of EI in the GAI era. Section III revisits the EI framework tailored for GAI, while Section IV discusses key techniques, and Section V includes two case studies in the EI framework oriented towards GAI. Some future directions are proposed in Section VI, followed by the article's conclusion.

II. When EI meets GAI

In this section, we delve into the benefits, challenges, and state-of-the-art efforts of bringing GAI services into EI. Based on these insights, we propose three core principles for the future EI framework optimized for GAI.

A. The Benefits of EI for GAI

Due to the resource-intensive nature of GAI, it is currently hosted on the cloud. The centralized approach faces potential flaws that will hinder GAI expansion in the near future. For example, the long propagation delay to the remote cloud center deters user experience. This is illustrated when players of the game "Mount&Blade 2" endure a three to five-second delay during dialogues with GPT-enhanced NPCs[1]. Moreover, the current cloud-based approach places restrictive limits on usage. Even for paying users of ChatGPT, the limit stands at fifty messages every three hour[2]. Additionally, the cloud lacks adequate and up-to-date data to train new GAI models, which leads to output hallucinations that are inconsistent with real-world facts [4]. Current high-quality language data for GAI training is projected to be exhausted by 2026, with image data expected to run out between 2030 and 2050[3].

In contrast to the centralized cloud, EI deploys GAI models in pervasive edge resources near users, substantially amplifying GAI computational capacity. This also allows users to seamlessly enjoy real-time and reliable GAI service connections through various wireless communication technologies, including 5G/6G mobile communication, WiFi, and so on. Moreover, EI can offer a vast pool of data, collected daily from billions of mobile users and IoT devices. This data assists in consistently training GAI models to reduce hallucinations, adapt to societal shifts, customize for downstream applications, and personalize outputs for diverse users. By processing data directly where it is generated, EI also ensures privacy preservation. In summary, EI provides a paramount infrastructure for GAI, and is poised to emerge as a prominent catalyst for future GAI development.

B. Challenges and Start-of-the-Art Efforts

Despite the above EI advantages for GAI, the current EI designs are still based on traditional AI, lacking adaptability to the evolving GAI models. This mismatch between current EI and GAI primarily stems from differences between traditional AI and the new GAI in two dimensions: model design and resource consumption.

- *Model design:* GAI models are intricately crafted, incorporating billions of neurons to emulate complex human neural responses. Consequently, the model parameters and size of GAI drastically surpass those of traditional AI. As illustrated in Table I, we take the leading transformer-based GAI model GPT-4.0[4] and the typical diffusion-based model Stable Diffusion[5] for example. GAI models exceed traditional AI in size by over 1,000 times.
- *Resource consumption:* For GAI models to mirror real-world data distributions accurately,

[1] https://www.polygon.com/ai-artificial-intelligence/23650693/chatgpt-generative-ai-video-game-development
[2] https://community.openai.com/t/chatgpt4-now-with-50-messages-every-3-hours/304697
[3] https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset
[4] https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/
[5] https://en.wikipedia.org/wiki/Stable_Diffusion

rigorous training is indispensable. As shown in Table I, training or fine-tuning GPT-4.0 demands substantially more data and resources than traditional AI, with GAI inference also being resource-intensive.

The above unique GAI characteristics pose challenges for current EI:

- *Significant communication cost.* To train or fine-tune GAI models aligned with modern societal norms and new businesses, it is imperative to consistently gather multi-modal data from diverse industries and individuals. This mandates constant data sensing from end devices and data collection at the edge, leading to significant communication overheads. Moreover, given the extensive parameter count in GAI models, collaborative training or fine-tuning across multiple edge devices incurs substantial communication costs for weight aggregation.
- *Resource-intensive computation.* The resource-intensive GAI training, fine-tuning, and inference raise issues related to hardware capacity, memory footprint, computational efficiency, and energy utilization for resource-constrained edge devices.
- *Security and privacy concerns.* GAI's ability to produce realistic content necessitates securing its training dataset to prevent biased or low-quality outputs. There's also a risk of users inadvertently revealing personal data during GAI interactions. Protecting data from sensitive sectors when adapting GAI models is essential.

AI researchers are striving to compress GAI models and reduce computational costs to facilitate their operation on mobile devices. For example, the Stable Diffusion model has been compressed to 1.7MB [5]. LoRA [6] reduced the number of GPT-3's trainable parameters by 10,000 times and the GPU memory requirement by 3 times. DreamBooth [5] can fine-tune a personalized Stable Diffusion model using only a few images within seconds. [7] enables LLM inference on a single GPU. These studies lay a foundation for the feasibility and practicality of hosting GAI in future EI.

C. Design Principles

While the foundation of hosting GAI in future EI has been preliminarily explored, the emphasis has been on computation acceleration. There still lacks a systematic and clear EI framework tailored for GAI. To address this, we first present key design principles in this section and then take a comprehensive revisit of EI framework in the next section. The potential solutions to the challenges are discussed in Section IV.

The following design principles need to be considered:

- *Balance between computation efficiency and accuracy loss:* Given edge devices' constraints, GAI models must leverage some computation-efficient techniques to save on memory and energy. However, efficiency often comes at the cost of some accuracy loss or may even lead to hallucination during fine-tuning and inference. The framework should consider user requirements and ensure the computation efficiency without sacrificing accuracy.
- *Collaboration across heterogeneous devices with multi-dimensional resources:* Devices within EI vary in computation speed, memory, storage capacity, wireless channel status, and data distribution. The framework should optimize the devices' collaborative fine-tuning and inference processes while jointly considering sensing, communication, and computation resources.
- *Security, factuality, and privacy:* With GAI's capacity to create highly realistic content, the framework must prioritize the security and factuality of training data, ensuring its authenticity, impartiality, and non-toxicity. Additionally, the framework should rigorously protect the privacy of training datasets from sensitive sectors such as healthcare and finance.

III. Shifting EI Framework Towards Embracing GAI

Although embracing GAI in EI offers benefits like reduced latency, augmented service capacity, data enrichment, and privacy safeguards as described in Section II-A, the large scale and complexity of GAI models hinder satisfactory GAI training and inference within current EI frameworks. Emerging EI systems must therefore streamline GAI lifecycle operations and jointly optimize communication, computational, and storage resources. Following the design principles in Section II-C, we propose the future EI framework oriented towards GAI era shown in Fig. 1. The cloud center with powerful cloud servers is located remotely. AI enterprises host their GAI models within this cloud infrastructure and are responsible for GAI model pre-training, management, and dispatching to the EI layer. On the other hand, the EI layer consists of edge servers and end devices. Edge servers, including microdata centers at mobile network base stations, routers on vehicles or UAVs, offer GAI services to end devices. End devices, such as mobile phones, IoT devices, and vehicles, request GAI services from edge servers. Crucially, the EI layer manages the entire lifecycle of GAI involving the following phases. Note that the mentioned key techniques in new EI are depicted in Fig. 2 and are elaborated in the next section.

- *Data collection phase:* EI consistently provides GAI fine-tuning and inference with vast data collected and sensed from pervasive end devices. Key factors include transmission cost and data security. To reduce transmission costs, data can be pre-processed at end devices to extract features before transmission to edge servers. If resources permit, end devices can aggregate data locally for fine-tuning and inference. For data security and privacy, local storage on end devices is ideal. When local devices lack sufficient resources, data aggregation at edge servers is required, which should ensure unbiased and representative datasets. Mathematical tools can assess dataset biases [4]. Overall, the EI framework should optimize communication and security by strategically placing data, considering resource constraints and user distribution.
- *Model fine-tuning phase:* GAI is fine-tuned for customization, personalization, and to avoid failures in generalizing to new knowledge fields. Given their significant scale compared to smaller AI models deployed in current EI, GAI requires advanced computational and storage resource management techniques. As shown in Fig. 1, when end device resources are sufficient, fine-tuning can be performed locally. Model

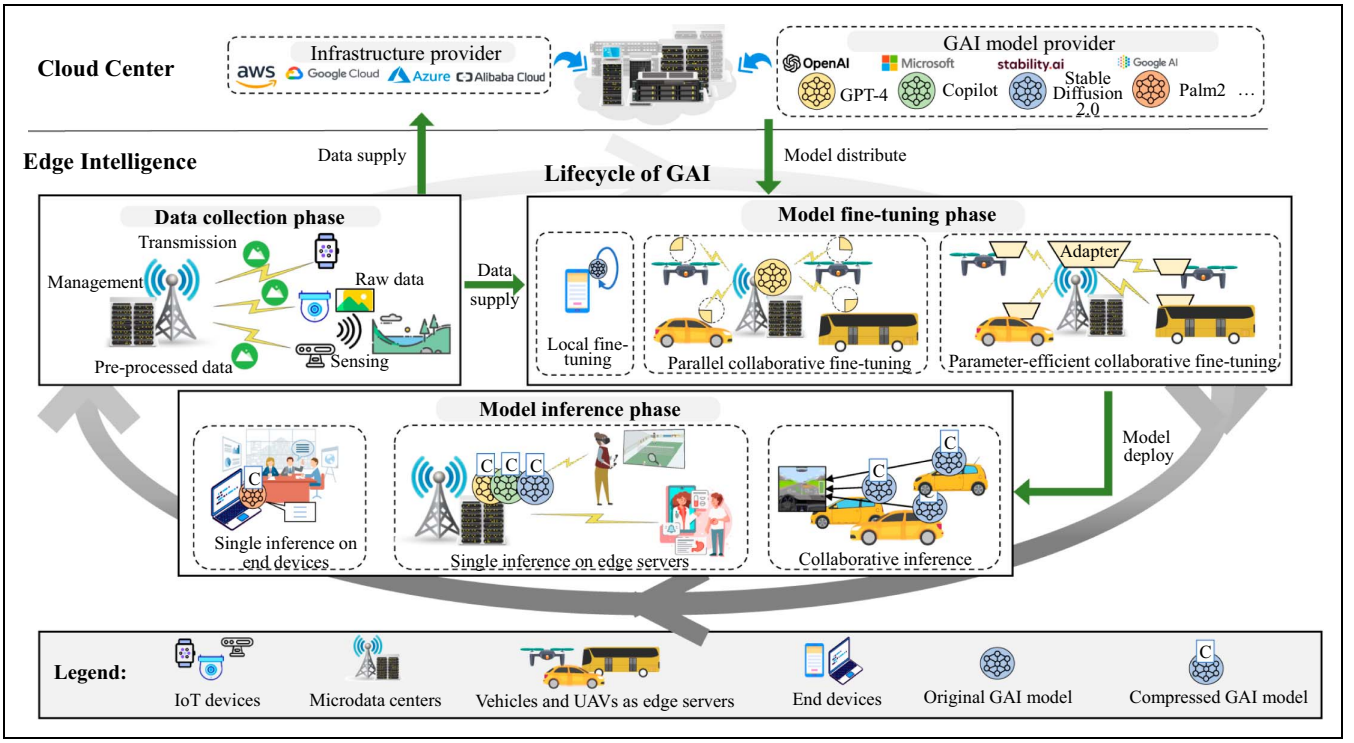

FIG. 1. Future EI framework, with techniques of compression, acceleration, collaboration, and offloading, is tailored for the entire lifecycle of GAI, including data collection, fine-tuning, and inference phases.

quantization techniques (Section IV-A) can be tailored to reduce costs without sacrificing quality by leveraging GAI's tolerance to weight quantization and sensitivity to activation quantization [8]. Prompt tuning (Section IV-B3) is also an effective technique, guiding the model's output without modifying its parameters. If local resources or datasets are insufficient, collaborative fine-tuning across multiple servers and devices is required. In such cases, parameter-efficient techniques, such as parallel lightweight fine-tuning in federated learning (Section IV-B1) or transfer learning with adapters (Section IV-B2), should be employed. Unlike traditional EI, the design of adapters in the new EI should prioritize GAI's generation diversity, style consistency, and output quality.

- *Model inference phase:* Inference is the process that feeds input prompts into the GAI model and generates outputs. The computational complexity of GAI models makes it challenging for current EI to perform inference on common devices within acceptable time frames. For single-device inference, hardware acceleration (Section IV-C1) should be employed, optimized for GAI architectures like the multi-head attention module, to efficiently distribute workloads across CPUs, GPUs, and storage. Additionally, early-exit mechanisms (Section IV-C2) should incorporate novel quality detection strategies for GAI to skip redundant inference steps. Model quantization can also be utilized (Section IV-A). For collaborative inference across multiple devices, new EI should explore offloading techniques (Section IV-D) to minimize communication overhead, accounting for the specific architecture of GAI models, such as the auto-regression process in transformer decoders.

IV. Key Techniques in the EI Framework for the GAI Era

In this section, we delve into the key techniques highlighted in the framework and describe them comprehensively.

A. Post-Training Model Quantization

Model quantization is one of the most extensively studied techniques in GAI model compression. Among its varied methods, post-training quantization (PTQ) is preferred for LLMs, as it avoids the significant cost of retraining such a large model [8]. As shown in Fig. 2 Part A, PTQ quantizes LLMs' weights and activations to lower precision data types to reduce memory footprint and computation costs. While LLMs exhibit greater tolerance to weight quantization compared to traditional AI models, they present challenges in activation quantization. This is attributed to LLMs' vast hidden activation dimensions and broader dynamic range, complicating the quantization of activations. The performance of activation quantization is closely related to LLM types and specific activation quantization details [8]. Thus, edge devices deploying LLMs can utilize weight quantization methods to reduce resource footprint but should carefully choose activation quantization methods.

B. Parameter-Efficient Fine-Tuning

1) Collaborative fine-tuning for specialized GAI with federated learning: Contrary to cloud-based GAI, EI-enabled GAI services aim to offer customized contents tailored for diverse intelligent services in different edge areas. To achieve this within the resource-limited wireless EI networks, collaborative fine-tuning is essential, with federated learning (FL) standing out as a compelling solution. FL harnesses specialized end-device data without the necessity of cloud transmission,

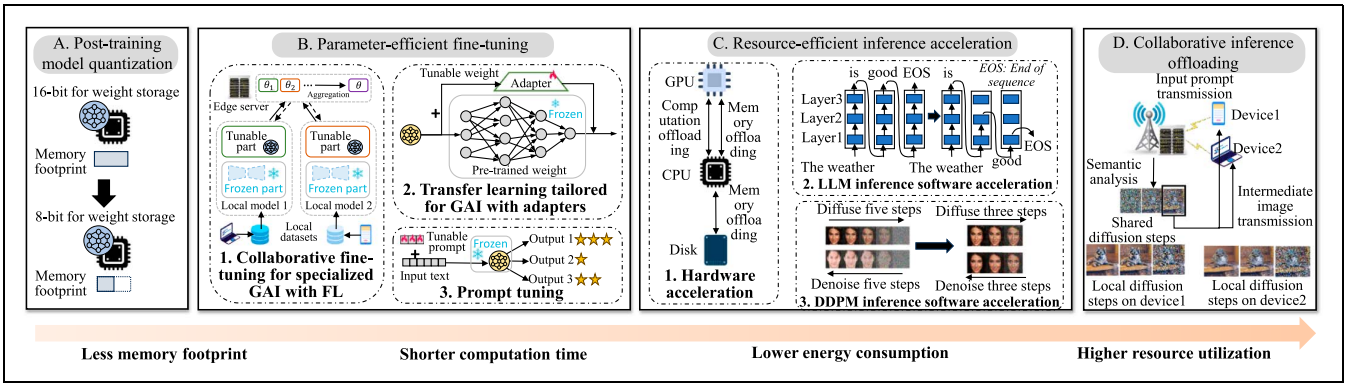

FIG. 2. Key techniques in the future EI framework oriented towards the GAI era.

preserving privacy [9]. Additionally, emerging lightweight fine-tuning methods like DreamBooth [5] provide resource-efficient fine-tuning solutions for each end device participating in FL. Specifically, DreamBooth allows adjustments to pre-trained diffusion models using only 3-5 specialized images and ensures the output diversity using a prior-preservation loss strategy, providing more flexible and parameter-efficient capabilities to FL-based GAI. Leveraging efficient fine-tuning methods on individual devices, collaborative FL fine-tuning can better utilize valuable locally distributed data, aligning with flexible training approaches like one-shot and few-shot FL, to optimize data and parameter efficiency within EI.

2) Transfer learning tailored for GAI with adapters: When specific tasks do not have sufficient data to train a model from scratch, transfer learning is a powerful method to leverage a generic pre-trained model and further refine the model on a new and smaller dataset. Fine-tuning is an approach to implement transfer learning. While fine-tuning the full model can be resource-intensive, recent advances[10] proposed fine-tuning with adapters, which freezes most of the weights, inserts small adapter modules between the existing layers, and only trains the parameters of the adapter as depicted in Fig. 2 Part B. Given the huge size of GAI models, using adapters can improve both the parameter efficiency and flexibility of transfer learning. However, the previous transfer learning with adapters approach cannot directly used to GAI. While traditional AI primarily evaluates the discriminative performance (e.g., classification accuracy) during transfer learning, GAI must uphold generation diversity, style consistency, and content quality, adding complexity to the transfer learning process. Moreover, considering the multi-head attention mechanism, generative AI requires a more complex adapter structure than traditional AI, such as adding a specific loss function or regularization. The VL-adapter [11], tailored for vision-and-language generative tasks, showcased this advanced approach. It was demonstrated that such adapter-based methods could match the performance of fine-tuning the entire model like VL-BART and VLT5, signifying a promising direction for efficient GAI fine-tuning in future EI.

3) Prompt tuning: Prompt tuning refers to adjusting the input prompts (i.e., task descriptions or a few canonical instances) to guide the model toward generating desired outputs. In Fig. 2 Part B, it sidesteps the need to alter the model's weights, thus offering a unique way to adapt models to new tasks with minimal changes and computational costs. It is particularly suitable for large-scale GAI models. Research[12] has indicated that for language models boasting billions of parameters, prompt tuning outperforms few-shot prompts and narrows the performance disparity with model tuning. However, crafting an effective prompt can be tricky and labor-intensive since it might require many iterations. The size and inherent bias of the model limit the results of prompt tuning, which must be factored into future EI considerations.

C. Resource-Efficient Inference Acceleration

1) Hardware-based acceleration: As illustrated in Fig. 2 Part C, hardware-based acceleration distributes inference tasks across various hardware components, including CPUs, GPUs, and disks. By leveraging increased memory and computational power, hardware acceleration methods can expedite inference without compromising accuracy [7]. This provides an efficient hardware resource utilization approach for end devices in EI.

2) Software-based acceleration: This focuses on skipping redundant inference procedures to enhance speed. Apart from well-known parallelism techniques [13], early-exit mechanisms have emerged as a promising solution. For example, diffusion models iteratively denoise a randomly sampled Gaussian noise to generate a high-quality image, while hundreds or even thousands of denoising steps needed in practice lead to significantly low inference efficiency. To address this issue, the early-stop strategy [14] in Fig. 2 Part C considers only few initial diffusing steps and the reverse denoising process starts from a non-Gaussian distribution. By skipping numerous denoising steps, it greatly accelerates the inference speed.

D. Collaborative Inference Task Offloading

Given the limited resources on a single edge device, it is often more practical to offload resource-intensive inference tasks across multiple edge servers and end devices. This approach better coordinates the sensing, communication, and computation resources throughout the EI network. One critical challenge, arising from the unique architectural design of GAI models, is determining whether the inference process is divisible and, if so, how to segment the inference computation. In this context, an edge-end offloading technique for diffusion-based image generation tasks using text prompts in [15] is demonstrated in Fig. 2 Part D. They divide the inference procedures based

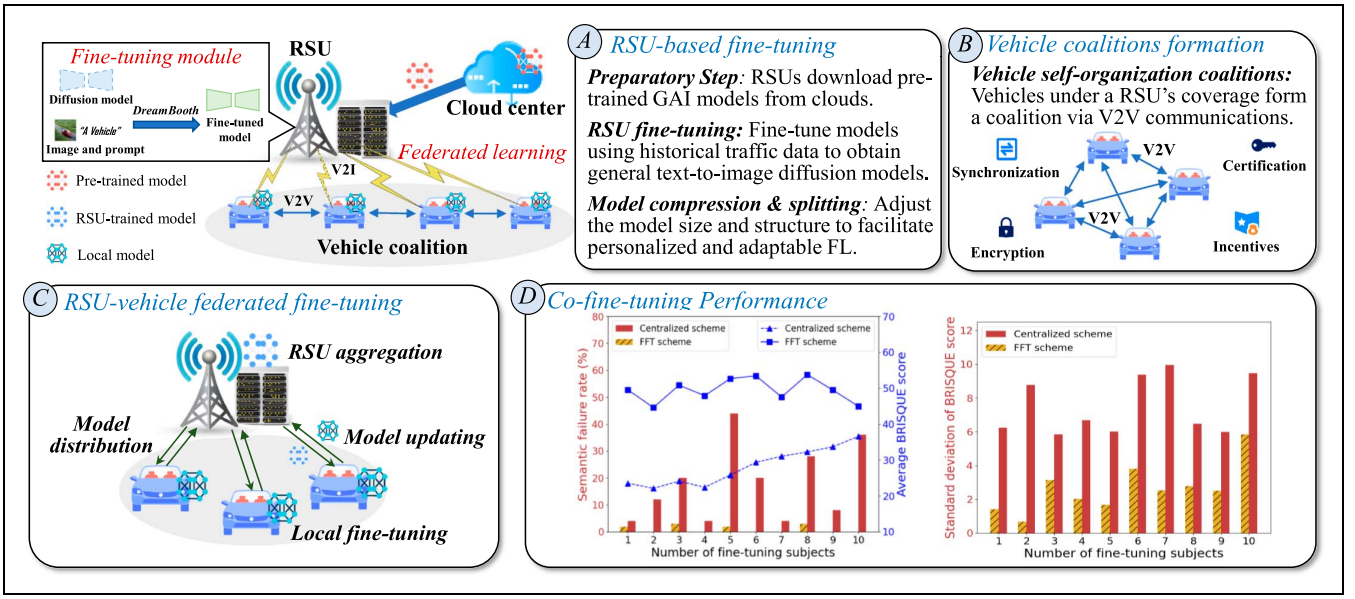

FIG. 3. Text-to-image diffusion model co-fine-tuning experiments in vehicular networks for service-specific content generation. In the experiments, each vehicular device is equipped with an Intel Xeon CPU with 12 cores and 32GB memory and an NVIDIA GeForce RTX 3060 GPU with 12.74 TFLOPS. The experimental model and dataset are the Stable Diffusion v1-4 model and the German traffic sign recognition benchmark. We simulate a environment where the centralized method utilizes RSU high-quality historical datasets and the FFT leverages real-time service-specific data collected by vehicles.

on diffusion denoising steps. The edge server undertakes the semantic analysis to identify similarities among prompts from different end devices, and it manages the shared denoising steps due to its superior resource availability. Intermediate outputs are then relayed from the edge server to end devices for further processing, catering to distinct text semantics.

V. Case Studies

In this section, we present two case studies: collaborative fine-tuning based on the diffusion model in vehicular networks, and real-time inference offloading based on LLM in wireless networks. These case studies exemplify the future EI framework detailed in Section III.

A. Collaborative Fine-Tuning in Wireless Vehicular Networks

Vehicular networks exemplify a typical EI scenario as shown in Fig. 1, where vehicles and roadside units (RSUs) collaboratively provide drivers with vehicular services like augmented reality (AR) navigation, in-vehicle 3D entertainment, and even vehicular metaverse. In this case, GAI is utilized to create diverse, novel, and customized content. However, fundamental GAI models, trained on data from a broad spectrum of life, are too general to meet users' specific personalization needs. Thus, in this subsection, we introduce a fine-tuning method for diffusion-based vehicular network image generation, showcasing how to adapt GAI efficiently for personalization as outlined in Section III. To mitigate privacy risks from vehicles uploading real-time locations and surrounding images to a central database, our approach utilizes FL. This method differs from current EI approaches by fine-tuning only a small part of the model with prompts tailored to vehicular tasks, such as traffic objects, guideposts, and speed limit signs, rather than the time-consuming and resource-intensive fine-tuning of the entire GAI model with extensive data on each device. Our method minimizes resource and data usage on individual devices, enabling low-latency GAI fine-tuning suitable for devices with limited capabilities.

In Fig. 3, each RSU is equipped with edge servers and maintains extensive historical traffic datasets, whereas vehicles with diverse computational capacities are distributed across urban areas. We adopt the DreamBooth technique as an exemplary efficient FL fine-tuning method on distributed devices. The process is divided into two stages to further ease computational burden on end devices and improve output quality: an RSU-based fine-tuning stage and an RSU-vehicle federated fine-tuning (FFT) stage. In the first stage, illustrated in Fig. 3 Part A, RSUs retrieve pre-trained diffusion models from cloud centers and refine them with historical traffic data. They then implement model compression and splitting to modify the model dimensions and structures. These tailored models are disseminated to vehicles, enabling a personalized and flexible FL process. Before the second stage, as depicted in Fig. 3 Part B, nearby vehicles form self-organizing coalitions via vehicle-to-vehicle (V2V) communications for collaborative FFT with the associated RSU. Vehicle mobility is considered by calculating its coverage time under an RSU, based on speed and acceleration. Vehicles join the RSU coalition if this time exceeds the FL task's computation and transmission duration; otherwise, they do not. Advanced mechanisms like certification, synchronization, encryption, and incentivization ensure the robustness and security of these vehicle groupings. In the second stage, models previously refined by RSUs are locally enhanced with service-specific images. Fig. 3 Part C shows that each FFT iteration updates and aggregates these local models, achieving specialized generation capabilities tailored for vehicular contexts. If a vehicle's connection with the RSU deteriorates or disconnects during an iteration, the

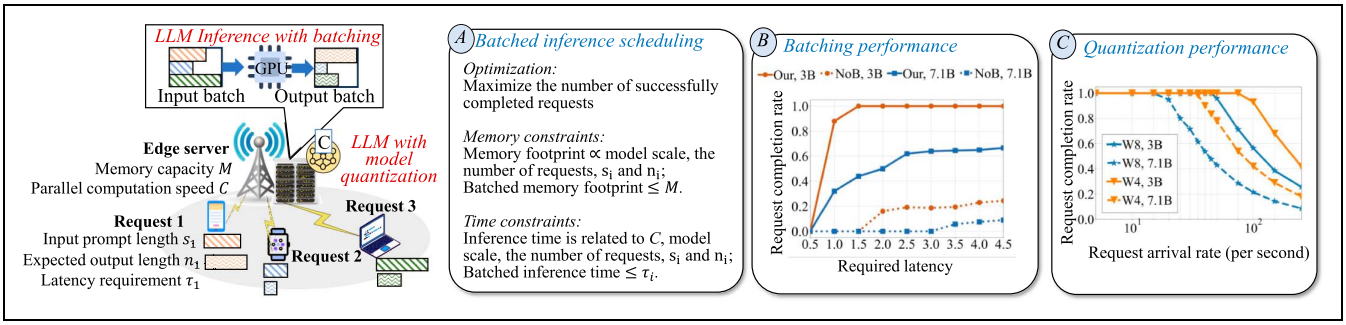

FIG. 4. Formulation and simulation of LLM inference offloading in wireless networks. In the simulation, the edge server is equipped with 20 NVIDIA JETSON TX2 GPUs, each with 1.33 TFLOPs computation speed and 32GB memory. The channels follow Rayleigh fading. The input prompt length and expected output length is randomly selected from 128, 256, and 512 tokens. Inference period length is 2 seconds. Request arrival rate is 50 per second.

RSU may retransmit the task, allocate it to another vehicle, or even discard its parameter modifications.

In the simulation, we undertake multiple fine-tuning processes using images spanning different traffic subjects to cater to a range of vehicular GAI services. We examine the semantic failure rate of the generated images by measuring the proportion of images that are semantically incorrect relative to the overall output. Besides, the quality of the generated images is measured with the blind/referenceless image spatial quality evaluator (BRISQUE), a pivotal metric for evaluating human-perceived image naturalness, where a lower BRISQUE score indicates better image quality.

As depicted in Fig. 3 Part D, the FFT scheme achieves a 22-fold decrease in semantic failure rate compared to the centralized scheme relying solely on RSU historical data. This is by capitalizing on service-specific data to refine the generating capability of GAI models. Note that this enhancement results in a drop in the naturalness of the produced images, given the subpar quality of local training data compared to RSU datasets. Moreover, considering the standard deviation of BRISQUE scores, the images generated by the centralized scheme display more variability than those from the FFT scheme. This variance can be attributed to the ability of FFT to continuously employ local, service-specific data, whereas the centralized scheme faces challenges adapting to diverse service needs in real-time, leading to fluctuating performance.

B. LLM Inference Offloading in Wireless Networks

In this subsection, we present a detailed case study on LLM inference offloading in new EI. We consider various end devices within the wireless range of multiple edge servers, each equipped with LLMs tailored to its resources and user requirements. Using transformer decoder-based LLMs as an example, we highlight key differences from traditional EI offloading. Firstly, unlike traditional EI that supports model splitting across devices, distributing parts of an LLM to different edge servers incurs high communication overhead, as each auto-regression iteration requires all prior processing results to pass through all model layers, rendering LLM splitting infeasible for offloading. Secondly, LLM output size is unpredictable, contrasting with the typically small output size in current EI. Lastly, LLM inferences are notably more resource-demanding than traditional AI, placing significant burdens on resource-limited edge devices. Given these issues, we suppose an LLM inference request can only be computed on a single edge server. To address resource limitation on edge servers, we adopt model quantization to reduce memory footprint and utilize batching to reuse the GPU-loaded model weights during inference and improve inference throughput.

We take the inference scheduling on one edge server for example. As shown in Fig. 4 Part A, given an inference period and a batch of requests, the edge server arranges inference based on the request information while guaranteeing its memory constraints and requests' time constraints. The request information includes input prompt length, expected output length, and required latency. Memory footprint during batched inference depends on the LLM's parameters and batched request information, such as the number of requests, and input prompt/output lengths. Inference time is influenced by the server's parallel processing capability, LLM parameters, and batched request information.

We use CPLEX[6] to solve this binary assignment problem. In the simulation, we employ two open-source transformer decoder-based LLMs: BLOOM with 3 billion parameters (3B) and 7.1 billion parameters (7.1B). Given LLMs' resilience to weight quantization, we apply two quantization methods: storing weights in 8 bits (W8) and 4 bits (W4), detailed in [8]. Request completion rate is defined as the ratio of successfully completed requests within latency requirements to total requests. In Fig. 4 Part B, we benchmark against a no-batching (NoB) inference approach, where requests wait in a first-in-first-out queue and get computed when the edge server is idle. Results show our batched inference consistently surpasses NoB in request completion rate, underlining the benefits of batching in throughput enhancement. As required latency rises, the completion ratio increases due to extended batch computation time. The larger LLM consistently has a lower completion ratio than the smaller one due to its higher memory and processing demands. The throughput improvement from a more powerful quantization method is shown in Fig. 4 Part C. Transitioning from W8 to W4 precision not only reduces weight memory consumption but also boosts computation efficiency for weight matrices. Yet, as the request arrival rate increases, the completion rate declines because of the edge server's capacity constraints.

VI. Future Directions

In this section, we highlight three promising future directions for exploration within the EI framework in the GAI era.

[6] https://ibm.com/products/ilog-cplex-optimization-studio/cplex-optimizer

A. Feature-Based Data Processing for Multi-Modal Content Generation

EI enables localized data collection and processing, making it well-suited for time-sensitive applications like VR/AR and the Metaverse games, where multi-modal inputs–such as posture, voice, and body shape–are collected to create personalized avatars. However, the sheer volume of multi-modal sensor data strains both EI processing capabilities and wireless network bandwidth. To mitigate this, instead of transmitting raw sensor data, multi-modal GAI models should push their feature extraction and alignment modules to the distributed end or edge. The placement of small embedding models, which utilize RNN/CNN and transformer architectures to extract features, should be determined by balancing device capabilities, energy consumption, and bandwidth usage. The placement of multi-modal fusion and alignment modules should also consider device resource limits as well as the input precision and fidelity.

B. Hybrid Federated Split Learning-Based Collaborative Fine-Tuning With Retrieval-Augmented Generation (RAG)

To enable GAI customized capabilities within EI while preserving user data privacy, federated learning (FL) and lightweight fine-tuning techniques can be employed to train models efficiently across decentralized edge devices without sharing data. However, EI still faces challenges, such as resource constraints and data quality limitations in FL-based training. In response, integrating split learning and RAG techniques can further optimize fine-tuning on ubiquitous edge devices. Split learning allows GAI components to be assigned based on each device's hardware and physical topology, while RAG addresses the issue of insufficient high-quality data at each learning node. Nevertheless, ensuring seamless cooperation between split learning and FL in complex wireless environments and constructing dynamically specialized edge RAGs remains an urgent challenge.

C. Mixture-of-Expert (MoE) Based Joint Inference Offloading

In EI environments with diverse services and device heterogeneity, deploying full-scale GAI models often results in resource inefficiencies. MoE models address this by employing specialized sub-modules (experts) and a gating network that directs input tokens to necessarily activated experts for efficient inference. However, native MoE models adopt token-expert routing and parallel processing, lacking adaptation to wireless edge environments. To address this, the gating network's routing logic must be designed to accommodate dynamic network conditions, edge resource availability, and user service requirements. Additionally, expert module deployment should minimize data transfer costs and latency, and inference offloading strategies should support collaborative processing across heterogeneous devices, dynamically allocating experts among the edge based on task requirements and network conditions.

VII. Conclusion

In this article, we have revisited EI in the new GAI era. We have first delineated the design principles for the future EI framework tailored for the GAI landscape and then revisited the EI framework, emphasizing its evolution in supporting the entire lifecycle of GAI. We have then discussed several key techniques with the potential to ensure the feasibility and practicality of fostering GAI within EI. To further elucidate the EI framework designed for GAI, we have presented two case studies and underscored potential research trajectories. We expect that this article should offer insights into advancing the adoption and popularity of GAI services and encourage further research and innovation in this critical field.

References

[1] D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, and P. Hui, "Edge intelligence: Empowering intelligence to the edge of network," *Proc. IEEE*, vol. 109, no. 11, pp. 1778–1837, Nov. 2021.

[2] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," 2023, *arXiv:2303.04226*.

[3] Y.-C. Wang, J. Xue, C. Wei, and C. C. J. Kuo, "An overview on generative AI at scale with edge–cloud computing," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2952–2971, 2023.

[4] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Nov. 2024.

[5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22500–22510.

[6] E. J. Hu et al., "LoRa: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–13.

[7] Y. Sheng et al., "FlexGen: High-throughput generative inference of large language models with a single GPU," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2023, pp. 31094–31116.

[8] Z. Yao, C. Li, X. Wu, S. Youn, and Y. He, "A comprehensive study on post-training quantization for large language models," 2023, *arXiv:2303.08302*.

[9] W. Zhuang, C. Chen, and L. Lyu, "When foundation model meets federated learning: Motivations, challenges, and future directions," 2023, *arXiv:2306.15546*.

[10] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 2790–2799.

[11] Y.-L. Sung, J. Cho, and M. Bansal, "VI-Adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5227–5237.

[12] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empirical Methods in Natural Lang. Process.*, 2021, pp. 3045–3059.

[13] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[14] Z. Lyu, X. Xu, C. Yang, D. Lin, and B. Dai, "Accelerating diffusion models via early stop of the diffusion process," 2022, *arXiv:2205.12524*.

[15] H. Du, R. Zhang, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, X. S. Shen, and H. V. Poor, "Exploring collaborative distributed diffusion-based AI-generated content (AIGC) in wireless networks," *IEEE Netw.*, vol. 38, no. 3, pp. 178–186, May 2024.

Biographies

Xinyuan Zhang received the B.S. degree in communication engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2019, and the Ph.D. degree from the State Key Laboratory of Networking and Switching Technology, BUPT, in 2024. She was a visiting Ph.D. student with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore, since 2022. She is currently an Assistant Researcher with Zhongguancun Laboratory. Her current research interests include satellite-terrestrial integrated networks, edge intelligence, and generative AI.

Gaochang Xie received the Ph.D. degree from the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China, in 2025. He was a

Visiting Scholar with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore, from 2023 to 2025. He is currently a Postdoctoral Research Fellow with Purple Mountain Laboratories, China. His current research interests include edge intelligence and network optimization.

YUDONG HUANG received the B.S. degree in communication engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2019, and the Ph.D. degree from BUPT, in 2024. He was a visiting Ph.D. student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, from 2022 to 2023. His current research interests include time-sensitive networks, deterministic networks, and network intelligence.

ZEHUI XIONG received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He is currently an Assistant Professor with Singapore University of Technology and Design, and also an Honorary Adjunct Senior Research Scientist with Alibaba-NTU Singapore Joint Research Institute, Singapore. He was the Visiting Scholar with Princeton University and University of Waterloo. His research interests include wireless communications, internet of things, blockchain, edge intelligence, and metaverse.

JIANG LIU received the B.S. degree in electronics engineering from Beijing Institute of Technology, Beijing, China, in 2005, the M.S. degree in communication and information system from Zhengzhou University, Zhengzhou, China, in 2009, and the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, in 2012. He is currently a Professor with Beijing University of Posts and Telecommunications. His current research interests include network architecture, network virtualization, satellite networking, software-defined networking (SDN), information-centric networking (ICN), and network testbed.

SHUGUANG CUI (Fellow, IEEE) received the Ph.D. degree from Stanford, in 2005. He is currently a X.Q. Deng Presidential Chair Professor with The Chinese University of Hong Kong, Shenzhen, China. His current research interest is data driven large-scale information analysis and system design. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the Worlds' Most Influential Scientific Minds by ScienceWatch in 2014. He was the recipient of the IEEE SP Society 2012 and ComSoc 2023 Marconi Best Paper Awards. He is member of Both Royal Society of Canada and Canadian Academy of Engineering.

SUMEI SUN (Fellow, IEEE) is the Executive Director of the Institute for Infocomm Research (I^2R), Agency for Science, Technology, and Research (A*STAR), Singapore. She also holds an Adjunct appointment with the National University of Singapore, and joint appointment with the Singapore Institute of Technology, both as a full professor. Her current research interests include next-generation wireless communications, joint communication-sensing-computing-control design, industrial internet of things, and applied artificial intelligence. She is a member of the IEEE Vehicular Technology Society Board of Governors (2022–2027) and fellow of the Academy of Engineering Singapore.

XUEMIN SHERMAN SHEN (Fellow, IEEE) is a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, the internet of things, 5G and beyond, and vehicular networks. He was the President of IEEE Communication Society.