Resource-Aware Diffusion Model Compression for Mobile and Embedded Multimodal Intelligence Systems

Malcolm J. Perez
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
malcolmjperez@colostate.edu

Seamor Parekh
School of Computing, Clemson University, Clemson, SC, USA.
parekh1974@clemson.edu

Ankit R. Rha
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
ankitj@ucf.edu

Martins M. Parzez
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
perezmartins@oregonstate.edu

Abstract

Diffusion models have emerged as powerful generative frameworks capable of producing high-fidelity multimodal content that spans text, images, speech, and sensor streams. Their deployment on mobile and embedded platforms, however, is severely constrained by memory footprints, energy budgets, and real-time latency requirements that conflict with the iterative denoising process at their core. This paper presents a system-level analysis of resource-aware diffusion model compression, moving beyond isolated algorithmic contributions to examine the structural trade-offs, architectural implications, and socio-technical dimensions of shrinking these models for edge-native multimodal intelligence. We survey a continuum of compression paradigms—quantization, pruning, knowledge distillation, neural architecture search, and sampling trajectory reduction—and evaluate them through the lenses of hardware-software co-design, on-device orchestration, sustainable computing, fairness, and governance. The discussion foregrounds infrastructure-level considerations such as heterogeneous accelerator utilization, memory hierarchies, and federal learning pipelines that distribute compression across devices. By integrating case illustrations from mobile vision-language systems, on-device audio synthesis, and sensor-conditioned scene generation, we identify persistent tensions between fidelity, latency, and equity. We argue that compression must be treated not as a post-hoc optimization but as a first-class design constraint that shapes the entire training-inference-operations lifecycle. The paper concludes by outlining policy implications and future research directions for robust, inclusive, and energy-proportional diffusion intelligence at the extreme edge.

Keywords

diffusion models, model compression, mobile AI, edge computing, multimodal intelligence, resource-aware systems, sustainable computing, fairness, governance.

1. Introduction

The rapid evolution of diffusion models has redrawn the boundaries of generative artificial intelligence, enabling the synthesis of realistic images, coherent speech, and multimodal narratives with unprecedented quality. Starting from foundational formulations that define a forward diffusion process gradually corrupting data and a learned reverse process that restores structured information, these models have scaled to billions of parameters and routinely power applications in content creation, scientific simulation, and assistive technologies [1]. The development of latent diffusion architectures that operate in compressed perceptual spaces significantly reduced the computational overhead relative to pixel-space generation without sacrificing output fidelity, making diffusion practical for server-class hardware [2]. Despite these advances, the formidable computational demands of the iterative denoising procedure—often requiring hundreds to thousands of functional evaluations—remain a fundamental barrier to adoption on mobile phones, microcontrollers, augmented reality headsets, and embedded vision systems where multimodal reasoning is most transformative.

The tension between model capacity and resource availability has motivated a rich body of compression research that spans quantization, pruning, knowledge distillation, and neural architecture search. Early efforts in deep neural network compression established that aggressive parameter reduction combined with low-precision arithmetic can shrink a model by an order of magnitude with negligible accuracy loss [4]. These techniques were later extended to ternary and binary representations, demonstrating that even extreme discretization can preserve representational power when coupled with appropriate training or fine-tuning strategies [5]. Knowledge distillation, wherein a compact student model learns to mimic the softened output distributions of a larger teacher, offered another pathway to embed complex dependencies into lightweight architectures, although its application to generative models raises unique challenges around distribution matching and perceptual alignment [3]. As diffusion models became dominant, research communities developed dedicated compression pipelines that exploit the iterative nature of the denoising trajectory, enabling on-device image generation in under two seconds through architectural redesign, operator fusion, and step distillation [6]. Post-training quantization schemes specifically designed for diffusion backbones further demonstrated that 8-bit and even 4-bit inference can retain visual quality, opening the door to deployment on NPUs and DSPs commonly found in mobile SoCs [7].

While these algorithmic contributions are valuable, they rarely engage with the broader systems perspective that is essential for real-world deployment in multimodal contexts. Compression is too often studied as an isolated accuracy-efficiency tradeoff on a single modality, ignoring how different sensor streams—camera, microphone, inertial measurement unit, lidar—fuse at the edge under tight memory and energy envelopes. This paper argues that resource-aware compression of diffusion models for mobile and embedded multimodal intelligence must be understood as a socio-technical systems problem that interweaves hardware-software co-design, distributed orchestration, lifecycle carbon costs, fairness implications, and regulatory governance. We structure the analysis around five thematic pillars. Section 2 contextualizes compression within the unique computational profile of multimodal diffusion. Section 3 presents a system-level taxonomy of compression strategies, emphasizing structural trade-offs rather than algorithmic minutiae. Section 4 discusses deployment architectures, including co-design with edge accelerators and federated

refinement. Section 5 examines sustainability, fairness, and governance dimensions that are frequently overlooked. Section 6 provides cross-domain case illustrations, and Section 7 concludes with policy-oriented recommendations.

2. Diffusion Multimodality and Resource Pressure at the Edge

Diffusion models generate data by learning to invert a noise-injection process that transforms a clean signal into an isotropic Gaussian distribution. The sampling phase requires iteratively applying a denoising network, often a U-Net or transformer backbone, producing one refined output at each step. For a single image generation at standard resolution, this can involve hundreds of network evaluations, each entailing billions of floating-point operations, memory transfers, and intermediate activation storage. When extended to multimodal tasks—such as generating an audio track conditioned on an image, producing haptic feedback from a video stream, or fusing text prompts with spatial sensor readings—the complexity compounds because multiple encoders, cross-attention mechanisms, and modality-specific decoders must operate in tandem. A mobile system that must simultaneously handle video decoding, voice interaction, and real-time rendering cannot afford to allocate a dedicated high-performance GPU core to a slow autoregressive or iterative generative process; the workload must be compressed, scheduled, and scheduled across heterogeneous compute units while preserving perceptual coherence across modalities.

Latent diffusion models partially mitigate the computational burden by shifting the denoising process into a compact latent space, typically obtained via a pretrained variational autoencoder [2]. This reduces the dimensionality by a factor of four to eight, but the denoising network itself remains large, and the encoder-decoder pair adds its own memory and latency footprint. For embedded platforms such as smart glasses or hearables, even a 50-million-parameter U-Net running for 20 sampling steps can exceed the thermal design power and cause frame drops or battery drain. The challenge is not simply one of shrinking the parameter count; it is about designing a compression regime that respects the temporal dynamics of the denoising chain, the interplay between modalities, and the need to shut down portions of the model when certain sensors are inactive. From a systems perspective, adaptive precision, conditional execution, and modality-wise compute skipping become as important as static weight pruning.

A further complication arises because multimodal alignment often relies on high-capacity representation spaces that are sensitive to compression artifacts. Quantizing the cross-modal attention layers, for instance, can disproportionately affect the model's ability to associate a spoken command with a visual object, resulting in semantic drift that may not be captured by standard image quality metrics. Thus, resource-aware compression must be evaluated not only on unimodal fidelity indices such as FID or CLIP score but also on multimodal consistency measures that reflect end-user task success. This shift in evaluation philosophy has downstream consequences for how compression-aware training objectives are formulated and how on-device schedulers decide which parts of a model to quantize aggressively and which to leave in full precision.

3. A System-Level Taxonomy of Compression Strategies

Compression techniques for diffusion models can be organized along several axes that intersect with system-level concerns. The first axis is representational efficiency, which encompasses quantization, prunning, and low-rank factorization of weight tensors. Quantization reduces the bit-width of parameters and activations, thereby decreasing memory

footprint, memory bandwidth usage, and energy per operation. While uniform 8-bit post-training quantization is well established for discriminative models, diffusion models exhibit sensitivity to quantization error accumulation across iterative steps, requiring step-aware calibration and mixed-precision assignment [7]. Pruning removes entire neurons, channels, or attention heads that contribute minimally to the loss landscape. Structural pruning that eliminates hardware-unfriendly sparsity patterns is favored for mobile inference because it aligns with the vectorized execution units of DSPs and NPUs. Low-rank decomposition approximates large weight matrices by products of smaller matrices, reducing both storage and computation, though it can impair the expressivity needed for cross-modal fusion.

The second axis is algorithmic efficiency, which targets the iterative sampling process itself. Progressive distillation trains a student to replicate two teacher denoising steps in a single student step, halving the number of required iterations with each distillation stage; this approach has been shown to reduce sampling steps from over a thousand to as few as four while retaining image quality [8]. Trajectory truncation methods learn shorter noise schedules, while advanced solvers based on differential equations can trade asymptotic accuracy for wall-clock speed. From a systems standpoint, reducing the number of sampling steps directly cuts the energy budget and allows the accelerator to enter a low-power state sooner, a particularly valuable property for battery-operated devices where idle power dominates. However, aggressive step reduction can harm the model's ability to generate diverse multimodal outputs, increasing the risk of mode collapse in conditional generation tasks such as producing varied responses to the same prompt.

The third axis concerns architectural co-design, where compression is embedded into the model architecture itself rather than applied post hoc. Mobile-optimized convolutional backbones with depthwise separable convolutions and inverted residuals have become default building blocks for efficient vision models [9], [10]. Their extension to the diffusion U-Net via compound scaling that simultaneously adjusts depth, width, and input resolution can yield families of models that span a wide efficiency spectrum while preserving multimodal alignment [11]. Neural architecture search has been employed to automatically discover compact diffusion networks that meet latency and memory constraints, but the search cost is high, and the resulting architectures may not generalize well across different embedded hardware targets. A more pragmatic approach is to develop hardware-aware super-networks that can be sub-sampled at inference time according to the device profile, effectively realizing dynamic compression without retraining.

The interaction between these axes produces structural trade-offs that are best managed at the system level rather than through isolated per-axis optimization. For example, applying aggressive structural pruning to fit a model in the L2 cache of a microcontroller may eliminate channels that are critical for the audio-visual synchronization pathway, which can be compensated by increasing the number of distillation steps during training to redistribute representational capacity. A system-level orchestrator can decide, based on the current battery level and task priority, to select a pruned-quantized checkpoint with 8 sampling steps for casual image captioning but reserve a higher-fidelity variant with 16 steps for a medical assistive application. Such orchestration requires standardized model descriptors that expose accuracy-latency-energy profiles of compressed variants, a topic we return to in Section 4.

4. Hardware-Software Co-Design and Deployment Architectures

The deployment of compressed multimodal diffusion models on mobile and embedded systems depends intimately on the characteristics of the underlying hardware. Contemporary

smartphone SoCs integrate heterogeneous compute blocks: large CPU cores for control and sequential operations, GPUs for parallelizable neural kernels, NPUs optimized for quantized convolution and transformer primitives, and DSPs for audio and sensor signal processing. Each unit has distinct memory subsystems, supported data types, and power gating behaviors. Efficient compression strategies must therefore produce models that can be partitioned across these compute units, with high-throughput blocks like early convolutional layers assigned to the NPU in 8-bit integer mode, attention layers mapped to the GPU using 16-bit floating point, and the final lightweight decoder running on the CPU to minimize context-switching overhead. This partitioning is not static; it must adapt to thermal throttling, concurrent workloads, and battery state, underscoring the need for runtime systems that can re-bind operators on the fly.

Recent developments in edge-native training further complicate the deployment landscape by enabling model fine-tuning directly on resource-constrained devices, often leveraging China-developed AI accelerators that offer competitive throughput at lower power envelopes [13]. Such platforms enable a paradigm where a globally compressed base model is personalized on-device using local multimodal data, merging the benefits of federated learning with compression-aware adaptation. The federated setting introduces additional constraints: the compression scheme must be compatible with secure aggregation protocols that rely on the homomorphic properties of integer arithmetic [18], and the update payload communicated over cellular or WiFi networks must be minimal. This synergy between compression and federated learning is a fertile area where sparsified fine-tuning, low-bit gradient transmission, and differential privacy noise injection must be jointly optimized to maintain generative quality without leaking sensitive information.

On the infrastructure side, managing a fleet of heterogeneous devices—each with a distinct capability profile—requires a model registry that stores multiple quantized, pruned, and distilled variants of a base diffusion model, along with metadata about their performance characteristics. A device-side profiler can periodically sample available memory, CPU utilization, and thermal headroom, then request the most appropriate model variant from a cloud or edge portal. This approach echoes the logic of content delivery networks but adapted for compute payloads. An important design consideration is version consistency: when a user interacts with a multimodal assistant that relies on a compressed vision-language diffusion model, the alignment between the image encoder, text encoder, and denoising decoder must be preserved across variant swaps. Automated validation pipelines that run standardized multimodal benchmarks on each new compressed variant before deploying to the fleet are therefore essential to prevent silent regressions in fairness or safety.

5. Sustainability, Fairness, and Governance in Compressed Systems

The pursuit of efficiency in resource-aware compression carries both sustainability promises and risks. On one hand, moving diffusion workloads from energy-hungry data centers to edge devices can significantly reduce the carbon footprint associated with data transmission and cloud inference, especially when the edge device itself is powered by renewable energy or operates under strict thermal caps [14]. The lifecycle analysis must account for the training cost of multiple compressed variants and the overhead of the compression pipeline itself, including the energy expended during distillation, quantization-aware fine-tuning, and neural architecture search. As a community, we lack standardized benchmarks that report the total carbon cost from initial data collection through compressed model deployment, making it difficult to assess whether a tenfold reduction in inference energy justifies a twofold increase

in one-time training emissions. Research on energy and policy considerations for NLP models has highlighted that efficiency improvements can paradoxically increase overall resource consumption through demand elasticity, a phenomenon likely to replicate in generative multimodal systems unless accompanied by usage governance [15].

Compression also amplifies fairness concerns that are already present in large diffusion models. Studies on compressed discriminative models have revealed that pruning and quantization can disproportionately degrade performance on underrepresented subgroups, effectively causing the model to forget rare patterns that are not reinforced by the majority training distribution [12]. In multimodal diffusion, this risk extends to linguistic minorities, regional accents, cultural artifacts, and assistive augmentations for people with disabilities. A compressed model that reliably generates high-quality images of European cityscapes may fail to render architectural details characteristic of Southeast Asian villages if the compression scheduler has eliminated channels that encoded those visual features. Mitigating such disparities demands fairness-aware compression objectives that penalize degradation on protected subgroups, coupled with diverse evaluation datasets that span geographies, languages, and abilities. From a governance standpoint, regulatory frameworks such as the EU AI Act and emerging standards on algorithmic accountability require that the entire model supply chain, including compression operations, remains auditable, with documented impact assessments for affected communities [19].

The governance of compressed multimodal diffusion also raises questions about content provenance and misuse. Lighter models are easier to distribute surreptitiously on consumer devices, potentially facilitating the generation of deepfakes or disinformation without server-side safety filters. This shifts the locus of responsibility to on-device watermarking, output filtering, and user consent mechanisms that must themselves be lightweight enough to coexist with the compressed generator. Hardware-backed attestation and secure enclaves can enforce usage policies, but they increase silicon area and cost, creating a governance trade-off between device affordability and societal safety. As edge-native training on domestic accelerators becomes more prevalent [13], nations may impose localization requirements on model compression pipelines, further complicating the global harmonization of fairness and sustainability standards.

6. Case Illustrations and Cross-Domain Comparisons

To concretize the system-level perspective, we consider three representative deployment scenarios: a mobile multimodal assistant for content creation, a hearable device for real-time audio-visual translation, and an industrial embedded system for sensor-conditioned anomaly visualization. In the mobile assistant case, a user captures a photo and provides a text prompt to generate a stylized visual storyboard using a compressed latent diffusion model. The system leverages MobileNet-style feature extractors for the image encoder and a distilled denoising U-Net that runs eight sampling steps on the GPU. Quantization-aware training is restricted to layers that do not intersect with the cross-modal attention pathway, preserving the semantic alignment between image regions and text tokens [7]. The orchestrator monitors battery level and, below 20 percent, switches to a more aggressively pruned variant that sacrifices some decorative background detail but maintains facial and textual fidelity, demonstrating graceful degradation.

In the hearable translation scenario, a compression strategy must account for the tight coupling between audio encoding, neural machine translation, and speech synthesis, all of which must complete within the latency budget of a natural conversation. Here, progressive

distillation reduces the synthesis model to a single sampling step, converting the diffusion process into a deterministic mapping that, while less expressive, achieves sub-10-millisecond synthesis latency on a dedicated DSP [8]. The memory footprint is further reduced by merging the audio encoder and the diffusion decoder into a shared backbone using low-rank adaptation. However, the compressed model exhibits a higher word error rate for tonal languages, highlighting the fairness challenge discussed earlier. Retaining a slightly larger variant for supported tonal languages and switching dynamically based on language identification is a system-level compromise that respects both latency constraints and equity.

The industrial embedded system employs a diffusion model conditioned on sensor time-series data from vibration, temperature, and pressure gauges to generate a visual heatmap of anticipated equipment failure. Because the platform is a microcontroller-class device with 512 KB of SRAM, the model is compressed to an extreme degree using ternary weight quantization and structural pruning guided by a sensitivity analysis that preserves the sensor fusion pathways [5]. The sampling trajectory is further shortened by training a dedicated lightweight predictor that skips early noise levels entirely. Although the generative quality is lower than a server-side equivalent, the trade-off is acceptable because the task requires anomaly localization, not photorealistic rendering. This case illustrates how domain-specific utility metrics must drive compression decisions rather than generic image quality scores.

Cross-domain comparisons reveal recurring patterns. Compression techniques that work well for visual modalities—such as channel pruning in convolutional layers—do not transfer seamlessly to audio spectrogram generation or sensor fusion, where temporal dependencies and frequency-domain representations are more vulnerable to quantization drift. Similarly, distillation that minimizes the KL divergence between teacher and student output distributions can inadvertently amplify biases present in the teacher, a phenomenon observed across vision, speech, and text modalities. These cross-domain commonalities underscore the value of developing compression-aware multimodal frameworks that treat each modality's sensitivity as a tunable parameter in the overall system optimization, rather than applying a uniform compression recipe.

7. Conclusion

This paper has examined resource-aware diffusion model compression through a system-level lens that integrates architectural design, hardware-software co-deployment, sustainability accounting, fairness analysis, and governance frameworks. We argued that effective compression for mobile and embedded multimodal intelligence cannot be achieved by stacking algorithmic techniques in isolation; it requires an orchestrated approach that trades off fidelity, latency, energy, and equity in a context-aware manner. The analysis identified key structural tensions, including the conflict between representational efficiency and cross-modal alignment, the need for dynamic model variant selection under varying resource profiles, the double-edged sustainability implications of moving inference to the edge, and the amplification of subgroup disparities through aggressive pruning and quantization.

Looking forward, several priorities emerge. First, the community must develop standardized lifecycle assessment protocols for compressed generative models that capture both operational and embedded carbon costs, allowing system designers to make genuinely informed trade-offs. Second, fairness-aware compression objectives should become mandatory components of multimodal AI pipelines, supported by curated evaluation benchmarks that cover underrepresented languages, cultures, and sensor conditions. Third, device-side orchestration frameworks need standardized interfaces to negotiate model variant selection with cloud

operators, including protocols for secure model delivery, usage auditing, and revocation in case of safety failures. Fourth, the integration of edge-native training ecosystems with domestic accelerator supply chains demands international dialogue on technology sovereignty, export controls, and collaborative governance without stifling innovation. By addressing these priorities, the research community can steer the deployment of compressed diffusion intelligence toward a future that is not only efficient but also equitable, robust, and environmentally responsible.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.

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

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

4. 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.

5. Zhu, C., Han, S., Mao, H., & Dally, W. J. (2016). Trained ternary quantization. arXiv preprint arXiv:1612.01064.

6. Li, Y., Wang, Y., Chen, X., Li, Z., & Wang, Y. (2023). SnapFusion: Text-to-image diffusion model on mobile devices within two seconds. Advances in Neural Information Processing Systems, 36.

7. Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., & Keutzer, K. (2023). Q-Diffusion: Quantizing diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 17535–17544).

8. Salimans, T., & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512.

9. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & 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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4510–4520).

11. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105–6114). PMLR.

12. Hooker, S., Courville, A., Clark, G., Dauphin, Y., & Frome, A. (2019). What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248.

13. 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.

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

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

16. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

17. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

18. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175–1191).

19. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

20. David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Regev, S., Rhodes, R., Wang, T., & Warden, P. (2021). TensorFlow Lite Micro: Embedded machine learning on TinyML systems. Proceedings of Machine Learning and Systems, 3, 800–811.