Energy-Efficient Multimodal Foundation Models for Real-Time Visual Content Generation on Edge Devices

Shaowenhao Yuan
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
yuan1983@uc.edu

George Edaems
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
george.work@binghamton.edu

Abstract

The rapid proliferation of multimodal foundation models has enabled unprecedented capabilities in visual content generation, yet their deployment on edge devices is severely constrained by energy budgets, thermal envelopes, and latency requirements. This paper provides a system-level analysis of energy-efficient multimodal foundation models designed for real-time visual synthesis at the edge. We examine the structural trade-offs between model expressiveness and power consumption, survey architectural innovations in model compression, distillation, and quantization, and discuss hardware-software co-design strategies that leverage heterogeneous accelerators and edge-native compute fabrics. The deployment infrastructure is analyzed through the lens of the edge-cloud continuum, workload partitioning, and orchestration for low-latency inference. Beyond technical optimization, the paper addresses sociotechnical dimensions including robustness, fairness, and bias in generative outputs, lifecycle sustainability accounting, and the evolving landscape of governance and policy instruments that shape responsible edge AI. Throughout the discussion, cross-domain comparisons between mobile vision, data center generative models, and emerging edge-native foundations illuminate the systemic challenges and opportunities. The paper concludes with forward-looking perspectives on standardization, regulatory frameworks, and the long-term viability of energy-proportional multimodal intelligence at the extreme edge, arguing that sustainable visual AI demands a holistic integration of model design, accelerator ecosystems, and socio-environmental accountability.

Keywords

multimodal foundation models, energy efficiency, edge computing, real-time visual generation, model compression, hardware-software co-design.

1. Introduction

Foundation models have reshaped the landscape of artificial intelligence by offering general-purpose representations that can be adapted to a wide array of downstream tasks through minimal fine-tuning or prompting [1]. Simultaneously, the exponential growth of edge computing has pushed computation, storage, and intelligence toward the data-producing periphery, motivated by bandwidth constraints, privacy requirements, and the demand for ultra-low latency responses [2]. The confluence of these two trends gives rise to the pursuit of edge-native multimodal foundation models capable of real-time visual content generation, a

capability that carries profound implications for augmented reality, assistive technologies, interactive media, and industrial digital twins. Realizing such models on resource-constrained devices, however, exposes fundamental tensions between the prodigious computational demands of generative architectures and the severe energy, memory, and thermal limits imposed by mobile, wearable, and embedded platforms.

Research on edge intelligence has proposed a spectrum of solutions ranging from lightweight neural architectures to collaborative inference between edge and cloud [3]. Yet the extraordinary scale of foundation models, particularly those that fuse vision and language modalities for high-fidelity image synthesis, necessitates rethinking the entire system stack. Contemporary text-to-image models such as Stable Diffusion [4] and DALL-E 2 [5] routinely involve billions of parameters, iterative denoising steps, and cross-attention mechanisms that consume multiple watt-hours per generation on GPU-class hardware. Translating such workloads to milliwatt-scale edge devices forces a critical re-examination of model expressiveness, inference scheduling, memory bandwidth, and the energy proportionality of each subsystem. This paper adopts a system-level, interdisciplinary perspective to dissect these challenges, moving beyond isolated algorithmic improvements to encompass hardware architectures, deployment topologies, fairness considerations, lifecycle carbon accounting, and governance frameworks that together define the viability of energy-efficient generative AI at the edge.

2. System-Level Challenges for Edge-Centric Generative AI

Deploying multimodal generative models at the edge gives rise to a multitude of interacting constraints that must be addressed concurrently. The primary challenge is the energy-delay product: real-time visual content generation, particularly when conditioned on text or other modalities, imposes strict latency budgets on the order of tens to hundreds of milliseconds, while battery-operated devices can allocate only a few joules per inference session. Empirical studies of large-scale deep learning have shown that training and inference for generative models can incur substantial carbon footprints and electricity costs, raising urgent questions about the sustainability of scaling such models without corresponding efficiency gains [6, 7]. The tension between rich multimodal representations and frugal computation is exacerbated by the stochastic nature of iterative generative processes, which often require multiple forward passes through a denoising network, amplifying the effective compute per output sample.

Beyond raw compute, memory footprint constitutes a key bottleneck. Diffusion models typically maintain high-resolution feature maps throughout the denoising chain, and transformer-based text encoders demand substantial parameter memory and activation storage. On edge devices with limited DRAM and no virtual memory, these requirements can lead to excessive data movement between off-chip memory and compute units, incurring significant energy penalties. The system designer must therefore orchestrate a careful balance between spatial and temporal tiling, activation recomputation, and aggressive compression, all while preserving the generative quality that users expect. Furthermore, edge environments exhibit high variability in thermal conditions, battery states, and concurrent workloads, requiring adaptive strategies that can gracefully degrade generation fidelity or switch to lighter model variants without violating real-time guarantees. These challenges collectively demand a holistic design philosophy that integrates model architecture, compression, hardware acceleration, and runtime scheduling into a unified energy-aware framework.

3. Architectural Design Space and Energy-Aware Model Compression

The architectural design space for energy-efficient visual generation spans convolutional, transformer, and hybrid backbones, each offering distinct trade-offs between parameter efficiency, computational regularity, and perceptual quality. A long line of work on efficient convolutional networks has demonstrated that depthwise separable convolutions, inverted residuals, and compound scaling can drastically reduce multiply-accumulate operations while maintaining competitive accuracy [8, 9, 10, 11]. Although these designs originated in discriminative vision tasks, their principles have been adapted to the generative domain through carefully pruned and structurally re-parameterized diffusion U-Nets. In multimodal settings, however, the presence of cross-attention between textual embeddings and visual feature maps introduces irregular memory access patterns and dynamic tensor shapes that are less amenable to the regularization offered by pure convolutional blocks, necessitating hybrid architectures that interleave efficient local operations with sparse or linearized attention.

Knowledge distillation has emerged as a particularly powerful technique for transferring the generative quality of large teacher models into compact student networks amenable to edge deployment. The foundational concept of distillation, where a student is trained to match the softened output distributions of a teacher, has been extended to iterative generative processes through progressive distillation, which reduces the number of sampling steps required by a diffusion model while preserving sample fidelity [12, 13]. By training a student to predict the output of multiple teacher denoising steps in a single forward pass, progressive distillation directly attacks the inference cost multiplier inherent in diffusion models, offering a path toward real-time generation at the edge. When combined with post-training quantization to 8-bit or even 4-bit integer formats, such compressed generative models can achieve significant reductions in energy per sample without catastrophic quality degradation. Crucially, these compression pipelines must be aware of the heterogeneous compute resources on the target edge system, as the optimal quantization scheme and layer fusion patterns depend heavily on whether the device relies on a CPU, GPU, NPU, or custom accelerator.

4. Hardware-Software Co-Design and Accelerator Integration

Achieving real-time visual generation on milliwatt-scale devices demands deep co-design between model architectures and the underlying hardware. A comprehensive survey of efficient processing techniques for deep neural networks highlights the necessity of algorithmic transformations that align with dataflow architectures, minimizing off-chip memory accesses through tiling, loop ordering, and weight-stationary or output-stationary dataflows [14]. Custom accelerators for edge inference, including neural processing units integrated into mobile system-on-chips and discrete AI accelerators, offer orders of magnitude better energy efficiency than general-purpose CPUs or GPUs, but they impose constraints on operator support, numerical precision, and memory hierarchy that must be internalized during model design. The recent emergence of edge-native text-to-image foundation models trained entirely on domestically developed AI accelerators provides a compelling case study in full-stack co-optimization [15]. In such deployments, quantization-aware training is combined with structural pruning, and the denoising backbone is tailored to exploit the accelerator’s systolic array dimensions and on-chip scratchpad memory, enabling sustained throughput at sub-watt power envelopes. These systems illustrate that energy-efficient multimodal generation is not solely a problem of compressing a pretrained cloud model but rather a holistic design exercise that considers accelerator microarchitecture, compiler-level optimizations, and model topology as a single integrated search space.

The integration of specialized accelerators also introduces challenges related to software stack maturity and ecosystem fragmentation. Memory allocation, operator scheduling, and model partitioning across heterogeneous compute units must be orchestrated by a runtime that dynamically accounts for energy states and thermal throttling. Furthermore, the security implications of running generative foundation models on edge devices, including model inversion attacks, adversarial perturbations, and the potential misuse of generated content, necessitate hardware-enforced trusted execution environments and secure model provisioning pipelines. These concerns, while often peripheral in cloud-centric discourse, become central in edge deployment scenarios where devices operate in uncontrolled physical environments and handle sensitive user data.

5. Multimodal Fusion Pipelines and Real-Time Inference on the Edge

Multimodal foundation models for visual generation rely on the tight coupling between language understanding backbones and visual decoders. Architectures that bootstrap vision-language pretraining by connecting frozen image encoders to large language models through lightweight querying transformers have demonstrated remarkable few-shot capabilities, yet their computational profiles remain dominated by the transformer layers of the language model and the visual decoder [16]. Visual instruction tuning further enriches the multimodal interplay by aligning generated visual content with complex user intents, but it simultaneously increases the intra-inference latency due to additional cross-attention blocks and token generation steps [17]. For edge deployment, researchers have explored decomposing these pipelines into asynchronous stages, where the computationally heavy language encoding can be offloaded to a nearby edge server or precomputed for common prompts, while the lightweight visual decoder runs entirely on the device. Such decoupling, however, introduces synchronization overheads and can break the tight conditioning that defines high-quality multimodal generation, requiring careful pipeline parallelism and cached key-value states to preserve semantic coherence.

Real-time operation demands not only low average latency but also predictable tail latency under stochastic scheduling. Edge operating systems and inference engines must be designed with preemptive multi-task capabilities that allow critical generation tasks to meet deadlines even when co-located with other signal processing workloads. The energy implications of buffering, data movement, and repeated context switching can dominate the total energy budget if not explicitly managed. Techniques drawn from real-time systems research, such as energy-aware earliest deadline first scheduling and dynamic voltage and frequency scaling, have begun to be applied to neural inference pipelines, but their interaction with the iterative, non-deterministic nature of generative models remains an underexplored frontier. A systems perspective thus mandates tight integration between the runtime scheduler, the model's internal sampling trajectory, and the hardware power management unit.

6. Deployment Infrastructure and Edge-Cloud Continuum

No single edge device can encompass the full generality of foundation models, and a realistic deployment must exploit the edge-cloud continuum for workload partitioning and lifecycle management. A canonical pattern involves using cloud resources for model training, large-batch distillation, and the compilation of optimized model variants, while edge devices perform inference, fine-tuning, and personalization on locally generated data. The edge intelligence paradigm envisions a hierarchical orchestration layer that decides, based on device energy states, network conditions, and task criticality, which portions of the multimodal pipeline execute locally and which are delegated to proximal edge servers or the

cloud [3]. Such a continuum requires standardized model interchange formats, versioned model registries, and robust fallback mechanisms that maintain generation quality even under intermittent connectivity.

Federated learning and split learning present additional avenues for collaborative model improvement without centralizing sensitive visual and textual prompts. However, the energy cost of participating in federated training rounds can be prohibitive for battery-powered devices, and the communication overhead of exchanging model updates must be carefully balanced against the gains in personalization. Research on communication-efficient federated optimization, combined with progressive model shrinking and quantization, can lower the barrier, but the system designer must also account for the non-IID nature of user-generated multimodal data and the resulting fairness implications across diverse demographic groups. In this light, the deployment infrastructure becomes not merely a technical scaffold but a governance mechanism that distributes computational burden and data agency among stakeholders.

7. Robustness, Fairness, and Sociotechnical Governance

Generative foundation models inherit and amplify biases present in their training corpora, and these biases can manifest in visual content that reinforces stereotypes or produces harmful representations. A comprehensive survey on fairness in machine learning underscores that fairness is not a monolithic property but a multidimensional construct that includes demographic parity, equalized odds, and individual fairness, each with its own trade-offs and measurement challenges [18]. In the context of edge-deployed multimodal generation, these concerns are magnified because the models operate in intimate, context-rich environments where biased outputs can cause immediate social harm. System-level interventions must therefore embed fairness-aware filtering, prompt sanitization, and output auditing directly into the inference pipeline, without incurring prohibitive energy overhead. Lightweight, learnable fairness adapters that can be toggled on-device and updated via the edge-cloud continuum represent a promising direction, but they remain nascent.

Broader governance frameworks provide ethical guardrails for the deployment of autonomous content generation systems. The AI4People initiative articulated foundational principles of beneficence, non-maleficence, autonomy, justice, and explicability that should guide the development and deployment of AI technologies [19]. Translating these high-level principles into concrete design requirements for energy-efficient edge AI involves embedding explainability modules that can attribute generated visual features to specific prompt elements and training data influences, all while operating within tight energy budgets. Moreover, the proliferation of edge-deployed generative models raises jurisdictional questions about content liability, intellectual property of generated visuals, and the enforcement of platform-level content moderation across heterogeneous device ecosystems. A system-level approach to governance must therefore consider not only the technical properties of the model but also the institutional and regulatory environments in which devices operate.

8. Lifecycle Sustainability and Energy Accountability

Sustainability assessments of AI systems must encompass the full lifecycle, from data curation and model training to inference, device manufacturing, and end-of-life disposal. While much attention has focused on the carbon footprint of large-scale training runs, the aggregate energy consumption of billions of edge inference queries may dominate the lifecycle emissions of widely deployed generative models [6, 7]. Quantifying the carbon

emissions of machine learning operations has stimulated the development of software tools that report energy usage and equivalent carbon emissions per inference call [20]. When integrated into edge runtime environments, such tools can provide real-time feedback to users and system orchestrators, enabling carbon-aware scheduling that defers non-urgent generation tasks to periods of high renewable energy availability on the grid. Transparency in energy reporting also supports regulatory compliance and consumer choice, as device manufacturers and application developers can be required to disclose the energy intensity of AI features.

The hardware supply chain further complicates the sustainability equation. Edge AI accelerators rely on semiconductor fabrication processes that are energy- and water-intensive, and the geopolitics of rare-earth mineral extraction adds a layer of social sustainability concerns. Designing for longevity, modularity, and firmware-level upgradability of AI accelerators can extend device lifetimes and reduce electronic waste, aligning with circular economy principles. From a system design perspective, the energy efficiency of a multimodal foundation model should be evaluated not only in operations per watt but also in useful generative work per embodied energy of the entire hardware-software stack over its operational life.

9. Policy Implications and Standardization Efforts

The global landscape of AI ethics guidelines reveals a converging consensus around principles of transparency, accountability, and human oversight, yet the operationalization of these principles for energy-efficient edge AI remains fragmented [21]. Policymakers face the challenge of setting performance benchmarks that simultaneously incentivize low energy consumption, high generation fidelity, and fairness, without stifling innovation through overly prescriptive technical mandates. Standardization bodies such as the IEEE and ISO have begun to develop frameworks for assessing the environmental impact of AI systems, but these standards typically address cloud-scale training rather than the distributed, heterogeneous world of edge inference. A future standards track could define energy proportionality metrics that normalize generation quality against joules per sample, carbon-aware model cards that report lifecycle emissions, and conformance testing suites that validate the fairness and robustness of compressed models on diverse hardware.

Incentive structures also matter. Energy labeling programs analogous to the ENERGY STAR certification for consumer electronics could be extended to AI-accelerated applications, creating market pull for efficient on-device generative capabilities. Public procurement policies that require disclosed energy and fairness audits for AI systems used in education, healthcare, and public services would further accelerate adoption of responsible edge AI. International coordination will be essential to prevent a race to the bottom where jurisdiction-shopping by developers undermines sustainability and fairness commitments. In this regulatory milieu, the edge-native model development paradigm, exemplified by initiatives that train entirely on domestic accelerator ecosystems [15], may also carry geopolitical dimensions that intersect with technology sovereignty and data governance policies, requiring careful multidisciplinary dialogue.

10. Conclusion

Energy-efficient multimodal foundation models for real-time visual content generation at the edge sit at the intersection of deep learning, computer architecture, systems engineering, and sociotechnical governance. This paper has argued that achieving sustainable, performant, and fair on-device generation demands a layered approach that co-optimizes model architecture,

compression, hardware acceleration, and runtime orchestration while embedding accountability mechanisms for energy consumption, bias, and lifecycle impacts. The move from cloud-centric model scaling toward edge-native systems thinking opens opportunities for radical energy proportionality but also introduces new structural tensions around fairness, security, and regulatory coherence. As the ecosystem continues to mature, interdisciplinary collaboration among researchers, standards bodies, and policymakers will be critical to ensure that the proliferation of generative visual AI on personal devices aligns with broader societal values of sustainability, equity, and trustworthiness.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646.

3. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762.

4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695.

5. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125.

6. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

7. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., … & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

8. Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv preprint arXiv:1510.00149.

9. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510–4520.

11. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, 6105–6114.

12. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

13. Salimans, T., & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. Proceedings of the International Conference on Learning Representations.

14. Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295–2329.

15. Chen, C., Wang, C., Li, Y., Wan, Z., Geng, M., Xiao, J., ... & Peng, Y. (2026). JuZhou 1.0 Technical Report: The First Edge-Native Text-to-Image Foundation Model Trained Entirely on China-Developed AI Accelerators. arXiv preprint arXiv:2606.28421.

16. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning, 19730–19742.

17. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36.

18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

19. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

20. Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.

21. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.