When AIGC Meets MEC: A Novel Diffusion-Based Collaborative Inference Paradigm

Hongjia Wu, Xinyi Zhuang, Jiaqi Wu, Lin Gao, Dusit Niyato, and Tse-Tin Chan

Abstract

Driven by rapid advancements in mobile edge computing and model compression technologies, generative diffusion models (GDMs) are increasingly being deployed at the edge to support a broad range of AI-generated content (AIGC) applications. However, delivering efficient and high-quality AIGC services remains a significant challenge due to the inherent limitations of edge resources and the growing diversity of personalized user demands. To address these challenges, collaborative inference has emerged as a promising paradigm. By partitioning model execution between edge servers and end devices, this approach effectively leverages the complementary strengths of both platforms in terms of scalability, computational efficiency, and personalization. In this article, we first provide a comprehensive overview of the paradigm evolution of collaborative inference tailored for GDMs, with a particular focus on how different paradigms trade off efficiency, reliability, and quality. Building upon this foundation, we propose a semantic-aware cross-task collaborative inference (SemCT) framework that examines how prompt semantic similarity and shared inference proportion affect content quality, and leverages these insights for more effective prompt clustering and denoising step allocation. We further present a case study to validate the effectiveness of SemCT and offer practical insights into efficient AIGC service delivery. Finally, we identify and discuss key challenges in this domain and outline promising directions for future research. In summary, this work aims to deepen the understanding of collaborative inference for GDMs, highlighting critical methodologies and open issues to guide further exploration.

Introduction

The rapid development of AI-generated content (AIGC) technologies is reshaping digital media by facilitating high-quality generation of images, videos, and multimodal content. As a core driver behind these capabilities, generative diffusion models (GDMs) have attracted widespread attention for their outstanding performance in numerous applications [1]. However, their substantial computational demand results in prohibitively high inference costs and energy consumption. For example, running Stable Diffusion XL on a cloud platform for 8 hours per day over 20 working days costs approximately $310, while even a small-sized model like Stable Diffusion 1.5 consumes roughly 1.38×10^{-3} kWh per inference (comparable to driving a typical electric car about eight meters) [2]. Moreover, cloud-based inference introduces significant latency due to long-distance data transmission and raises potential privacy risks, making it unsuitable for real-time and interactive AIGC applications.

To address these limitations, industrial deployments are increasingly placing model training in the cloud and pushing model inference to the edge to meet real-time performance demands. This approach effectively reduces transmission latency and bandwidth consumption, alleviates the cloud's computational burden, and enhances data privacy. However, deploying compute-intensive inference on resource-constrained edge devices poses a significant challenge. To bridge this gap, mobile edge computing (MEC) offers a practical platform by providing flexible and scalable computational support at the edge. Meanwhile, advanced model compression techniques (e.g., pruning, distillation, and quantization) significantly reduce the resource requirements for model deployment. Together, these developments accelerate the large-scale deployment of GDMs from the cloud to the edge and end devices, paving the way for efficient and scalable AIGC services.

Many studies in this field focus on the independent inference paradigm [3–6], where inference tasks can be performed at end-user devices locally, or be offloaded to edge servers to alleviate the resource constraints of end devices. However, relying on a single computing node is often inadequate for AIGC services. On the one hand, edge-side inference suffers from throughput degradation and increased latency under heavy traffic and may raise privacy concerns in sensitive scenarios. On the other hand, end-side inference, though privacy-preserving, is constrained by on-device resources, resulting in slower inference and degraded content quality (e.g.,

Hongjia Wu and Tse-Tin Chan (corresponding author) are with the Department of Mathematics and Information Technology, The Education University of Hong Kong, China; Xinyi Zhuang and Lin Gao (corresponding author) are with the School of Electronics and Information Engineering and the Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology, Harbin Institute of Technology, Shenzhen, China; Jiaqi Wu is with the School of Computer Science, Guangdong University of Finance, China; Dusit Niyato is with the College of Computing and Data Science, Nanyang Technological University, Singapore. Hongjia Wu and Xinyi Zhuang contributed equally to this work.

Digital Object Identifier: 10.1109/MCOM.001.2500620

0163-6804/26/$25.00 © 2026 IEEE

lower resolution and missing details in the generated content). These limitations motivate the development of more flexible and adaptive inference paradigms. Cross-layer collaborative inference [2] has emerged as a promising paradigm for addressing the above limitations by properly distributing the total denoising steps among different computing nodes. This distributed paradigm enhances resource efficiency by leveraging the computational capabilities of heterogeneous computing nodes, improves scalability and fault tolerance, and increases user privacy by allowing localized data processing. Building on these insights, prior works [7–9] have investigated such an inference paradigm, primarily focusing on dynamically optimizing the allocation of denoising steps according to resource availability and performance constraints. However, despite such cross-layer collaboration, each request is still processed independently. Under high-concurrency conditions, this can cause redundant computation across tasks and limit overall system resource efficiency (e.g., energy consumption and GPU utilization).

To further improve efficiency, researchers recently proposed cross-task collaborative inference [10] to exploit the interdependencies between tasks. In this approach, the whole inference process is usually divided into two phases: a *shared inference phase* to construct common characteristics for each user cluster (e.g., early denoising steps of the inference process), followed by a *personalized inference phase* to refine outputs according to individual needs. In such an inference paradigm, it is crucial to design the shared and personalized inference phases properly. Du *et al.* [11] randomly selected a user-provided prompt as a public prompt for shared inference. In contrast, Xie *et al.* [12] performed prompt clustering based on entity information and selected a random prompt from each cluster to generate shared intermediate results. Moreover, Zhuang *et al.* [13] proposed a dynamic sharing mechanism, allowing individual users to select the most suitable intermediate result for personalization. However, these approaches fail to exploit semantic consistency as the fundamental driver for cross-task computational reusability. As a result, the generated intermediate results may not align well with all users' expectations, thereby diminishing the effectiveness of personalization.

In this article, we first provide a comprehensive review of the three inference paradigms mentioned above and summarize their limitations. Motivated by these observations, we propose a semantic-aware cross-task collaborative inference (SemCT) framework that introduces semantic awareness to improve service performance while reducing resource consumption. Unlike random public-prompt selection [11] and entity-level clustering [12, 13], SemCT performs fine-grained clustering based on the semantic similarity of user requests, with the goal of maximizing within-cluster reuse of shared intermediate results. Guided by cluster semantics, SemCT further adaptively allocates denoising steps and edge resources between shared and personalized inference for better overall inference performance.

Our contributions are as follows:

- **Novel Taxonomy of Inference Paradigms:** We provide an in-depth analysis and systematic classification of inference strategies across independent, cross-layer, and cross-task paradigms. By evaluating their building designs, we reveal key strengths, limitations, and trade-offs between performance and efficiency, offering critical insights for developing more intelligent AIGC services.
- **Semantic-Aware Collaborative Framework:** We present SemCT, a framework that leverages the critical correlation between prompt semantic similarity and the reusability of shared inference. SemCT
  - –Formulates a fine-grained clustering objective that transforms semantic information into quantifiable metrics for computation sharing
  - –Optimizes the division of denoising steps between shared and personalized inference to balance latency, energy, and content quality.
- **Case Study and Future Directions:** We implement SemCT in simulations using real data, demonstrating significant improvements in both content quality and resource efficiency compared with existing frameworks. We also discuss practical challenges and outline future research directions to advance collaborative inference, paving the way for scalable, efficient, and intelligent distributed AI systems.

The Evolution of Inference Paradigms: From Independence to Collaboration

As illustrated in Fig. 1, inference strategies can be broadly categorized into three paradigms:
1. Independent inference,
2. Cross-layer collaborative inference, and
3. Cross-task collaborative inference.

Each paradigm represents a distinct trade-off between inference performance and resource efficiency. In this section, we systematically review the core principles and technical underpinnings of both independent and collaborative inference, while critically assessing their respective contributions and inherent limitations.

Independent Inference

As shown in Fig. 1 (Part A, Left), independent inference is a self-contained paradigm in which each inference task can be performed on a single computing node (e.g., a cloud server, edge server, or end device), but cannot be separated and distributed to multiple computing nodes. This paradigm serves as a foundational baseline, offering flexible and straightforward control, but it suffers from limited scalability in resource-constrained environments.

Prior studies have attempted to enhance system performance from three primary perspectives.

1. *Stage-wise parameter adjustment approach* dynamically selects models of different sizes for distinct denoising stages: large models are employed in early stages to build accurate semantic representations, while lightweight models accelerate later iterations, trading a marginal quality loss for significant resource savings [3]. These techniques, however, optimize only the model-internal schedule and remain oblivious to time-varying network conditions.
2. *Dynamic load-balancing approach* copes with multi-user edge scenarios by continuously monitoring computational and communica-

This paradigm serves as a foundational baseline, offering flexible and straightforward control, but it suffers from limited scalability in resource-constrained environments.

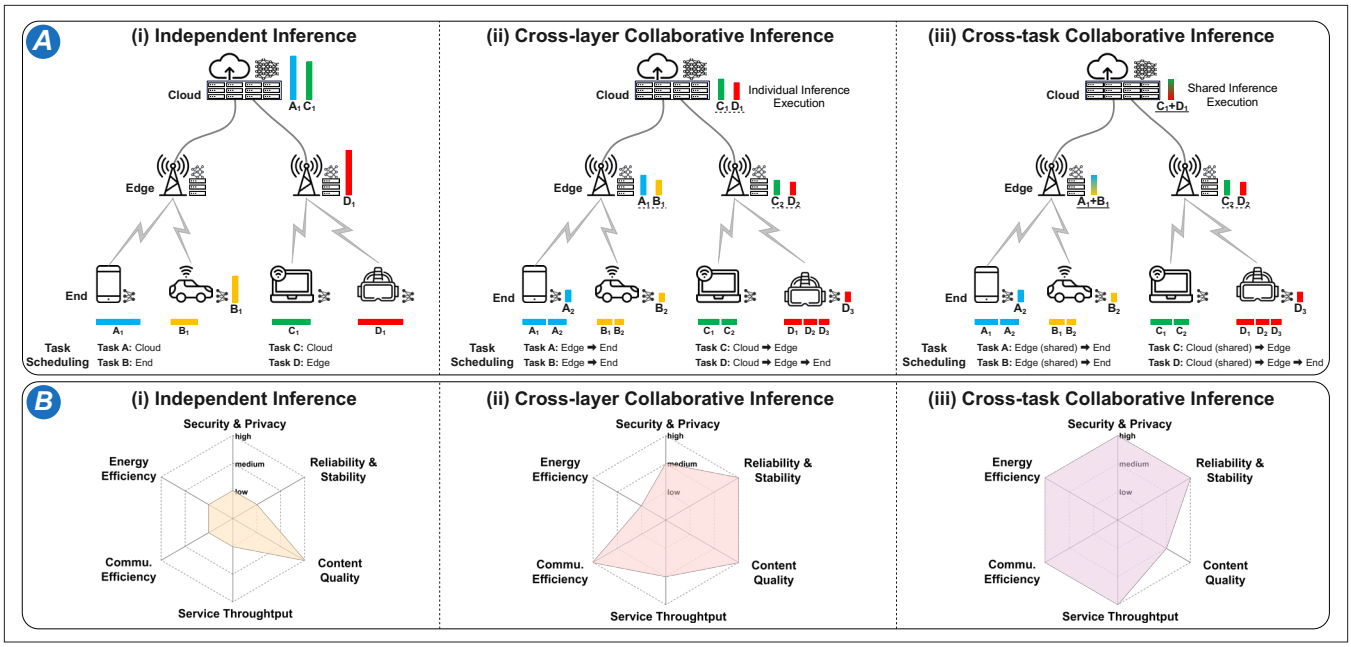

FIGURE 1. Workflow comparison (**Part A**) and key performance indicator (KPI) radar charts (**Part B**) of three inference paradigms: (i) independent inference, (ii) cross-layer collaborative inference, and (iii) cross-task collaborative inference. For the cross-task collaborative inference paradigm, an example is shown where tasks A ("A fluffy white cat with blue eyes sitting on a windowsill.") and B ("A gray cat with green eyes sitting on a wooden porch.") share similar semantics; the edge server performs shared denoising steps (A1+B1) for their common intents, and each end device continues local refinement (A2 for white cat and B2 for gray cat).

tion resources, then dispatching inference tasks to the most suitable edge servers [4]. This alleviates instantaneous bottlenecks and boosts both aggregated throughput and per-user experience, yet treats each request independently.

3. *Inference-step reduction approach* minimizes the total number of denoising steps per task through adaptive-scheduling strategies [5, 6]. When jointly combined with load-balancing decisions, this approach further reduces computational overhead while preserving acceptable content quality.

Although these strategies improve the efficiency of independent inference, they remain fundamentally constrained by the isolation of single-node computation. The lack of resource integration across different computing nodes imposes significant limitations in service throughput, energy efficiency, and other performance dimensions, making it difficult to scale under high-concurrency scenarios.

Cross-Layer Collaborative Inference

As shown in Fig. 1 (Part A, Center), cross-layer collaborative inference is a distributed paradigm in which the inference process (i.e., denoising steps) of each task can be separated into multiple parts and strategically allocated to different computing nodes.

Existing studies have explored various collaboration architectures in such an inference paradigm, including *cloud-edge collaboration* [2, 7], *edge-end collaboration* [8], and *end-edge collaboration* architectures [9]. For example, Yan *et al.* [2] proposed Hybrid SD, a cloud-edge collaboration architecture in which a heavyweight cloud model performs early denoising for high-level semantic planning. In contrast, a lightweight edge model finishes late visual refinement, ensuring both content quality and resource efficiency. To better adapt to heterogeneous resources and network dynamics, Zhuang *et al.* [7] introduced a joint optimization of model inference and task offloading, where scheduling decisions are guided by multi-dimensional quality metrics under delay and energy constraints. However, cloud-edge designs inevitably involve frequent data exchanges between the cloud and the edge, which introduces higher network latency and poses greater privacy leakage risks. To address these concerns, Yang *et al.* [8] explored a framework in which inference is divided between the edge server and the end device through a tunable split point. In this design, the edge runs a shared model to extract coarse semantic features, while the end device employs a personalized model to refine the output. In contrast, Feng *et al.* [9] proposed a framework where inference begins on the end device with preliminary denoising, and the intermediate result is then transmitted to the edge for further computation.

By distributing inference workloads across heterogeneous computing nodes, such a vertical collaboration overcomes the limitations of independent inference and significantly improves communication efficiency, reliability, and scalability.

Cross-Task Collaborative Inference

As illustrated in Fig. 1 (Part A, Right), cross-task collaborative inference reduces computational cost by uncovering and exploiting shared intent across different users. The key observation is that many personal prompts (i.e., user-provided prompts) contain a reusable semantic core. Rather than processing each prompt in isolation, this paradigm first groups personal prompts based on semantic similarity to derive a public prompt. It then performs a shared inference phase (based on the public prompt) to produce intermediate results, followed by a personalized inference phase (based on personal prompts) that adapts these shared results to each user's specific intent.

This design strikes an effective balance between efficiency and personalization.

Cross-task collaboration has evolved from coarse-grained static frameworks to fine-grained dynamic optimizations. An early study [11] adopted *public prompt selection and uniform sharing*. In such a framework, a personal prompt is randomly selected as the public one, and all users share the same intermediate result from a fixed split point (e.g., step 10). Although simple to implement, this method fails to capture differences across user intents, resulting in degraded content quality. To improve sharing accuracy, Xie *et al.* [12] proposed an entity-level prompt clustering method that employs classification models (e.g., CNNs) to extract key entities and relationships from personal prompts. Prompts with the same entities are grouped into a cluster, and each cluster reuses the same intermediate result from a randomly selected public prompt. However, this approach oversimplifies the diversity of user intents by using the same intermediate result, making it difficult to meet the personalized needs of all users.

Building on this, Zhuang *et al.* [13] introduced *public prompt generation and dynamic sharing*, which enables a more flexible and cost-effective sharing strategy. It generates a high-level public prompt for each cluster and allows each user to select the most suitable intermediate result based on the similarity between personal and public prompts. For instance, some tasks may reuse the intermediate result from denoising step 9, while others may start from step 11. This fine-grained and task-adaptive mechanism significantly improves resource efficiency while enhancing system flexibility. As a complementary approach to cross-task collaboration, Wang *et al.* proposed a *cache-based mechanism* [14] that precomputes and stores a large set of intermediate results on cloud or edge servers. Users retrieve the most relevant entry based on their personal prompts and directly reuse cached intermediate results to continue denoising. This method reduces the latency and computational load. However, maintaining a large and frequently updated cache incurs considerable storage and management overhead, especially in high-concurrency or memory-constrained environments.

By leveraging similarity in user intent across tasks, cross-task collaborative inference enables shared computation. This approach maintains personalized generation while significantly reducing redundant denoising steps and enhancing overall efficiency. However, since most existing cross-task frameworks perform clustering primarily at the entity level, their granularity remains too coarse to preserve fine-grained user intents, leaving room for semantic-aware refinement.

Performance Comparison of Three Inference Paradigms

As shown in Fig. 1 (Part B), the three inference paradigms exhibit distinct performance characteristics. Independent inference performs the whole generation pipeline for each request, ensuring high content quality. However, the lack of collaboration across computing nodes and tasks leads to low communication efficiency and high energy costs. Under high concurrency, computation and bandwidth demands scale with the number of tasks, which may cause request queuing and unstable service, ultimately degrading service reliability. Moreover, in cloud-hosted independent inference, transmitting generated content between the cloud and the edge exposes a significant attack surface, resulting in weak privacy protection.

Cross-layer collaborative inference distributes denoising steps across multiple computing nodes. By balancing the workload, it improves service throughput and communication efficiency, enabling higher reliability under concurrent requests. Moreover, keeping the final denoising step on the end device prevents leaked processed content, thereby providing a medium level of privacy protection. However, because each task is still executed independently, computational redundancy remains unaddressed, leading to limited energy efficiency gains. This redundancy becomes a primary energy bottleneck, revealing the need for semantic-aware methods that can reuse intermediate results across tasks.

Cross-task collaborative inference further reduces computation overhead through semantic aggregation and reuse of intermediate results. This allows the system to sustain high throughput and reliability even under heavy concurrency. By sharing only intermediate results derived from public prompts and storing final outputs locally, it achieves the strongest privacy protection among the three paradigms. However, the shared intermediate results limit fine-grained modeling of individual user intent, which can reduce personalization and lead to a moderate decline in content quality.

By balancing the workload, it improves service throughput and communication efficiency, enabling higher reliability under concurrent requests.

SemCT: Semantic-Aware Cross-task Collaborative Inference Framework

To overcome the semantic limitations of entity-level prompt clustering, we propose **SemCT**, a novel **Semantic-aware Cross-task Collaborative Inference** framework. By integrating deep semantic understanding into the collaborative inference process, SemCT enables more accurate prompt clustering and smarter reuse of intermediate results, achieving a better balance between efficiency and personalization.

Workflow of the SemCT Framework

Figure 2 illustrates the SemCT framework under a general network architecture, which consists of the following steps:

Step 1: Semantic-level Prompt Clustering and Public Prompt Generation: Since the public prompt is unknown in advance, directly maximizing public-personal prompt similarity is non-trivial. To address this, we reformulate the objective as maximizing intra-cluster similarity (i.e., the average pairwise similarity among personal prompts within each cluster), which is empirically shown to be strongly (approximately linearly) correlated with the resulting public-personal prompt similarity. Based on this reformulation, we adopt hierarchical agglomerative clustering (HAC) as an efficient clustering heuristic, and then extract shared semantic features from each cluster to generate a representative public prompt.

Step 2: Shared Inference: During the shared inference phase, the GDM generates a series of intermediate results based on each public prompt. These results encode the general semantic characteristics and are reused for all users within the cluster. By merging initial inference across cluster members, this process reduces computational

The network consists of an edge server equipped with an industrial-grade NVIDIA A100 GPU cluster and 50 users, each with a consumer-grade NVIDIA RTX 3090 Ti GPU, uniformly distributed within a radius of 50 to 300 meters from the edge server.

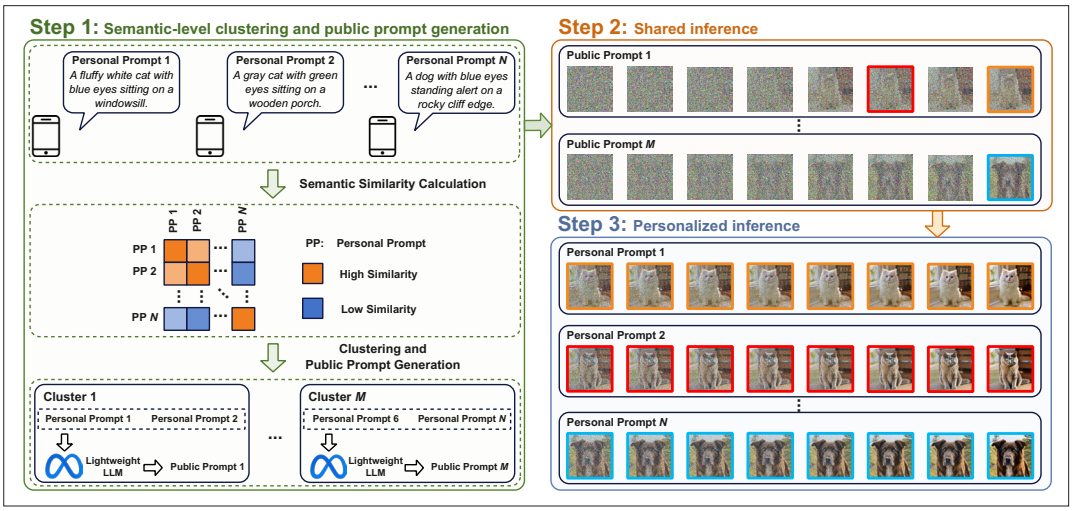

FIGURE 2. Workflow of the SemCT framework. **Step 1:** Cluster personal prompts and construct a high-level public prompt for each cluster. **Step 2:** Generate and share intermediate results from public prompts to reduce computational overhead. **Step 3:** Personalize the shared results with each personal prompt to generate user-specific content.

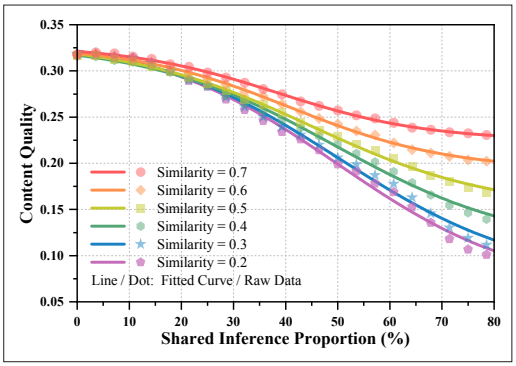

FIGURE 3. Effect of shared inference proportion on content quality across different semantic similarities.

overhead while establishing a semantic foundation for personalized inference.

Step 3: Personalized Inference: In the personalized inference phase, the system dynamically selects the most appropriate intermediate result for each user and combines it with the personal prompt for personalization. This process ensures efficient and customized content generation, producing final outputs that align with individual user needs.

With this three-step workflow, the system can reduce computational overhead while preserving task-specific accuracy. However, increasing the shared inference proportion introduces a trade-off that may degrade the quality of personalized content.

General Model of Content Quality

In SemCT, we define content quality as the combination of subjective quality (i.e., the semantic alignment between generated content and personal prompt) and objective quality (i.e., fidelity and aesthetic appeal). Theoretically, the content quality is affected by two factors:

1. Shared inference proportion and
2. Semantic similarity between personal and public prompts.

To investigate the impact of these two factors, we conducted extensive experiments based on state-of-the-art GDMs (e.g., Stable Diffusion 3 Medium). Our findings reveal that these factors primarily influence subjective quality, while exhibiting negligible effects on objective quality. Therefore, focusing on the subjective dimension, we employ the pre-trained ViT-L/14@336px CLIP model to quantify the content quality.

As shown in Fig. 3, we apply a sigmoid-based curve fitting method to capture the non-linear impact of shared inference proportion on content quality. By analyzing the fitted curves, we derive the following two key findings:[1]

1. As the shared inference proportion increases, the content quality tends to decline; moreover, the lower the semantic similarity, the faster this degradation occurs.
2. For a fixed shared inference proportion, higher semantic similarity leads to improved content quality, with more gains observed as similarity increases.

The modeling results underscore the crucial role of semantic alignment and the benefits of fine-grained prompt clustering, providing empirical guidance for optimizing model inference and resource allocation in MEC-empowered AIGC services.

Case Study

In this section, we present a case study to evaluate the performance of SemCT in an MEC network. The network consists of an edge server equipped with an industrial-grade NVIDIA A100 GPU cluster and 50 users, each with a consumer-grade NVIDIA RTX 3090 Ti GPU, uniformly distributed within a radius of 50 to 300 meters from the edge server. The number of clusters K balances intra-cluster similarity and clustering overhead. As shown in Fig. 4, too small K decreases intra-cluster similarity and limits sharing gains, while too large K increases overhead. Since the utility is near-optimal for moderate K, we set $K = 3$ to achieve high utility with low overhead. To enable effective cross-task collaboration, identical GDMs are deployed on both the edge server and the end devices.

The collaborative process begins with each user uploading a personal prompt to the edge server. To cluster users for shared inference, the edge server (Step 1) computes a prompt similarity matrix using a lightweight encoder (e.g., all-MiniLM-L6-v2 [15]). It then applies aver-

age-linkage HAC: starting from individual clusters, it iteratively merges the pair of clusters with the highest average inter-cluster similarity until K clusters remain. Once clustering is complete, the edge server uses a lightweight LLM (e.g., Llama-3-8B) to extract semantic features from each cluster and generate a high-level public prompt that captures common intents within each cluster.

With the public prompts established, the system proceeds to the shared and personalized inference phases (Steps 2 & 3), guided by a coordination mechanism that optimizes the overall network utility. We define the utility as a weighted composite metric that balances three critical factors: content quality, service latency, and energy consumption, providing a holistic assessment of AIGC service performance from both technical and user-experience perspectives. This design allows us to flexibly reflect the relative importance of each factor (e.g., prioritizing quality in user-centric scenarios or energy efficiency in resource-constrained edge environments).

To maximize network utility, the system optimizes the trade-off between shared and personalized inference. Leveraging the edge server's capability to execute both shared inference (cluster-level) and personalized inference (task-level), it dynamically adjusts four key variables:

1. *The number of shared denoising steps executed on the edge server;*
2. *The number of personalized denoising steps executed on the edge server;*
3. *The number of personalized denoising steps performed locally on each end device;*
4. *The allocation of the edge server's computational resources across clusters.*

This joint optimization enables SemCT to better balance content quality, latency, and energy consumption under heterogeneous tasks and device capabilities.

To address the optimization problem characterized by multiple interdependent decision variables, we propose a two-tier framework that employs DRL for complex inference scheduling and efficient optimization for resource allocation, enabling scalable and practical decision making. Specifically, the first tier employs the edge server as a centralized controller, which collects user observations in parallel and applies a DRL-based independent proximal policy optimization (IPPO) algorithm [13] to determine optimal inference decisions, i.e., decision variables 1–3. In the second tier, we utilize convex-based optimization techniques to derive a closed-form resource allocation strategy [7] by relaxing the decision variable 4. This enables the edge server to distribute its computing resources among clusters according to the inference decisions made in the first tier, achieving a balanced trade-off among content quality, service latency, and energy consumption.

Simulation results show that the SemCT framework achieves a better trade-off compared with the following three representative benchmarks:

- Semantic-Free Inference Framework (SFIF) [10, 12]: This framework leverages entity-level prompt clustering. Within each cluster, a randomly selected personal prompt is designated as the public prompt.
- Clustering-Free Inference Framework (CFIF) [11]: This framework treats all prompts as a single cluster, from which a personal prompt is randomly selected as the public one.
- Cross-Layer Inference Framework (CLIF) [7–9]: This framework is designed without shared inference.

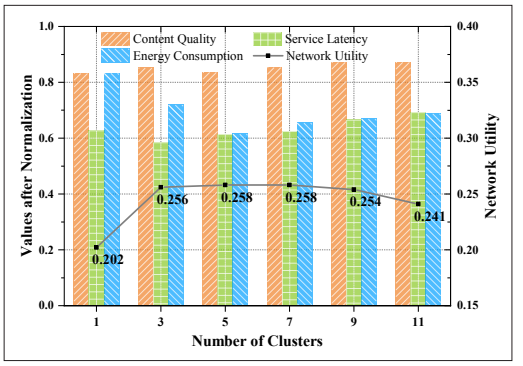

FIGURE 4. Impact of the number of clusters on network utility.

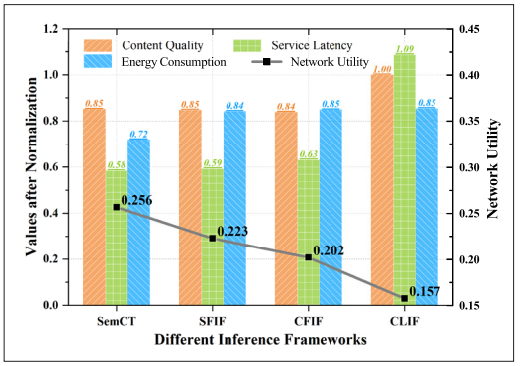

FIGURE 5. Values of content quality, service latency, energy consumption, and network utility under different inference frameworks.

As illustrated in Fig. 5, the proposed SemCT framework demonstrates a superior trade-off among content quality, service latency, and energy consumption. Moreover, it improves network utility by 14.8%, 26.7%, and 63.1% compared to SFIF, CFIF, and CLIF, respectively. Notably, SemCT not only provides content quality comparable to SFIF and CFIF, but also achieves lower service latency and energy consumption. This performance gain is supported by the semantic-level clustering approach, which allows for more accurate prompt clustering based on their underlying intents. In addition, as summarized in Table 1, SemCT further reduces the computational overhead to 9792 TFLOPs, corresponding to reductions of 13.5% against SFIF, 21.5% against CFIF, and 41.7% against CLIF. This advantage stems from its efficient utilization of shared inference, which enables a larger number of denoising steps to be executed in the shared inference phase without compromising content quality. Consequently, the overall inference workload is alleviated, leading to faster response times and lower energy consumption.

Open Challenges and Future Work

Caching for Asynchronous Collaborative Inference

During collaborative inference processes, caching intermediate results effectively reduces redundant computation and significantly enhances resource efficiency. Existing research predominantly focuses on synchronous collaboration, where the

To address the optimization problem characterized by multiple interdependent decision variables, we propose a two-tier framework that employs DRL for complex inference scheduling and efficient optimization for resource allocation, enabling scalable and practical decision making.

[1] Our code is available at: https://github.com/iimxinyi/SemCT-Quality

	SemCT	SFIF	CFIF	CLIF
Computational Overhead (Required TFLOPs)	9792	11316	12468	16800

TABLE 1. Computational overhead across inference frameworks.

design of caching strategies is relatively straightforward, leveraging the latest intermediate results for efficient reuse and updates. In asynchronous collaborative environments, users often progress at different speeds and request different outputs, leading to intermediate results that vary in computation stages and timeliness, which makes traditional caching mechanisms inadequate. To address this challenge, caching must evolve from passive storage into proactive schedulers with predictive and decision-making capabilities. Such mechanisms should carefully balance computational cost, data volume, and access probability to select cached content dynamically. Furthermore, under resource constraints, systems need to adjust cache priorities and employ intelligent eviction policies to optimize global computation-communication trade-offs, thereby ensuring high throughput and low latency for asynchronous inference tasks.

Cross-Domain and Heterogeneous Model Collaboration

Most existing collaborative inference frameworks rely on homogeneous models, which share the same architecture but differ in scales, such as varying parameter sizes. This uniformity simplifies task coordination and ensures predictable performance. However, such deployments cannot keep up with advances in specialized models or leverage the growing variety of domain-specific heterogeneous models. Future collaborative inference frameworks should incorporate heterogeneous models to unleash the full potential of collaboration. The key is enabling these models to "understand" each other. Semantic alignment techniques can map intermediate representations from different models into a shared semantic space, making cross-model information compatible and usable. Building upon this, the inference process can be strategically divided into a collective semantic-planning stage, handled by a general-purpose large model, and a subsequent fidelity-improving stage, executed by heterogeneous domain-specific smaller models. By harnessing advances across various domain-specific models, the system can achieve unparalleled flexibility, scalability, and performance, paving the way for truly global and all-domain AI applications.

Shared Inference for Transformer-Based Models

Beyond diffusion models, a promising direction is to extend collaborative inference to other generative AI architectures, such as transformer-based text generation. Unlike image denoising, text generation exhibits high diversity and strong context dependence, which makes it more challenging to design shared computation without sacrificing personalization. Meanwhile, multimodal generative models (e.g., text-to-video, speech-to-text) are rapidly emerging, and exploring cross-modal collaborative inference remains an attractive yet underexplored challenge. A potential direction is to investigate shared-backbone layer-wise strategies: earlier layers capture general semantic representations that can be shared across users, while deeper layers maintain task-specific personalization through lightweight adapters (e.g., LoRA) or dedicated heads. Combined with dynamic KV-cache management, parameter factorization, and knowledge distillation, such designs may offer a viable path to striking a balance between efficiency and quality.

Conclusion

In this article, we have investigated collaborative inference for GDMs in MEC networks. First, we have reviewed the architecture and evolution of collaborative inference tailored for GDMs, tracing its progression from independent inference to cross-layer and cross-task collaboration, and highlighting the advantages and limitations of each paradigm. Building on these insights, we have developed SemCT, a semantic-aware framework designed to exploit the full potential of collaboration. Then, we have presented a case study to validate its effectiveness and offer practical insights for efficient AIGC service delivery. Finally, we have discussed open issues and outlined future research directions. As AIGC continues to expand its role in digital life, developing efficient and scalable inference strategies will be critical to meeting the rapidly growing and diverse user demands. We hope this work serves as both a reference and an inspiration for further advancements in collaborative inference for next-generation AIGC services.

Acknowledgment

This work was supported in part by the Natural Science Foundation of Guangdong Province (Grant No. 2024A1515010178), in part by the Shenzhen Science and Technology Program (Grant No. KQTD20190929172545139, GXWD20231129103946001, and KJZD20240903095402004), and in part by the Research Matching Grant Scheme from the Research Grants Council of Hong Kong.

References

[1] M. Xu *et al.*, "Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, 2nd Quarter 2024, pp. 1127–70.
[2] C. Yan *et al.*, "Hybrid SD: Edge-Cloud Collaborative Inference for Stable Diffusion Models," 2024, arXiv:2408.06646.
[3] S. Yang *et al.*, "Denoising Diffusion Step-Aware Models," *Proc. ICLR*, May 2024.
[4] H. Du *et al.*, "Diffusion-Based Reinforcement Learning for Edge-Enabled AI-Generated Content Services," *IEEE Trans. Mobile Comput.*, vol. 23, no. 9, Sept. 2024, pp. 8902–18.
[5] H. Liu *et al.*, "Joint Communication and Computation Scheduling for MEC-Enabled AIGC Services Based on Generative Diffusion Model," *Proc. WiOpt*, Oct. 2024.
[6] Y. Wang, C. Liu, and J. Zhao, "Offloading and Quality Control for AI Generated Content Services in 6G Mobile Edge Computing Networks," *Proc. IEEE VTC*, June 2024.
[7] X. Zhuang, J. Wu *et al.*, "Joint Optimization of Model Inferencing and Task Offloading for MEC-Empowered Large Vision Model Services," *Proc. IEEE INFOCOM*, May 2025.
[8] W. Yang *et al.*, "Efficient Multi-User Offloading of Personalized Diffusion Models: A DRL-Convex Hybrid Solution," *IEEE Trans. Mobile Comput.*, vol. 24, no. 9, Sept. 2025, pp. 9092–9109.
[9] W. Feng *et al.*, "Exploring Collaborative Diffusion Model Inferring for AIGC-Enabled Edge Services," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 2, Apr. 2025, pp. 946–60.
[10] H. Du *et al.*, "Exploring Collaborative Distributed Diffusion-Based AI-Generated Content (AIGC) in Wireless Networks," *IEEE Networtk*, vol. 38, no. 3, May 2024, pp. 178–86.
[11] H. Du *et al.*, "Reinforcement Learning with LLMs Interaction for Distributed Diffusion Model Services," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, Oct. 2025, pp. 8838–55.
[12] G. Xie *et al.*, "GAI-IoV: Bridging Generative AI and Vehicular

Networks for Ubiquitous Edge Intelligence," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, Oct. 2024, pp. 12,799–12,814.
[13] X. Zhuang *et al.*, "QoS-Driven Hybrid Inference Scheme for Generative Diffusion Models in MEC-Enabled AI-Generated Content Networks," *Proc. IEEE ICC*, June 2025.
[14] H. Wang *et al.*, "SC-TSDRL: A Cloud-Edge Collaboration Framework for Diffusion Model Inference Acceleration," *Proc. IFIP NPC*, Dec. 2024.
[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," *Proc. EMNLP*, Nov. 2019.

Biographies

HONGJIA WU (whongjia@eduhk.hk) is a Postdoctoral Fellow with the Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China. Her research interests include edge intelligence, game theory, and Internet of Things.

XINYI ZHUANG [GSM] (zhuangxinyi@stu.hit.edu.cn) is pursuing his Ph.D. degree with the School of Electronics and Information Engineering, Harbin Institute of Technology, Shenzhen, China. His research interests include networks for large AI models and AI for networking.

JIAQI WU (wjqeasy@163.com) is a Lecturer with the School of Computer Science, Guangdong University of Finance. His research interests include network optimization, deep reinforcement learning, edge intelligence, and generative AI.

LIN GAO [SM] (gaol@hit.edu.cn) is a Professor at the School of Electronics and Information Engineering, Harbin Institute of Technology, Shenzhen, China. His main research interests are in the interdisciplinary area between game theory, network optimization, and artificial intelligence.

DUSIT NIYATO [F] (dniyato@ntu.edu.sg) is a Professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. His research interests are in the areas of sustainability, edge intelligence, decentralized machine learning, and incentive mechanism design.

TSE-TIN CHAN [M] (tsetinchan@eduhk.hk) is an Assistant Professor with the Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China. His research interests include wireless communications and networking, age of information (AoI), and artificial intelligence (AI)-native wireless communications.