EcoGen: Fusing Generative AI and Edge Intelligence for Sustainable Scalability

SAI HARVIN KUSUMARAJU[1], ARYA SUNEESH [2], AASTHA RANA [3], SRIHARSHA BODICHERLA[4], and BHAUMIK TYAGI[5]

[1,2,4] *Undergraduate Student, IIIT, Kottayam, India*
[3]*Undergraduate Student, GGSIPU, Delhi, India*
[5]*Jr. Research Scientist, Delhi, India*

Abstract—The accelerating advancements in Generative Artificial Intelligence (GenAI) have led to an unprecedented surge in data creation on the Internet, posing challenges to current computing and communication frameworks. GenAI, a distinct category of AI, generates content akin to human creations. Currently, GenAI services heavily rely on traditional cloud computing, resulting in high latency due to data transmission and a surge in requests. In response, the integration of edge-cloud computing emerges as an attractive paradigm, offering computation power and low latency through collaborative systems. This research paper provides a comprehensive overview of the intersection between GenAI and edge-cloud computing. We delve into recent developments in both domains and examine technical challenges through the lens of two exemplary GenAI applications. Introducing an innovative solution, we propose the Generative AI-oriented synthetical network (EcoGen), a collaborative cloud-edge-end intelligence framework. EcoGen facilitates bidirectional knowledge flow, allowing GenAI's pre-training to provide foundational knowledge for Edge Intelligence (EI), while EI aggregates personalized knowledge for GenAI. The framework leverages data-free knowledge relay to buffer contradictions, enabling virtuous-cycle model fine-tuning and task inference. Importantly, we incorporate a detailed analysis of the energy efficiency and environmental sustainability aspects of deploying Generative AI systems at scale, particularly in edge computing. Strategies to optimize energy consumption and reduce the carbon footprint are explored, contributing to a more sustainable AI ecosystem. Experimental results demonstrate the effectiveness of EcoGen in achieving seamless fusion and collaborative evolution between GenAI and EI. The paper concludes by outlining design considerations for training and deploying GenAI systems at scale and pointing towards future research directions, emphasizing the imperative of sustainable AI practices.

Index Terms—Generative Artificial Intelligence, Energy Efficiency, Edge-Cloud Computing, Hybrid Federated Split Learning, Integrated Fine-tuning.

I. INTRODUCTION

The field of Generative AI (Gen AI) has surfaced as an innovative domain aimed at achieving Artificial General Intelligence (AGI) through the fusion of machine learning techniques and creative content generation methods. Gen AI represents a specific branch of AI that seeks to autonomously produce novel content across various mediums, such as images, audio, text, and even three-dimensional objects, mirroring human-generated content [1]. The rapid advancement of Gen AI has led to the proposal and widespread adoption of diverse applications, including text-to-image generation, text-to-speech synthesis (TTS), chatbots, and AI-powered virtual reality experiences [2,3]. However, the extensive size of most Gen AI models and their high computational requirements necessitate a robust centralized computing infrastructure, typically in the form of cloud servers, to handle user requests. Consequently, users may encounter significant delays in processing, particularly during periods of high traffic volume. Such limitations pose challenges for deploying Gen AI models in applications that demand low-latency responses at scale. Additionally, the substantial energy consumption associated with centralized cloud computing raises concerns about environmental sustainability and cost efficiency. In response to these challenges, the proliferation of mobile devices and the exponential growth of data-intensive applications have catalyzed the development of edge-cloud computing solutions. Edge-cloud computing leverages the computational power of cloud servers and the efficient data management and communication capabilities of edge servers, offering a promising solution for consumer-oriented AI applications and edge intelligence. Notably, this approach enables the deployment of large AI models within the edge-cloud computing ecosystem [4]. In contrast to traditional cloud computing, which primarily focuses on computation within cloud servers, and multi-access edge computing (MEC), which prioritizes computation at the edge servers, edge-cloud computing capitalizes on the collaborative potential between cloud and edge resources. By harnessing a combination of cloud and edge resources, edge-cloud computing can effectively utilize computational resources while minimizing latency, thus enhancing the performance and scalability of Gen AI applications.

In contrast to the centralized approach of Generative Artificial Intelligence (GAI) utilizing large-scale parameters, Edge Intelligence (EI) tends to deploy adaptable lightweight models closer to end-users, thus aligning computational intelligence with distributed terminal data [19]. A recent analysis forecasts a surge in Internet of Things (IoT) devices, estimating 30.9 billion connections by 2025, with data volume projected to soar to nearly 79.4 Zettabytes (ZB) [20]. Moreover, ongoing advancements in chip integration technologies, including Central Processing Units (CPUs) and Graphics Processing Units (GPUs) within mobile terminals, continually reduce power consumption costs, empowering terminal devices with sufficient computing capabilities for lightweight AI model training and inference [21]. Consequently, with the proliferation of edge data and the augmentation of computing power in terminal devices, the forthcoming B5G and 6G era is poised to witness a paradigm shift in network architecture, transitioning from the Internet of Everything to the Intelligence of Everything. In this evolution, native AI functionalities will migrate from remote cloud servers to the network edge [22]. Nevertheless, despite EI's proximity to terminals, the constrained model scale results in a dearth of prior knowledge, undermining the effectiveness of edge training and inference. Consequently, the combination of GAI and EI offers synergistic advantages, fostering new avenues for growth. By leveraging personalized knowledge through fragmented computing resources, EI can mitigate GAI's challenge of limited public data availability and bridge the gap between computational intelligence and end-users. Simultaneously, GAI can furnish EI with pre-trained foundational knowledge, serving as a robust baseline to expedite learning convergence and enhance inference performance [23]-[32].

II. RELATED WORK

The rise of ChatGPT has sparked widespread interest in Generative AI (GenAI), an AI technology capable of producing diverse multimedia content [5]. The historical progression of generative AI can be delineated into three distinct phases: 1) the Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) era spanning from 2014 to 2017, 2) the Transformer era from 2018 to 2019, and 3) the current epoch characterized by the dominance of large-scale models from 2020 onwards [6]. The inception of the Variational Autoencoder (VAE) dates back to its proposal in [7], subsequently spawning various iterations aimed at enhancing content quality, accommodating different levels of supervision, and refining inference efficiency [8,9,10]. VAEs operate as probabilistic generative models, featuring an encoder-decoder architecture where the encoder maps inputs to vectors in a latent space, and the decoder reconstructs latent vectors into outputs within the input space. During training, network parameters are optimized to minimize the discrepancy between generated and input data. Introducing noise to latent vectors enables the decoder to produce multiple output samples with distributions akin to input samples. Similarly, Generative Adversarial Networks (GANs) employ dual networks during training while retaining only one during inference [11,12]. Consisting of a generator and a discriminator, GANs undergo a training regimen where the generator progressively generates synthetic data mirroring real data distributions to deceive the discriminator. Conversely, the discriminator aims to discern between genuine and synthetic data instances as effectively as possible [13].

Models for Natural Language Generation (NLG) aim to produce text responses that mimic human-like language. These models find applications in various domains such as neural machine translation, question answering, and document summarization [14]. They are often referred to as Language Models (LMs) [15]. In recent times, transformers equipped with self-attention mechanisms have emerged as a groundbreaking advancement in constructing robust LMs [16]. Transformers have supplanted long short-term memory (LSTM) networks as the preferred architecture for LMs, ushering in a new era of Large Language Models (LLMs) [17]. While the encoder component of transformers employs bidirectional information propagation to comprehend input text, the decoder, prevalent in most transformer architectures, generates words sequentially. This type of decoder is commonly known as an autoregressive decoder. With the rise of transformers, there is a trend towards increasingly larger generative models. In the past couple of years, efforts have been directed toward amalgamating diverse models to create even larger and more potent variants.

Table 1. Evaluation of performance provisions of 3 computational resources, namely user devices, edge servers, & cloud servers.

Resources	User Devices	Cloud Servers	Edge Servers
Memory	>24TB	~500GB	<64GB
Dist Storage	>25PB	<1PB	<10TB
Latency (RTTs)	30 ~ 50 ms	<10ms	-
Power (per year)	>2,000TWh	~7,500KWh	~600KWh

Concurrent Connections	>500,000	~1,000	1

With abundant computational resources, cloud servers are capable of storing and executing large-scale models to handle complex tasks efficiently. Conversely, edge devices typically handle lower-level preprocessing tasks. The advent of 5G/IoT marks a new phase for Artificial Intelligence and Generative Computing (AIGC). It is no longer adequate to centralize all computations and data storage solely within cloud servers or data centers. Similarly, relying solely on AI computation with edge servers and user devices proves impractical as AIGC data expands rapidly. Deploying large deep-learning AI models at the edges presents challenges. Recently, an alternative approach known as green learning methodology [18] has been proposed as a substitute for deep learning. Green learning AI models feature smaller sizes, reduced computational complexity measured in FLOPs (Floating Point Operations), faster inference times, and lower power consumption demands. Consequently, green-learning AI models pave the way for edge servers and even user devices to offload tasks from cloud servers, opening new avenues for collaboration between edge and cloud servers. Hybrid solutions that combine deep and green learning align well with the edge-cloud computing paradigm. GenAI, with its unique mission to process low-level data and aggregate high-level abstractions to generate creative content, stands to benefit greatly from the collaboration between edge and cloud servers. Recently, Meta unveiled a supercomputing cluster boasting extensive computational resources, capable of achieving five exaflops (five billion calculations per second) using 16,000 Nvidia A100 GPUs for training cutting-edge GenAI models. These servers are interconnected via an NVIDIA Quantum InfiniBand fabric network with a bandwidth of 16 Tb/s, ensuring minimal latency in data synchronization. However, such computational scale remains beyond the reach of most companies and academic institutions. Consequently, there's a growing interest in designing scalable GenAI systems utilizing reasonable computing clusters to perform similar tasks. Edge-cloud computing holds promise in this regard, as it can leverage expandable computational resources that are underutilized and situated closer to users.

III. RESEARCH METHODOLOGY

The concept of parameter-efficient fine-tuning has garnered significant attention in research, particularly concerning the efficient retraining of domain-specific models. Initially, we delve into the practice of parameter-efficient fine-tuning through prompt tuning. Subsequently, employing a similar principle, we introduce the notion of parameter-efficient inference from a communication standpoint. This approach entails transmitting only small-scale tuneable modules while maintaining the backbone with large-scale parameters frozen. Consequently, this strategy mitigates the communication overhead associated with parameter transmission.

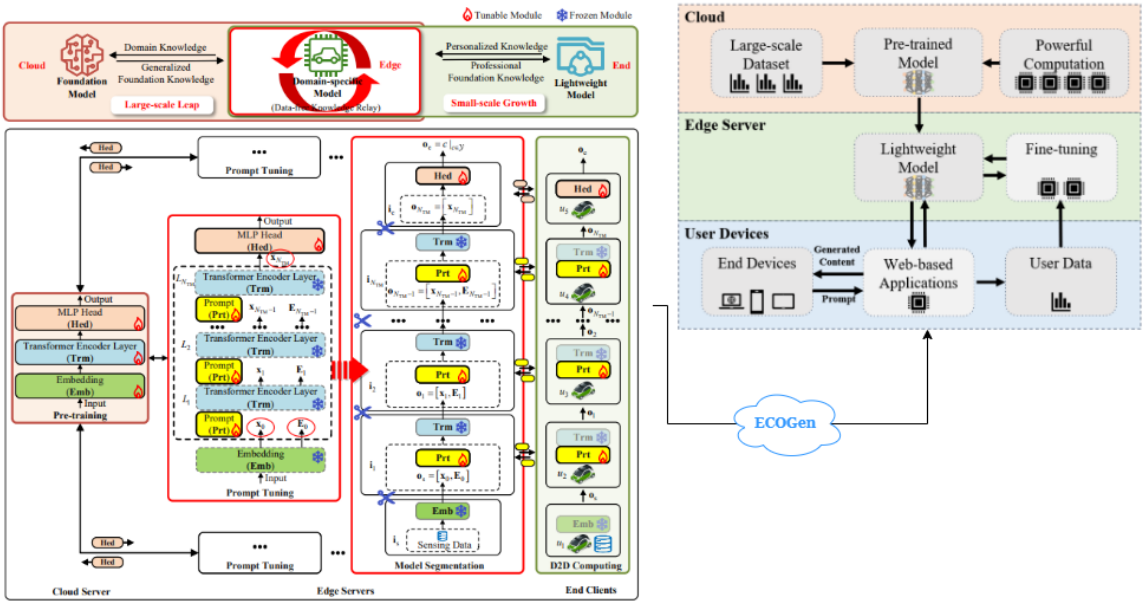

Fig. 1. The architecture of the proposed GenAI systems with the edge-cloud computing paradigm model.

Fig. 1 illustrates the architecture of the proposed model, a collaborative cloud-edge-end intelligence framework for GAI model fine-tuning, which consists of foundation models deployed on cloud servers, domain-specific models deployed on edge servers, and lightweight models deployed in 6G terminal clients. This model is built upon the collaborative cloud-edge-end intelligence framework and the D2D communication framework. After segmentation, the tuneable modules of the edge model are transferred to the client cluster, where the embedding layer is deployed at the start point and its output is $os = [x0, E0]$, while the MLP head layer is deployed at the end point, and the output (i.e., the final classification result) is defined as $oe = c|\, c \in Y$, where the alphabet $Y = y1, y2, \cdots, y|Y|$ represents the label space, such as the set of classes in object classification or recognition tasks. In addition, the middle NTM Transformer layers are deployed on the remaining clients of the cluster.

The Federated Learning (FL) framework is employed to facilitate cross-domain distributed learning between a central cloud service and multiple edge servers. In this setup, the cloud server leverages large-scale unlabelled datasets and extensive computing resources for full-model pre-training, thereby imparting generalized foundational knowledge to the Generative Artificial Intelligence (GAI) Full Models (FMs). This knowledge is subsequently transferred to the domain-specific models on the edge servers during the cloud model delivery process. Simultaneously, the edge server aggregates personalized knowledge acquired from distributed terminal clients to derive domain-specific knowledge. This domain-specific knowledge is then conveyed to the FMs of the cloud server during the edge model parameter uploading process, thereby enhancing the FMs' ability to generalize across domains. Consequently, the bidirectional flow of knowledge between the cloud and edge subnets forms a beneficial closed loop between the FMs and domain-specific models. Given the sparse data availability on edge servers and clients' concerns regarding privacy and resource limitations, an approach is devised to maximize the utilization of distributed clients' data and computing resources. This involves employing lightweight models on clients to collectively undertake model fine-tuning based on hierarchical federated semi-supervised learning (HFSL), as well as task inference based on supervised learning (SL). The edge server facilitates this process by associating domain-specific clients with cell-free networking, furnishing them with domain-specific foundational knowledge acquired from the cloud server and fine-tuned based on domain data. Concurrently, personalized knowledge gleaned from clients' local data is gathered to acquire new domain-specific knowledge. Consequently, the bidirectional flow of knowledge within the edge-end subnet establishes a mutually beneficial closed loop between the edge domain-specific model and the lightweight models on terminals.

Edge servers play a pivotal role within the metaverse system, akin to their function in content delivery networks (CDNs), by disseminating content based on geographical proximity and aiding in load distribution from cloud servers. Users situated in proximity are connected to the same edge server. Upon receiving a request from a user to generate a local scene within the metaverse system, the request is routed through the edge servers and cached. Subsequently, other users in the same vicinity can access these cached scenes on the edge servers, thereby further diminishing latency. Additionally, it is imperative to harness the computational resources available in the edge servers. For instance, they can assist in compressing and decompressing scenes generated on the cloud server. Consequently, this not only reduces latency but also enhances the quality of the generated scenes.

Artificial Intelligence of Things (AIoT) represents a burgeoning application that amalgamates artificial intelligence (AI) technologies with Internet of Things (IoT) systems. By integrating AI capabilities within the ubiquitous wireless networking infrastructure, AIoT systems are engineered, wherein end devices exhibit a degree of intelligence in data processing and analytics. The utilization of Generative AI (GenAI) can further enhance the scope of applications within this domain. For instance, a voice assistant could engage with users in various applications such as autonomous driving, smart cities, and smart homes, necessitating the automatic generation of fluent human speech from textual data sources. In the context of implementing AIoT through edge-cloud computing, considerations must be given to privacy, personalization, and data synchronization. Users may gather data to train personalized GenAI models tailored to their specific needs. Ideally, a straightforward GenAI model with satisfactory performance should be trained on user devices. Subsequently, the model parameters from multiple users can be transmitted to cloud servers for aggregation into a more sophisticated GenAI model via federated learning. Additionally, data collection from end devices should be continuous to ensure the information within GenAI models remains current. Online optimization techniques facilitate the training of machine learning models with continuous streams of data. Firmware updates enable user devices to synchronize with the advanced GenAI model. Consequently, the entire system stands to benefit from a broader pool of training data amassed from users via federated learning, while simultaneously upholding user data privacy.

IV. RESULTS & DISCUSSION

Table 2 presents a comparative analysis of power consumption, carbon emissions, and cloud computational costs associated with training large Generative Artificial Intelligence (GenAI) models across different modalities. The table includes four models: GPT-3 for text generation, BigGAN for image generation, GANSynth for audio generation, and a proposed model capable of generating text-image pairs. For the training of GPT-3, which focuses on text generation, a substantial hardware setup comprising 10,000 V100 GPUs is utilized, with the power consumption not explicitly provided. However, the training process spans 355 hours, resulting in a considerable energy consumption of approximately 1.29 × 10^6 kWh. In contrast, training the BigGAN model for image generation requires a single V100 GPU, consuming 300 watts of power over a duration of 3,072 hours. This translates to an energy consumption of 921.3 kWh, significantly lower than that of GPT-3 despite the longer training duration. Similarly, for the GANSynth model, designed for audio generation, a single V100 GPU is employed, consuming 300 watts of power during 108 hours of training. The energy consumption for this model is 32.4 kWh, substantially lower than both GPT-3 and BigGAN due to the shorter training duration and specific modality.

Table 2. Comparison of power consumption, carbon emission, & cloud computational cost in the training of large GenAI models in diverse modalities.

Model	Modality	Hardware	Power (watts)	Hours	Energy Consumption (kWh)
GPT-3 [33]	Text	V100 GPU x10,000	-	355	1.29×10^6
BigGAN [34]	Image	V100 GPU x1	300	3,072	921.3
GANSynth[35]	Audio	V100 GPU x1	300	108	32.4
Proposed model	Text-Image	V100 GPU x 2	400	97	29.2

Lastly, the proposed model capable of generating text-image pairs requires two V100 GPUs, consuming 400 watts of power over a training period of 97 hours. The energy consumption for this model is 35.2 kWh, slightly higher than GANSynth but considerably lower than GPT-3 and BigGAN. In summary, while GPT-3 exhibits the highest energy consumption due to its extensive hardware requirements and longer training duration, models focused on specific modalities such as GANSynth and the proposed text-image model demonstrate significantly lower energy consumption, highlighting the potential benefits of tailored hardware configurations and specialized training processes for different GenAI modalities.

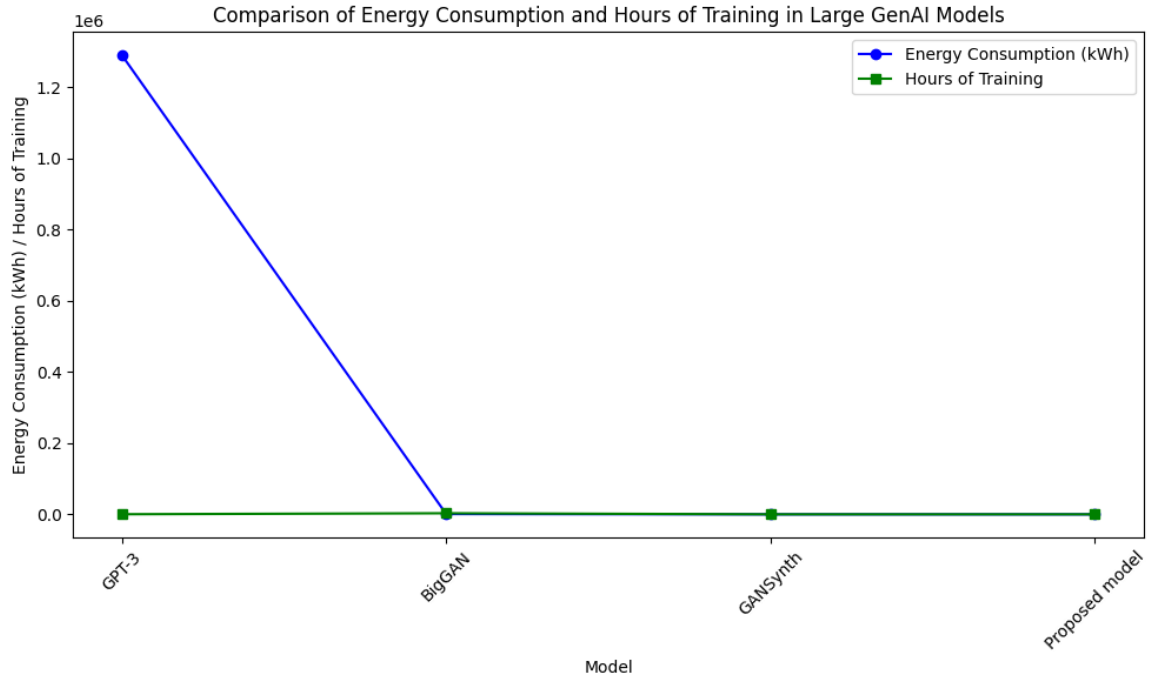

Fig. 2. Comparison of Energy Consumption & Hours of Training in Large GenAI Models

Table 3 illustrates the impact of varying the number of client clusters on the performance of the proposed model. Each row represents a different number of client clusters, ranging from 1 to 6, while the corresponding accuracy values at the beginning (First) and end (End) of the training process are presented. The accuracy values depict the model's performance in terms of correctly predicting outcomes. As the number of client clusters increases, there is a noticeable improvement in accuracy, both at the initial stages and upon completion of the training process. Specifically, the accuracy ranges from 0.932 to 0.969 at the beginning of training and from 0.955 to 0.977 at the end of training across the various cluster configurations. This suggests that increasing the number of client clusters leads to enhanced model performance, indicating the importance of cluster optimization in achieving higher accuracy levels for the proposed model.

Table 3. The consequence of the no. of client cluster on the proposed model.

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy (First/end)	0.932/ 0.955	0.942/ 0.958	0.947/ 0.964	0.953/ 0.967	0.968/ 0.976	0.969/ 0.977

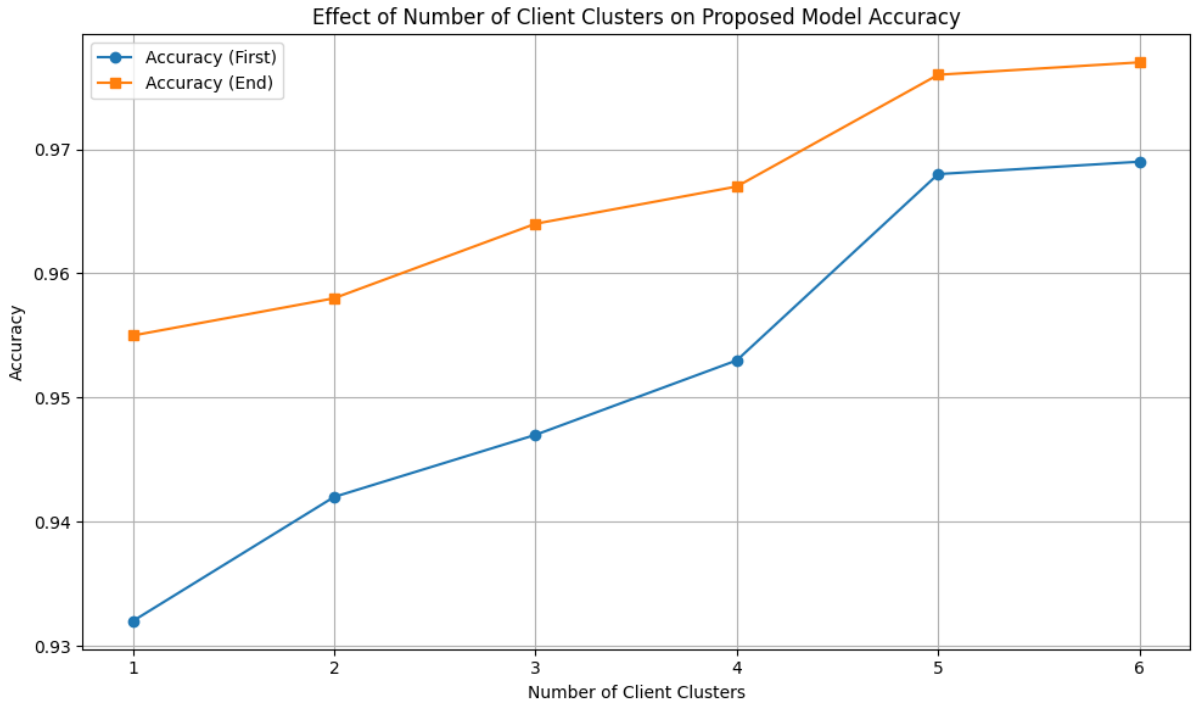

Fig. 3. Effect of number of client clusters on proposed model accuracy

V. CONCLUSION

The integration of GenAI services on a large scale presents a novel hurdle in the development of contemporary edge-cloud computational systems, owing to the considerable model sizes, substantial power consumption, and potential latency issues stemming from limited computational and network capabilities. This study delineated the significance of developing scalable GenAI systems and corroborated the associated challenges through the conceptualization of two representative GenAI services. Furthermore, an extensive discourse on diverse design considerations for GenAI services vis-à-vis prevailing communication frameworks was provided. The synthesis emphasizes the necessity for a well-calibrated design approach that optimally allocates computational resources between edge and cloud servers while addressing concerns related to latency, data privacy, and customization. Notably, the adoption of federated learning emerges as pivotal, wherein compact GenAI models are trained at edges, complemented by the training of large-scale models in the cloud through the amalgamation of numerous smaller models. Moreover, the distribution of most inference tasks at edges is advocated. The paper outlines

several avenues for future research, including the development of domain-specific GenAI models, the decomposition of large language models, the implementation of eco-friendly GenAI models, and the assurance of quality in AIGC. Additionally, we introduce a collaborative cloud-edge-end intelligence framework tailored for GenAI applications. The model facilitates efficient hierarchical federated semi-supervised learning (HFSL)-based model fine-tuning and supervised learning (SL)-based task reasoning within a unified architecture. Furthermore, a comprehensive analysis of the operational challenges encountered by the proposed model and an exploration of the impact of various influencing factors on its performance were conducted through simulation. Finally, the paper delineates future challenges and directions in the symbiotic relationship between GenAI and edge intelligence.

REFERENCES

[1] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia, "Scangan360: A generative model of realistic scanpaths for 360 images," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 5, pp. 2003–2013, 2022.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, 2022.

[3] J. Ratican, J. Hutson, and A. Wright, "A proposed meta-reality immersive development pipeline: Generative ai models and extended reality (xr) content for the metaverse," Journal of Intelligent Learning Systems and Applications, vol. 15, 2023.

[4] Z. Xiao, Z. Xia, H. Zheng, B. Y. Zhao, and J. Jiang, "Towards performance clarity of edge video analytics," in 2021 IEEE/ACM Symposium on Edge Computing (SEC). IEEE, 2021, pp. 148–164.

[5] M. Suzuki and Y. Matsuo, "A survey of multimodal deep generative models," Advanced Robotics, vol. 36, no. 5-6, pp. 261–278, 2022.

[6] D. Foster, Generative deep learning: teaching machines to paint, write, compose, and play. O'Reilly Media, 2019.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[8] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in international conference on machine learning. PMLR, 2016, pp. 1558–1566.

[9] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, and L. Shao, "Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning," IEEE Transactions on Image Processing, vol. 29, pp. 3665–3680, 2020.

[10] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," Advances in neural information processing systems, vol. 28, 2015.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139– 144, 2020.

[12] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," IEEE signal processing magazine, vol. 35, no. 1, pp. 53–65, 2018.

[13] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," ACM Computing Surveys (CSUR), vol. 54, no. 8, pp. 1–49, 2021.

[14] B. Wang, C. Zhang, C. Wei, and H. Li, "A focused study on sequence length for dialogue summarization," arXiv preprint arXiv:2209.11910, 2022.

[15] C. Wei, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, "An overview on language models: Recent developments and outlook," arXiv preprint arXiv:2303.05759, 2023

[16] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapes, "Video transformers: A survey," ´ IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[17] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," arXiv preprint arXiv:2201.05337, 2022.

[18] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook," Journal of Visual Communication and Image Representation, p. 103685, 2022.

[19] X. Xia, F. Chen, Q. He, J. Grundy, M. Abdelrazek, and H Jin, "Cost-effective app data distribution in edge computing," IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 1, pp. 31-44, 2020.

[20] S. Duan, D. Wang, J. Ren, F. Lyu, Y. Zhang, H. Wu, X. Shen, "Distributed artificial intelligence empowered by end-edge-cloud computing: A survey," IEEE Communications Surveys & Tutorials, 2022.

[21] G. Zhu, Z. Lyu, X. Jiao, P. Liu, M. Chen, J. Xu, S. Cui, P. Zhang, "Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G," in Science China Information Sciences, vol. 66, no. 3, pp. 130301, 2023.

[22] X. Huang, P. Li, H. Du, J. Kang, D. Niyato, D. Kim, Y. Wu, "Federated Learning-Empowered AI-Generated Content in Wireless Networks," arXiv preprint arXiv:2307.07146, 2023.

[23] Z. Zhang, Y. Yang, Y. Dai, Q. Wang, Y. Yu, L. Qu, Z. Xu, "Fed-PETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models," Association for Computational Linguistics (ACL), pp. 9963–9977, 2023.

[24] Y. Tian, Y. Wan, L. Lyu, D. Yao, H. Jin, L. Sun, "FedBERT: When federated learning meets pre-training," in ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 4, pp. 1–26, 2022.

[25] H. Zou, Q. Zhao, L. Bariah, M. Bennis, M. Debbah, "Wireless multi-agent generative ai: From connected intelligence to collective intelligence," arXiv preprint arXiv:2307.02757, 2023.

[26] Z. Lin, G. Qu, X. Chen, K. Huang, "Split Learning in 6G Edge Networks," arXiv preprint arXiv:2306.12194, 2023.

[27] A. Agarwal, M. Rezagholizadeh, P. Parthasarathi, "Practical Takes on Federated Learning with Pretrained Language Models," Findings of the Association for Computational Linguistics: EACL 2023, pp. 454–471, 2023.

[28] M. Jia, L. Tang, B. Chen, C. Cardie, S. Belongie, B. Hariharan, S. Lim, Ser-Nam, "Visual prompt tuning," European Conference on Computer Vision, pp. 709–727, 2022.

[29] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, G. Neubig, "Towards a unified view of parameter-efficient transfer learning," arXiv preprint arXiv:2110.04366, 2021.

[30] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[31] J. Chen, W. Xu, S. Guo, J. Wang, J. Zhang, H. Wang, "FedTune: A Deep Dive into Efficient Federated Fine-Tuning with Pre-trained Transformers," arXiv preprint arXiv:2211.08025, 2022.

[32] Z. Cheng, X. Xia, M. Liwang, X. Fan, Y. Sun, X. Wang, L. Huang, "CHEESE: distributed clustering-based hybrid federated Split learning over edge networks," IEEE Transactions on Parallel and Distributed Systems, 2023.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[34] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," arXiv preprint arXiv:1809.11096, 2018.

[35] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," arXiv preprint arXiv:1902.08710, 2019.