Strategies for Deploying High-Fidelity Generative Diffusion Models at Scale under Computational and Energy Constraints

Maria Jensen[1], Lars Holm[2], Søren Kristensen[3], and Chand Aline[1]

[1] Aarhus University, Department of Computer Science, Aarhus, Denmark
chand.aline@post.au.dk
[2] Technical University of Denmark, Department of Applied Mathematics and Computer Science, Lyngby, Denmark
lars.holm@dtu.dk
[3] University of Copenhagen, Department of Computer Science, Copenhagen, Denmark
soren.kristensen@di.ku.dk

Abstract. Generative diffusion models have emerged as a powerful class of probabilistic models capable of synthesizing high-fidelity data across diverse domains, including images, audio, video, and multimodal content. Their iterative denoising processes, grounded in stochastic differential equations and Markovian transitions, allow them to learn complex data distributions with remarkable accuracy. However, the deployment of these models in practical, large-scale applications is severely constrained by their computational and memory requirements, particularly in the context of big data environments where datasets are massive, heterogeneous, and continuously evolving. Quantization, the process of reducing the numerical precision of model parameters and activations, has recently gained attention as a crucial strategy to mitigate these challenges, offering substantial reductions in memory footprint, computational overhead, and energy consumption while maintaining the generative fidelity of the models. This survey provides a comprehensive analysis of quantization strategies for generative diffusion models, spanning post-training quantization, quantization-aware training, mixed-precision schemes, dynamic and adaptive bitwidth methods, and hybrid approaches that integrate complementary compression techniques such as pruning, low-rank factorization, and weight clustering. We systematically explore the mathematical foundations of quantization in the context of iterative denoising, formalizing error propagation, step-dependent sensitivity, and stochastic effects induced by low-precision arithmetic. Furthermore, we examine the system-level and hardware-aware implications of quantization, including memory alignment, tensor-core acceleration, cache utilization, distributed computation, and energy efficiency, highlighting the trade-offs that arise in heterogeneous big data pipelines. The survey also emphasizes the challenges unique to generative diffusion models, such as the amplification of quantization noise across timesteps, sensitivity to out-of-distribution and heterogeneous datasets, robustness to adversarial or rare events, and the complex interactions between precision reduc-

tion and model architecture. We review evaluation metrics and benchmarking strategies for quantized diffusion models, discussing traditional measures such as Fréchet Inception Distance and Inception Score alongside perceptual fidelity metrics, diversity assessments, robustness analyses, and hardware-aware efficiency measures. Finally, we outline open research directions and emerging trends, including adaptive and data-dependent quantization policies, integration with complementary compression methods, cross-platform optimization, fairness and robustness assurance, and energy-efficient design. By synthesizing advances across algorithmic, mathematical, hardware, and system-level perspectives, this survey provides a holistic framework for understanding, evaluating, and deploying quantized generative diffusion models in big data contexts, offering guidance for both researchers and practitioners seeking to balance computational efficiency with high-fidelity generative performance.

Keywords: Generative diffusion models, quantization, large-scale AI, big data, low-precision computing, mixed-precision training, adaptive quantization, post-training quantization, quantization-aware training, iterative denoising, memory-efficient AI, hardware-aware optimization, energy-efficient deep learning, robustness in generative models, system-level optimization, stochastic modeling, high-fidelity data generation, scalable AI pipelines, computational efficiency, model compression techniques

1 Introduction

The rapid evolution of large-scale generative models, particularly diffusion-based architectures, has ushered in a new era of artificial intelligence where the capacity to generate realistic images, coherent text, structured data, and even multimodal outputs is increasingly within reach. Unlike earlier generative paradigms such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), diffusion models have demonstrated remarkable stability, scalability, and fidelity in producing high-quality outputs. Their success has been evident across a wide spectrum of applications, including natural language processing, computer vision, drug discovery, speech synthesis, and multimodal reasoning. Yet, this success comes at a cost: diffusion models are computationally intensive, memory-demanding, and require vast energy resources for both training and inference. This computational burden becomes particularly salient in the era of big data, where the growing scale of datasets and the demand for real-time generative capabilities necessitate efficient deployment strategies. Among various optimization techniques, quantization has emerged as one of the most promising and widely explored methods for compressing and accelerating large generative diffusion models without severely compromising their performance. Quantization, in its essence, refers to the process of reducing the numerical precision of model parameters, activations, or gradients [1]. Traditionally, deep learning models rely on floating-point representations (e.g., FP32 or FP16) to capture the fine-grained details of learned weights and activations [2]. However, this comes

with high memory and computational costs, particularly when operating at scale. By representing model components with lower-precision formats such as INT8, INT4, or even binary quantization in extreme cases, significant reductions in storage requirements, bandwidth consumption, and energy usage can be achieved. While quantization has been extensively studied in the context of discriminative models like ResNets or Transformers for classification and recognition tasks, applying it to generative diffusion models introduces unique challenges. Unlike classification networks, generative diffusion models rely on iterative denoising processes across hundreds or even thousands of timesteps, where small perturbations introduced by quantization errors can accumulate, resulting in catastrophic degradation of sample quality. This sensitivity makes quantization in generative diffusion models both a critical necessity and a formidable research challenge. In the broader landscape of big data, the urgency of addressing these challenges cannot be overstated [3]. As data continues to grow exponentially in domains such as medical imaging, genomics, autonomous driving, climate modeling, and social media, the demand for generative AI systems that can learn from and synthesize vast amounts of information is rapidly intensifying. However, deploying large diffusion models on big data platforms introduces severe bottlenecks. Data centers, edge devices, and cloud services face growing pressures to handle high-throughput workloads under strict constraints of latency, bandwidth, and energy consumption. In many scenarios, particularly at the edge, resource limitations preclude the deployment of full-precision diffusion models altogether [4]. Quantization offers a pathway to bridge this gap, enabling generative diffusion models to scale down to resource-constrained environments while still preserving their generative fidelity [5]. Furthermore, quantization is not merely an optimization technique for deployment; it is increasingly being integrated into the training process itself, where quantization-aware training (QAT) and mixed-precision strategies allow models to adapt to low-precision constraints during learning, thereby mitigating accuracy degradation. The study of quantization for generative diffusion models in the context of big data encompasses a broad and complex set of considerations. These include the design of quantization schemes (uniform, non-uniform, logarithmic, mixed-precision), the choice of quantization granularity (per-layer, per-channel, per-tensor), the interplay with architectural features of diffusion models (U-Nets, attention mechanisms, time-embedding modules), and the integration of quantization with other model compression strategies (pruning, knowledge distillation, low-rank factorization). Moreover, the effectiveness of quantization cannot be measured solely in terms of model perplexity or classification accuracy, as in discriminative models; instead, evaluation must consider perceptual quality, fidelity of generated samples, distributional alignment with real data, and robustness under noisy or out-of-distribution conditions. In large-scale settings, this evaluation must also account for throughput, scalability across distributed systems, and compatibility with heterogeneous hardware accelerators such as GPUs, TPUs, and specialized AI chips [6]. Another dimension of complexity arises from the intersection of quantization and data heterogeneity in big data environments [7]. Unlike controlled

benchmark datasets, big data is characterized by diverse modalities, imbalanced distributions, and high levels of noise [8]. Quantized diffusion models must not only preserve generative quality but also maintain robustness when synthesizing from noisy, incomplete, or biased datasets. This raises new research questions: How does quantization affect generative fairness, diversity, and bias mitigation [9]? Can quantization be adapted dynamically based on data characteristics, computational budgets, or user requirements [10]? How do quantized models interact with distributed storage and retrieval systems in large-scale pipelines? These questions highlight the need for a comprehensive survey that integrates insights from model compression, big data systems, and generative modeling theory. In summary, quantization of generative diffusion models represents a rapidly emerging field situated at the intersection of large-scale AI, optimization, and systems research. It is driven by the twin imperatives of efficiency and scalability in an era defined by big data [11]. While the potential benefits of quantization are profound—enabling the deployment of powerful generative models across diverse hardware platforms, reducing carbon footprints, and democratizing access to generative AI—the path forward is fraught with challenges. The sensitivity of diffusion processes to quantization noise, the need for new evaluation metrics, and the integration with heterogeneous big data ecosystems all demand sustained research effort [12]. This survey aims to provide a comprehensive and detailed examination of these issues, synthesizing recent advances, identifying open challenges, and charting future directions for the quantization of generative diffusion models in the context of big data.

2 Background and Mathematical Foundations

To fully appreciate the unique challenges and opportunities associated with the quantization of generative diffusion models in the context of big data, it is essential to formalize the mathematical underpinnings of both diffusion models and quantization techniques. In this section, we provide an extended and detailed exposition of the relevant mathematical principles. We begin with a rigorous description of diffusion models, their forward and reverse stochastic processes, and then proceed to establish the formalism of quantization in the context of high-dimensional generative learning [13]. Throughout, we emphasize the interplay between mathematical precision and practical scalability, highlighting the points where quantization most critically impacts generative performance [14].

2.1 Generative Diffusion Models

Diffusion models are generative probabilistic models defined by a sequence of transformations that gradually corrupt a data sample $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ into a noise distribution through a forward diffusion process, followed by a learned reverse denoising process that reconstructs the data. More formally, given an initial data distribution $p_{\text{data}}(\mathbf{x})$, the forward process is defined as a Markov chain of length T that gradually adds Gaussian noise:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I}\big), \quad t = 1, \ldots, T, \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule with $\beta_t \in (0,1)$ controlling the noise intensity at step t. Through recursive application, one can derive the closed-form expression for sampling $\mathbf{x}_t$ at any timestep t directly from $\mathbf{x}_0$:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. This process ensures that as $t \to T$, the distribution of $\mathbf{x}_t$ approaches a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The generative process, or reverse process, seeks to approximate the time-reversal of the above Markov chain. Specifically, it defines a parameterized distribution $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ intended to reverse the corruption at each step:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\big)[15]. \tag{3}$$

The mean $\mu_\theta(\mathbf{x}_t, t)$ is predicted by a neural network, often a U-Net with attention layers, conditioned on the noisy sample $\mathbf{x}_t$ and the timestep t [16]. Training typically involves learning to predict the added noise $\boldsymbol{\epsilon}$ directly, where $\mathbf{x}_t$ can be reparameterized as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

The learning objective minimizes the expected squared error between the true noise $\boldsymbol{\epsilon}$ and the predicted noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right][17]. \tag{5}$$

This elegant mathematical formulation underlies the expressive power of diffusion models but simultaneously exposes them to vulnerabilities when subject to quantization noise.

2.2 Quantization as a Mathematical Operator

Quantization can be formalized as a function $Q : \mathbb{R} \to \mathcal{C}$, where $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ is a finite set of discrete representable values, typically integers or fixed-point numbers [18]. For a real-valued weight $w \in \mathbb{R}$, the quantized representation is:

$$Q(w) = \Delta \cdot \text{clip}\left(\text{round}\left(\frac{w}{\Delta}\right), q_{\min}, q_{\max}\right), \tag{6}$$

where $\Delta > 0$ is the quantization step size (also called the scale), and $(q_{\min}, q_{\max})$ defines the representable integer range (e.g., $[-127, 127]$ for 8-bit signed integers). The quantization error is then:

$$e(w) = Q(w) - w, \tag{7}$$

which is generally non-Gaussian, non-uniform, and potentially correlated across model parameters [19]. In the context of generative diffusion models, the accumulation of such errors across iterative timesteps poses significant risks.

2.3 Quantization of Model Parameters and Activations

Consider a diffusion model parameterized by weights $\mathbf{W}$ and biases $\mathbf{b}$ across L layers. Each weight tensor $\mathbf{W}_\ell \in \mathbb{R}^{d_\ell}$ is quantized as:

$$\hat{\mathbf{W}}_\ell = Q(\mathbf{W}_\ell) = \Delta_\ell \cdot \text{round}\left(\frac{\mathbf{W}_\ell}{\Delta_\ell}\right), \tag{8}$$

where Δ_ℓ may be chosen per-layer, per-channel, or per-tensor depending on the granularity of quantization. Similarly, activations $\mathbf{a}_\ell$ at each layer are quantized:

$$\hat{\mathbf{a}}_\ell = Q(\mathbf{a}_\ell). \tag{9}$$

The forward propagation in the quantized model thus becomes:

$$\hat{\mathbf{a}}_{\ell+1} = \sigma\big(\hat{\mathbf{W}}_\ell \hat{\mathbf{a}}_\ell + \hat{\mathbf{b}}_\ell\big), \tag{10}$$

where $\sigma(\cdot)$ denotes the non-linear activation function [20]. The difference between $\hat{\mathbf{a}}_{\ell+1}$ and its full-precision counterpart $\mathbf{a}_{\ell+1}$ defines the cumulative quantization-induced distortion.

2.4 Error Propagation in the Diffusion Process

The most striking challenge arises from the iterative nature of diffusion models [**?**]. Let us denote the quantized reverse process distribution as:

$$p_{\theta,Q}(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \mu_{\theta,Q}(\mathbf{x}_t, t), \Sigma_{\theta,Q}(\mathbf{x}_t, t)\big), \tag{11}$$

where $\mu_{\theta,Q}$ and $\Sigma_{\theta,Q}$ are computed using quantized weights and activations. The deviation in the generative trajectory can be formalized as the divergence between the quantized and full-precision joint distributions:

$$D_{\text{KL}}\big(p_\theta(\mathbf{x}_{0:T}) \,\|\, p_{\theta,Q}(\mathbf{x}_{0:T})\big), \tag{12}$$

which measures the cumulative effect of quantization noise across all timesteps. Unlike in discriminative models, where a single forward pass determines the output, here the generative trajectory involves repeated application of quantized operations, causing the error to potentially amplify as t decreases.

2.5 Quantization in the Big Data Regime

In big data scenarios, quantization becomes not just a model compression technique but also a systems-level optimization [21]. Let N denote the number of data samples, M the model size (in parameters), and B the bitwidth used after quantization. The total memory requirement can be expressed as:

$$\text{Memory}(Q) = \frac{M \cdot B}{8} \text{ bytes}, \tag{13}$$

compared to the full-precision requirement Memory(FP32) $= 4M$ bytes [22]. Similarly, the computational cost for matrix multiplications, a dominant operation in diffusion models, scales approximately as:

$$\text{Cost}(Q) \propto \mathcal{O}\big(N \cdot M \cdot B\big), \tag{14}$$

where lower B leads to substantial acceleration on hardware optimized for low-precision arithmetic (e.g., INT8 tensor cores) [23]. Thus, quantization not only alleviates the memory bottleneck but also reduces the energy footprint, a key concern in large-scale deployments [24].

2.6 Summary

From the above formulation, it becomes evident that quantization in generative diffusion models introduces a delicate trade-off: reducing bitwidth yields substantial memory and computational savings, but at the cost of potential error amplification across timesteps [25]. In the big data context, where efficiency is paramount, this trade-off becomes even more critical. The mathematical framework presented here sets the stage for a deeper exploration of quantization techniques, their empirical performance, and their integration into large-scale systems for generative modeling.

3 Taxonomy of Quantization Techniques for Diffusion Models

The field of quantization for large-scale generative diffusion models has witnessed an explosion of research directions, each proposing novel strategies to reduce the computational and memory burden while striving to retain the perceptual fidelity of generated outputs. Unlike conventional discriminative tasks such as classification or regression, where accuracy metrics provide a direct measure of performance degradation under quantization, generative diffusion models require an expanded set of metrics encompassing sample quality, diversity, fidelity to data distribution, and robustness to long iterative generation steps [26]. This complexity motivates the need for a systematic taxonomy of quantization techniques, one that is sufficiently broad to capture the diversity of approaches while being granular enough to identify subtle distinctions. In this section, we present such a taxonomy, categorizing quantization methods according to their level of precision, granularity, adaptivity, and integration into the training process [27]. Each category is associated with distinct mathematical trade-offs, computational efficiencies, and suitability for deployment in big data environments. At the highest level, quantization strategies may be broadly divided into *post-training quantization (PTQ)* and *quantization-aware training (QAT)*. PTQ techniques are particularly appealing for rapid deployment in resource-constrained settings, as they require no modification to the training pipeline and can be applied directly to pre-trained diffusion models. However, PTQ is typically more

prone to accuracy degradation, especially for models as sensitive as diffusion architectures, where even minor distortions can accumulate across thousands of denoising steps. QAT, in contrast, incorporates quantization effects during the training process, allowing the model to learn resilience against quantization noise. While QAT often achieves higher performance and robustness, it comes with significantly greater training costs, which in the context of diffusion models can be prohibitively expensive, particularly when dealing with massive datasets and billions of parameters. Beyond this binary division, quantization strategies also vary in terms of the precision levels employed. Traditional approaches rely on *uniform quantization*, where the step size Δ is constant across all representable values, leading to a linear partitioning of the real axis. This method is straightforward and highly hardware-friendly, but it can be suboptimal for distributions of weights and activations that exhibit non-uniform statistics, such as heavy-tailed distributions common in diffusion models. In response, researchers have explored *non-uniform quantization* methods, including logarithmic quantization, learned step-size quantization, and even vector quantization, where clusters of parameters are represented by shared centroids. Such non-uniform schemes are more complex to implement on hardware but often yield superior trade-offs between precision and compression. Another critical dimension concerns *bitwidth adaptivity*, where mixed-precision quantization allows different layers or components of the model to be quantized with different bitwidths based on their sensitivity to noise. For example, attention modules and time-embedding layers in diffusion models are often more sensitive and therefore require higher precision compared to convolutional layers in the U-Net backbone. The granularity of quantization is another key axis of differentiation. *Per-tensor quantization*, where a single scale is applied across the entire tensor, is computationally efficient but may poorly capture variability across channels. *Per-channel quantization*, by contrast, assigns independent scaling factors to each channel, significantly improving fidelity at the cost of increased complexity in implementation. Even finer-grained approaches include *per-group* and *per-element quantization*, though these are less common due to their prohibitive overhead [28]. Importantly, in the context of big data, such fine-grained methods may lead to severe inefficiencies in distributed systems, where synchronization overheads negate the theoretical benefits [29]. Thus, system-level considerations must accompany the mathematical optimization of quantization granularity. To provide a comprehensive overview of these strategies, we summarize the major categories of quantization techniques and their key properties in Table 1. This table encapsulates distinctions in terms of training requirements, computational cost, hardware compatibility, and impact on generative quality [30]. For clarity and to ensure it fits neatly within the width of the page, we use the `resizebox` environment to scale it appropriately.

This taxonomy underscores the diversity of available approaches and the trade-offs they entail [31]. For instance, PTQ provides a pragmatic solution for rapid deployment on edge devices where retraining is infeasible, but it risks catastrophic failure in iterative generative tasks such as diffusion. Conversely, QAT provides robustness but demands extensive training resources, which may be im-

Method	Precision Type	Granularity	Training Requirement	Hardware Compatibility	Impact on Generative Quality
Post-Training Quantization (PTQ)	Uniform (INT8/INT4)	Per-tensor or Per-channel	None (applied after training)	High (widely supported)	Moderate to severe degradation in sensitive diffusion models
Quantization-Aware Training (QAT)	Uniform or Non-uniform	Per-channel, Mixed-precision	Retraining with quantization simulation	Moderate (requires custom kernels)	High fidelity preservation, resilient to noise
Mixed-Precision Quantization	Hybrid (e.g., INT8 + FP16)	Layer-wise adaptive	Requires precision sensitivity analysis	High (with specialized accelerators)	Significant savings with minimal quality loss
Non-Uniform Quantization	Logarithmic, Learned step-size	Per-tensor or Per-group	Optional retraining (depending on method)	Limited (less hardware-friendly)	Better for heavy-tailed distributions, robust under noise
Vector Quantization	Codebook-based	Per-group or Per-block	Requires codebook training	Low to moderate (hardware complexity)	High compression but sensitive to codebook size
Dynamic Quantization	On-the-fly scaling	Per-tensor	None (applied dynamically during inference)	High (supported in runtime engines)	Moderate performance, suitable for real-time big data streams

Table 1. Taxonomy of quantization methods for diffusion models, categorized by precision type, granularity, training requirements, hardware support, and their qualitative impact on generative quality.

practical when handling massive datasets in big data pipelines. Mixed-precision quantization offers a middle ground, aligning well with modern heterogeneous accelerators that natively support FP16 and INT8 arithmetic. Meanwhile, non-uniform quantization techniques, though often less hardware-friendly, offer the promise of aligning quantization bins more closely with the statistical properties of weights and activations, thereby preserving fidelity under extreme compression [32]. Vector and dynamic quantization approaches extend these ideas further, enabling adaptive compression and scaling suited to data heterogeneity in big data settings [33]. In conclusion, the taxonomy presented here not only provides a structured map of the quantization landscape but also illuminates the unique challenges and opportunities specific to generative diffusion models. The iterative denoising process amplifies even minute quantization artifacts, demanding careful calibration of bitwidth, granularity, and adaptivity. Moreover, the big data context magnifies the importance of scalability, hardware compatibility, and system-level considerations, ensuring that quantization research must go beyond isolated model compression techniques and instead integrate with the broader ecosystem of distributed computation and heterogeneous accelerators. This synthesis of mathematical rigor, hardware-awareness, and systems integration forms the basis for the following sections, where we dive deeper into empirical results, case studies, and emerging research directions.

4 System-Level Implications and Hardware-Aware Quantization

The quantization of large generative diffusion models is not solely a mathematical or algorithmic problem; it has profound implications at the system level, particularly when these models are deployed in the context of big data. As the scale of both model parameters and dataset sizes grows exponentially, the interaction between quantization strategies and hardware constraints becomes increasingly critical. Modern data centers, edge devices, and heterogeneous AI accelerators impose diverse restrictions on memory bandwidth, compute throughput, and energy consumption. Quantization directly addresses these limitations by reducing numerical precision and thereby decreasing memory footprint, memory access costs, and computational intensity [34]. However, this reduction comes with systemic trade-offs: quantization may necessitate additional bookkeeping for scaling factors, introduce alignment issues for tensor cores, or require specialized kernels to maintain throughput. Consequently, understanding and optimizing the system-level behavior of quantized diffusion models is essential for

practical deployment, particularly in real-time or large-scale applications where the volume of data processed can be massive. From a systems perspective, the memory hierarchy plays a central role in the performance of quantized models. High-dimensional diffusion models, such as those with billions of parameters, are often memory-bound rather than compute-bound. In such cases, reducing parameter precision from FP32 to INT8 or lower can dramatically decrease the volume of data transferred between GPU memory and compute units, leading to substantial speedups. Let M denote the total number of model parameters, B_{full} the bitwidth of full-precision storage, and B_{quant} the bitwidth after quantization. Then, the memory footprint reduction can be expressed as:

$$\text{Reduction Factor} = \frac{M \cdot B_{\text{full}}}{M \cdot B_{\text{quant}}} = \frac{B_{\text{full}}}{B_{\text{quant}}}. \tag{15}$$

For INT8 quantization, this implies a $4\times$ reduction relative to FP32, while INT4 quantization achieves an $8\times$ reduction [35]. Beyond raw memory footprint, quantization also impacts cache utilization, memory alignment, and the ability to store intermediate activations during iterative denoising. These effects are particularly pronounced in diffusion models, where hundreds or thousands of intermediate latent representations must be maintained or recomputed across timesteps [36]. Efficient memory layouts, fused kernels, and careful scheduling of quantized operations are therefore essential to realize the theoretical benefits of reduced precision in practice [37]. Another dimension of system-level considerations involves the interaction between quantization and compute efficiency on specialized hardware. Modern accelerators, including NVIDIA tensor cores, Google's TPUs, and other AI-specific chips, offer native support for low-precision arithmetic such as FP16, BF16, INT8, and even INT4 in emerging architectures. Quantized diffusion models can leverage these capabilities to accelerate matrix multiplications, convolutions, and attention mechanisms. However, these benefits are contingent upon careful alignment of data structures, tensor shapes, and kernel implementations [38]. Misalignment or improper tiling can lead to underutilized compute units and diminished throughput. Additionally, heterogeneous systems that combine CPUs, GPUs, and memory hierarchies introduce further complexities, such as latency penalties for data transfers and contention for shared memory resources [39]. These factors necessitate a holistic approach where algorithmic quantization choices are co-designed with hardware capabilities and system-level scheduling strategies [40]. To visualize these system-level interactions, we provide a simple vertical schematic using TikZ that captures the memory-compute-energy trade-offs and the flow of quantized diffusion computations from data input to output generation. The figure is designed to fit the width of the page and maintain a vertical orientation for clarity in tall model pipelines.

In the context of big data, these system-level considerations become even more critical. High-throughput pipelines require the ability to ingest, process, and generate outputs for millions or billions of samples efficiently [42]. Quantization reduces the per-sample memory footprint, allowing larger batches to be

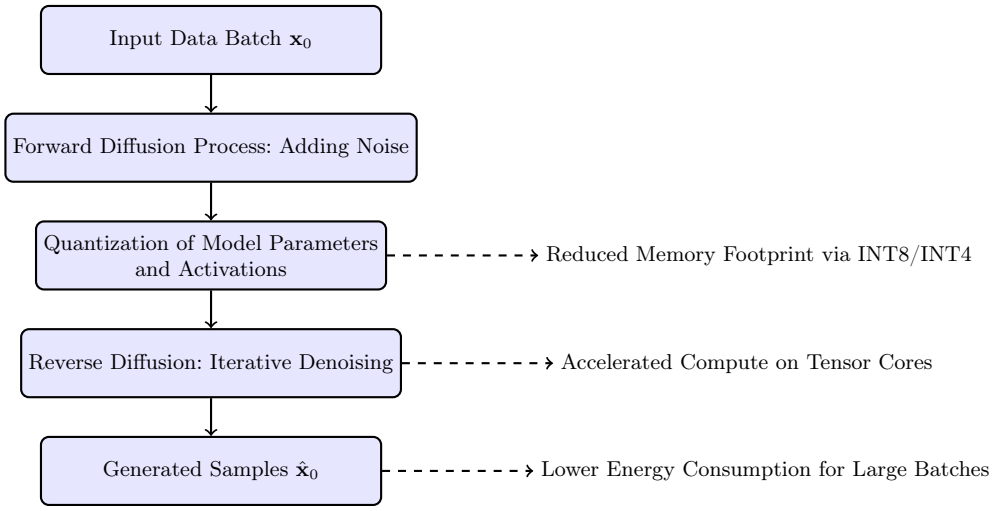

Fig. 1. Vertical schematic of system-level workflow for quantized diffusion models [41]. The pipeline illustrates the flow from input data through forward diffusion, quantization of parameters and activations, reverse iterative denoising, and final generated samples. Dashed arrows highlight system-level benefits of quantization including reduced memory footprint, accelerated compute, and lower energy consumption.

processed concurrently, thereby improving throughput and utilization of accelerator resources [43]. Furthermore, lower-precision arithmetic reduces power consumption, a crucial factor for sustainable deployment of large-scale AI systems [44]. Nevertheless, these advantages come with trade-offs. Quantization-induced errors, if not properly managed, can degrade generative fidelity and exacerbate model instability across timesteps [45]. Consequently, optimal deployment requires careful profiling of both computational performance and output quality, often necessitating a hybrid approach where critical layers remain in higher precision while less sensitive layers are aggressively quantized. In summary, system-level and hardware-aware quantization strategies are indispensable for scaling generative diffusion models in big data environments. Memory, compute, and energy constraints interact in complex ways with the mathematical properties of quantized diffusion, necessitating co-design across algorithm, architecture, and hardware layers. The vertical TikZ schematic presented here provides a high-level visualization of these interactions, emphasizing the key points at which quantization impacts the end-to-end pipeline. Recognizing and optimizing these interactions ensures that large-scale diffusion models can be deployed efficiently without compromising the generative fidelity that makes them uniquely powerful [46].

5 Challenges and Open Research Directions in Quantized Diffusion Models

Despite the remarkable progress in quantization techniques for generative diffusion models, a series of profound challenges persist, particularly when considering deployment in big data contexts. These challenges arise from the interplay between algorithmic sensitivity, iterative denoising processes, heterogeneous hardware constraints, and the scale and diversity of real-world datasets. Addressing these challenges is crucial to realize the full potential of quantized diffusion models, and they also reveal multiple avenues for future research that are currently underexplored [47]. One of the most fundamental challenges is the inherent sensitivity of diffusion models to quantization noise [48]. Unlike discriminative models, where a single forward pass produces an output that can tolerate minor perturbations, generative diffusion models operate through long iterative denoising chains, often comprising hundreds or even thousands of steps [49]. Mathematically, let $\hat{\mathbf{x}}_{t-1} = \mu_{\theta,Q}(\hat{\mathbf{x}}_t, t) + \Sigma_{\theta,Q}(\hat{\mathbf{x}}_t, t)\boldsymbol{\epsilon}$ denote the quantized reverse step at timestep t. The cumulative error introduced by quantization, defined as $\mathbf{e}_t = \hat{\mathbf{x}}_t - \mathbf{x}_t$, can propagate and amplify across timesteps, potentially leading to severe degradation in sample quality. This amplification is particularly problematic in early timesteps, where large deviations can misguide subsequent denoising operations and result in mode collapse or distorted outputs [50]. Understanding the precise dynamics of error propagation under quantization is a nontrivial problem that involves both stochastic process theory and high-dimensional numerical analysis. It remains an open research question how to design quantization schemes that minimize cumulative error while maintaining computational efficiency. Another significant challenge is the heterogeneity of modern big data. Unlike controlled benchmarks, real-world datasets exhibit a high degree of variability in scale, distribution, and modality [51]. For instance, generative models trained on medical imaging data must handle varying resolutions, imaging modalities, and noise levels, whereas models for autonomous driving must account for diverse lighting, weather conditions, and sensor characteristics [52]. Quantization exacerbates these challenges because lower-precision representations may fail to capture subtle statistical variations or outlier patterns, leading to artifacts or bias in generated samples. Formally, let $p_{\text{data}}(\mathbf{x})$ denote the underlying data distribution [53]. Quantization effectively introduces a perturbed generative distribution $p_{\theta,Q}(\mathbf{x})$ that may deviate from $p_{\text{data}}(\mathbf{x})$ not only in mean squared error but also in higher-order moments and distributional properties. Ensuring that $p_{\theta,Q}(\mathbf{x})$ retains fidelity across diverse data regimes, while maintaining computational tractability, is a pressing research frontier that intersects robust statistics, fairness in AI, and generative modeling theory [54]. Hardware heterogeneity introduces another layer of complexity. Low-precision arithmetic is implemented differently across CPUs, GPUs, TPUs, and custom AI accelerators, leading to variations in numerical behavior. While INT8 and FP16 are widely supported, emerging INT4 or mixed-precision schemes may not have consistent hardware support, resulting in platform-dependent performance and accuracy. Furthermore, large-scale pipelines often involve distributed computa-

tion across multiple nodes or clusters, where communication bandwidth, synchronization overhead, and memory alignment issues can interact with quantization-induced errors in unpredictable ways. The challenge is therefore twofold: designing quantization-aware algorithms that are robust to both stochastic noise and system-level constraints, and developing profiling tools that accurately predict end-to-end performance and quality trade-offs in heterogeneous big data environments. A particularly promising yet underexplored research direction involves adaptive and dynamic quantization [55]. Traditional approaches apply static quantization schemes determined before or during training, often using heuristics such as layer sensitivity analysis or per-channel statistics [56]. However, in dynamic data environments, where data distributions shift over time or model utilization varies across workloads, static quantization may be suboptimal [57]. Dynamic quantization strategies could adapt bitwidths, step sizes, or scaling factors on the fly based on real-time feedback from model performance, input characteristics, or hardware constraints. For example, critical layers such as attention heads in U-Net architectures may temporarily retain higher precision during high-variance input conditions, while convolutional backbone layers may use aggressive quantization during low-variance states [58]. Mathematically, this requires defining a control policy $\pi(\ell, t, \mathbf{x}_t)$ that maps layer indices, timestep, and input statistics to optimal quantization parameters, which can be learned or optimized using reinforcement learning or gradient-based methods. Developing efficient and stable policies for dynamic quantization remains a significant open challenge [59]. Another avenue for future research lies in the integration of quantization with other model compression techniques [60]. Pruning, low-rank factorization, knowledge distillation, and weight clustering each offer complementary benefits, and their combination with quantization could yield substantial improvements in memory efficiency, throughput, and energy consumption [61]. However, the interactions among these methods are highly nontrivial in diffusion models [62]. For instance, pruning may increase sensitivity to quantization by removing redundancy, while low-rank factorization could alter the statistical distribution of weights, impacting step-size selection in non-uniform quantization schemes. Formally analyzing these interactions requires a joint optimization framework, potentially combining second-order sensitivity analysis, quantization error modeling, and stochastic process analysis of diffusion dynamics. Such frameworks could provide principled guidelines for co-designing compression pipelines that balance fidelity, efficiency, and robustness [63]. Finally, evaluation metrics for quantized diffusion models remain an open problem [64]. While traditional metrics such as FID, IS, or PSNR provide some measure of generative quality, they often fail to capture subtle distortions induced by quantization or the distributional divergence in high-dimensional latent spaces. Moreover, system-level metrics such as throughput, energy efficiency, and memory footprint must be integrated with quality metrics to obtain a holistic view of trade-offs [65]. There is a growing need for composite metrics that jointly quantify perceptual quality, distributional fidelity, and computational efficiency, particularly in big data pipelines where millions of samples must be generated

and validated at scale. In summary, the challenges and open research directions in quantized diffusion models are multifaceted, spanning algorithmic, statistical, hardware, and system-level dimensions. Understanding and mitigating error propagation, accommodating heterogeneous and dynamic data, leveraging hardware efficiently, exploring adaptive and hybrid quantization strategies, and developing robust evaluation frameworks are all essential to advance the field [66]. Addressing these challenges will require interdisciplinary efforts, combining insights from machine learning theory, stochastic processes, numerical analysis, hardware architecture, and big data engineering [67]. The next generation of quantized generative diffusion models will likely emerge from such holistic approaches, achieving a balance between efficiency, fidelity, and scalability that is currently unattainable with existing techniques.

6 Applications and Impact of Quantized Diffusion Models in Big Data

The practical implications of quantized generative diffusion models extend far beyond theoretical considerations, encompassing a wide array of applications in domains where big data is pervasive. By reducing memory footprint, computational overhead, and energy consumption, quantization enables the deployment of large-scale generative models in scenarios that were previously infeasible, ranging from cloud-based high-throughput pipelines to resource-constrained edge devices. These capabilities open the door to transformative applications across fields such as healthcare, autonomous systems, multimedia content generation, scientific research, and industrial analytics, while simultaneously raising questions about scalability, fairness, and robustness. In healthcare, for example, medical imaging datasets are notoriously large and heterogeneous, encompassing modalities such as magnetic resonance imaging (MRI), computed tomography (CT), X-ray, and ultrasound, often accompanied by complex metadata. Generative diffusion models trained on these datasets can synthesize high-resolution images to augment scarce datasets, perform modality translation, or denoise corrupted scans. However, full-precision models with billions of parameters are impractical for hospital servers or edge devices embedded in imaging hardware due to memory and energy constraints. Quantization allows these models to operate efficiently on limited resources while maintaining the fidelity necessary for clinical decision support. Moreover, iterative denoising inherent to diffusion models benefits from reduced memory usage when processing large batches of patient scans, facilitating real-time inference and data augmentation [68]. Nevertheless, careful attention must be paid to the propagation of quantization errors, as subtle distortions in generated images could lead to misinterpretation in diagnostic contexts, emphasizing the need for domain-specific calibration and evaluation protocols. In the context of autonomous systems and robotics, quantized diffusion models enable real-time synthesis and prediction in complex environments [69]. Autonomous vehicles, drones, and industrial robots must process high-dimensional sensory input, including LIDAR point clouds, RGB-D images, and temporal

sensor streams. Generative models can provide predictive simulation, anomaly detection, or scenario augmentation to enhance training and planning pipelines. Here, quantization directly impacts latency and throughput: lower-precision operations reduce the time required to simulate multiple possible trajectories or environmental states, enabling on-device inference without offloading to high-performance cloud servers. For example, INT8 or mixed-precision quantization allows iterative denoising in real-time while maintaining the diversity and coherence of generated environmental scenarios [70]. This capability is particularly relevant for big data contexts, where continuous streams of sensor data must be processed, compressed, and synthesized for decision-making under tight computational budgets. Content generation in multimedia domains also benefits from quantized diffusion models. High-resolution image synthesis, video generation, text-to-image conversion, and multimodal content creation often require models with billions of parameters to capture complex structures and correlations [71]. When operating at scale—such as generating large image datasets for training or real-time rendering in interactive applications—quantization provides the necessary computational and memory efficiency [72]. Reduced precision enables larger batch sizes, faster sampling, and lower energy consumption, all critical for industrial-scale deployment. Moreover, in big data contexts such as social media, e-commerce, or digital entertainment, quantized generative models facilitate rapid personalization and adaptation to user-specific data, allowing services to dynamically generate content tailored to individual preferences [73]. Here, the challenge lies in balancing efficiency with quality: aggressive quantization may introduce artifacts or reduce fidelity, potentially impacting user experience or brand perception, necessitating careful design and tuning of quantization schemes [74]. Scientific research and simulation is another domain where quantized diffusion models have significant impact. Fields such as climate modeling, astrophysics, and computational chemistry generate enormous datasets representing complex spatiotemporal phenomena [75]. Generative models can be employed to interpolate sparse measurements, simulate potential future states, or explore high-dimensional parameter spaces. Quantization enables the scaling of these models to handle petabyte-scale datasets while running on clusters with constrained memory per node. In climate modeling, for instance, quantized diffusion models can generate high-resolution predictions for temperature, precipitation, and other environmental variables across multiple temporal horizons, facilitating ensemble forecasts and uncertainty quantification. However, maintaining the fidelity of physical laws and statistical correlations under reduced precision is nontrivial, requiring careful calibration of quantization parameters and potentially hybrid precision schemes where critical computations are maintained at higher bitwidths. Industrial analytics, including predictive maintenance, supply chain optimization, and anomaly detection, also benefit from quantized diffusion models. Large-scale sensor networks, IoT deployments, and manufacturing lines produce massive streams of high-dimensional data. Generative models can synthesize plausible scenarios, impute missing measurements, or forecast future operational states. Quantization enables these models to run efficiently on

edge devices embedded in industrial equipment or on centralized servers processing terabytes of sensor data, thus reducing latency and operational costs. Furthermore, big data in industrial contexts often exhibits heterogeneity and non-stationarity, highlighting the importance of adaptive quantization strategies that dynamically allocate precision based on input characteristics, model sensitivity, or resource availability [76]. Beyond specific application domains, the societal and environmental implications of quantized diffusion models are profound. Lowering the computational and energy requirements of large generative models directly reduces the carbon footprint associated with AI research and deployment, a critical consideration in the era of large-scale AI and climate consciousness. Additionally, democratizing access to high-quality generative capabilities through quantization allows smaller organizations, academic institutions, and developing regions to leverage advanced AI technologies without requiring supercomputing infrastructure [77]. However, these benefits also underscore the need for robust evaluation and ethical oversight. As generative models are deployed in critical or sensitive applications, quantization-induced artifacts or biases may inadvertently propagate misinformation, exacerbate social inequalities, or compromise safety, emphasizing the need for comprehensive governance, monitoring, and auditing mechanisms [78]. In conclusion, quantized diffusion models have transformative potential across a spectrum of big data applications, from healthcare and autonomous systems to multimedia content, scientific simulation, and industrial analytics. By enabling efficient computation, memory reduction, and energy savings, quantization makes large-scale generative models practically deployable, allowing them to operate on real-world datasets of unprecedented size and complexity. Nevertheless, the deployment of these models must carefully navigate the trade-offs between efficiency, fidelity, and robustness [79]. The confluence of algorithmic design, hardware-aware optimization, and domain-specific considerations defines a rich landscape of research and application opportunities, promising to extend the reach of generative AI while confronting the practical realities of big data and resource-constrained environments.

7 Evaluation Metrics and Benchmarking Strategies for Quantized Diffusion Models

The evaluation of quantized generative diffusion models is a multifaceted challenge that extends far beyond conventional measures of model performance used in discriminative tasks. Unlike classification or regression problems, where a single accuracy score or loss function can provide a reliable assessment, generative models require metrics that capture both statistical fidelity and perceptual quality across complex, high-dimensional distributions. The introduction of quantization adds an additional layer of complexity, as reduced precision can introduce subtle artifacts, distort distributional properties, and interact with the iterative dynamics of the diffusion process. Consequently, a comprehensive benchmarking framework must evaluate models along multiple axes, including sample quality, diversity, distributional alignment, computational efficiency, memory footprint,

energy consumption, and robustness under heterogeneous data conditions [80]. In this section, we provide an extensive discussion of the current state of evaluation metrics, the limitations of existing approaches, and potential directions for developing standardized benchmarks tailored to quantized diffusion models [81]. One of the most widely used metrics in generative modeling is the Fréchet Inception Distance (FID), which measures the distance between feature distributions of real and generated samples [82]. Formally, if $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$ denote the multivariate Gaussian approximations of real and generated feature distributions, FID is computed as:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\Big). \tag{16}$$

While FID provides a useful measure of distributional alignment, it is sensitive to sample size, feature extraction networks, and preprocessing, and it does not directly capture perceptual artifacts introduced by quantization [83]. For diffusion models, where thousands of iterative denoising steps are employed, quantization can introduce subtle spatial or temporal distortions that may not significantly affect FID but substantially degrade perceptual fidelity [84]. Complementary metrics such as the Inception Score (IS) and Kernel Inception Distance (KID) attempt to address some of these limitations, but they similarly focus on distributional properties rather than the accumulation of errors over iterative generation. Recent research emphasizes the need for perceptually-aware metrics, potentially leveraging learned perceptual similarity (LPIPS) measures or human-in-the-loop evaluation to capture visual fidelity in images or coherence in multimodal data. Another crucial dimension of evaluation is diversity. Generative models must avoid mode collapse and produce outputs that reflect the true variability of the underlying data distribution [85]. For quantized diffusion models, diversity may be compromised by the reduced representational precision of weights and activations [86]. Quantization-induced correlations or rounding effects can reduce variability across generated samples, particularly in sensitive layers such as attention heads or time embeddings [87]. Metrics for diversity include multi-scale structural similarity (MS-SSIM), entropy-based measures, and coverage metrics that assess how well generated samples span the support of real data [88]. A comprehensive benchmark should combine diversity measures with fidelity metrics to detect cases where high-quality outputs are generated at the expense of limited variation. Computational and system-level metrics are equally important in the context of big data applications. Key performance indicators include throughput (samples per second), memory usage (bytes or gigabytes per batch), energy consumption (joules per sample or watt-hours per training/inference cycle), and latency (time per iteration or end-to-end pipeline). Quantization directly affects these metrics: lower bitwidth reduces memory footprint and bandwidth requirements, accelerates matrix multiplications, and decreases energy consumption due to fewer bit operations. However, the relationship between bitwidth, batch size, and hardware utilization is nontrivial. For instance, aggressive INT4 quantization may require specialized kernels or memory alignment strategies, and in distributed settings, inter-node communication over-

heads may dominate performance gains. A robust evaluation framework must capture these interactions, ideally using hardware-agnostic metrics normalized across platforms to enable meaningful comparisons between models and quantization strategies [89]. Robustness evaluation is another critical consideration, particularly in big data environments characterized by heterogeneity, noise, and distributional shifts [90]. Let $\mathcal{D}_{\text{train}}$ denote the training dataset and $\mathcal{D}_{\text{test}}$ a test set drawn from a potentially shifted distribution. Quantized diffusion models must maintain generative fidelity under these conditions, ensuring that low-precision representations do not amplify errors when encountering rare or out-of-distribution samples. Adversarial robustness, sensitivity to input perturbations, and stability across multiple timesteps are all aspects that require careful assessment. Formal metrics may include worst-case deviation measures, perturbation-induced error accumulation across diffusion steps, and empirical robustness tests across synthetic or real-world data shifts [91]. Benchmarking strategies for quantized diffusion models must also account for the interplay between algorithmic choices and system-level optimizations. This includes comparisons between post-training quantization, quantization-aware training, mixed-precision strategies, and adaptive quantization schemes. Benchmark datasets should cover a wide range of modalities—images, text, audio, video, and multimodal data—to ensure generality of conclusions. Moreover, evaluations must consider both small-scale controlled experiments for reproducibility and large-scale pipelines reflective of production scenarios. Composite metrics that integrate generative fidelity, diversity, computational efficiency, and robustness offer a promising path toward standardized benchmarks that capture the multidimensional impact of quantization. Finally, the development of automated and reproducible benchmarking suites is an emerging direction. These suites may include pre-trained diffusion models, reference datasets, standardized quantization protocols, and automated scripts for computing metrics across multiple axes. Incorporating hardware-in-the-loop evaluation, energy profiling, and latency measurement tools will further enhance the relevance of benchmarks for real-world deployment. The combination of algorithmic, perceptual, and system-level evaluations provides a holistic view, enabling researchers and practitioners to make informed trade-offs between efficiency, fidelity, and scalability [92]. In conclusion, evaluating quantized diffusion models is a complex, multidimensional problem that requires metrics and benchmarks capturing both the generative quality and system-level efficiency [93]. Traditional metrics such as FID and IS must be complemented with diversity, perceptual fidelity, robustness, and computational performance measures to fully understand the impact of quantization [94]. Establishing standardized benchmarking frameworks, integrating hardware-aware evaluation, and developing composite metrics represent critical steps toward advancing the field, ensuring that quantized diffusion models can be deployed effectively in big data applications while maintaining the high fidelity and robustness that make these models uniquely powerful.

8 Future Directions and Emerging Trends in Quantized Generative Diffusion Models

As generative diffusion models continue to scale in size, complexity, and application scope, the role of quantization as a fundamental enabler of efficient, deployable AI becomes increasingly pronounced [95]. While current research has made substantial strides in post-training quantization, quantization-aware training, mixed-precision methods, and adaptive bitwidth schemes, the field is still in its infancy when it comes to fully exploiting the interplay between model architecture, data characteristics, hardware capabilities, and system-level optimizations. The future of quantized diffusion models will be shaped by interdisciplinary advances, combining innovations in algorithmic design, stochastic modeling, hardware acceleration, and software infrastructure, with an emphasis on achieving scalable, robust, and energy-efficient generative systems suitable for real-world big data environments [96]. One emerging direction involves the development of *fine-grained, layer-adaptive, and data-dependent quantization strategies*. Current approaches often assign static bitwidths or per-channel scaling factors based on sensitivity analysis or empirical heuristics [97]. However, as datasets become larger and more heterogeneous, static schemes may underperform, either by allocating excessive precision to insensitive layers or insufficient precision to critical components, such as attention mechanisms, time embeddings, or conditional input modules. Dynamic quantization policies that adapt bitwidths in real-time, conditioned on input characteristics, model state, or timestep in the diffusion process, could provide superior fidelity-efficiency trade-offs. Mathematically, this involves optimizing a control function $\pi : (\ell, t, \mathbf{x}_t) \mapsto B_\ell(t, \mathbf{x}_t)$ that maps layer index, timestep, and input statistics to an optimal precision assignment, balancing memory, throughput, and generative quality. Learning such policies may leverage reinforcement learning, gradient-based optimization, or Bayesian decision frameworks, integrating both model uncertainty and hardware constraints into the optimization loop. Another critical research frontier is the integration of *quantization with complementary compression techniques* such as pruning, low-rank factorization, knowledge distillation, and weight clustering [98]. While each method individually reduces memory or computational overhead, their interactions in iterative diffusion processes are complex and poorly understood [99]. For instance, pruning may amplify sensitivity to quantization noise by eliminating redundant pathways that would otherwise absorb rounding errors, while low-rank decomposition can alter the statistical distribution of weight matrices, affecting step-size selection in non-uniform quantization schemes [100]. Developing principled joint optimization frameworks that account for cumulative error propagation across diffusion steps, stochastic noise in iterative denoising, and distributional fidelity is essential. Such frameworks may combine second-order sensitivity analysis, stochastic process modeling, and information-theoretic measures to guide the co-design of hybrid compression pipelines capable of operating efficiently under extreme big data workloads. Hardware-aware quantization and cross-platform optimization constitute another key trend [101]. As AI accelerators continue to diversify, with GPUs, TPUs, FPGAs, and custom ASICs

offering varying support for INT8, INT4, BF16, and emerging sub-bit formats, it becomes imperative to co-design quantization strategies with hardware capabilities in mind [102]. This includes not only arithmetic precision but also memory layouts, cache hierarchies, instruction-level parallelism, and interconnect bandwidth. In heterogeneous big data environments, distributed inference pipelines must carefully manage synchronization overhead, communication bottlenecks, and load balancing to fully exploit low-precision acceleration. The development of standardized hardware-aware benchmarking suites, incorporating throughput, latency, energy consumption, and memory utilization alongside generative quality metrics, will be crucial for guiding research and deployment strategies. The integration of *robustness and fairness considerations* in quantized diffusion models is an additional emerging focus [103]. As models are increasingly deployed in sensitive domains such as healthcare, autonomous systems, and scientific research, it is essential to ensure that quantization does not introduce bias, exacerbate distributional skew, or compromise reliability. Research directions include designing quantization-aware regularization schemes, developing robustness-aware loss functions that account for quantization noise, and creating auditing frameworks for fairness and bias evaluation in low-precision generative outputs. Moreover, quantization strategies may need to adapt dynamically in response to dataset shifts, adversarial inputs, or rare event scenarios to maintain stability and fidelity, particularly in real-time big data pipelines. Finally, there is a growing interest in *energy-efficient and sustainable AI* as a research driver. Large generative models are energy-intensive, and full-precision training or inference can contribute significantly to carbon emissions. Quantization directly reduces the number of bit operations, memory accesses, and energy consumption, but there remains considerable room for optimization through mixed-precision pipelines, approximate arithmetic, dynamic scaling, and model sparsity [104]. Future work may explore co-designing diffusion models with energy-aware objectives, where quantization parameters, model architecture, and training schedules are jointly optimized to minimize energy while preserving generative fidelity [105, 106]. This direction aligns with broader societal goals of sustainable AI deployment and democratizing access to large-scale generative models in resource-constrained settings. In conclusion, the future of quantized generative diffusion models lies at the intersection of algorithmic innovation, hardware co-design, robustness, and sustainability [107]. Emerging trends emphasize adaptive, layer- and data-aware quantization, integration with complementary compression techniques, cross-platform optimization, fairness and robustness assurance, and energy-efficient design [108]. Addressing these challenges requires a holistic, interdisciplinary approach that simultaneously considers stochastic modeling, numerical precision, hardware constraints, and big data pipeline requirements. The realization of such comprehensive strategies promises to enable the deployment of powerful generative diffusion models at scale, making high-fidelity AI generation accessible, efficient, and robust in the era of big data.

9 Conclusion

The quantization of generative diffusion models represents a critical frontier in the intersection of machine learning, systems optimization, and big data analytics. Over the past several years, diffusion models have demonstrated unprecedented capabilities in high-fidelity generative tasks, ranging from image and video synthesis to multimodal content generation and scientific simulations [109]. However, the practical deployment of these models at scale has been hindered by their enormous computational and memory requirements [110]. Quantization offers a transformative solution, enabling models with billions of parameters to operate efficiently on heterogeneous hardware, reduce energy consumption, and process massive datasets in real time. Yet, the integration of quantization into diffusion pipelines is far from trivial. Unlike traditional discriminative models, the iterative denoising structure of diffusion models amplifies even minor numerical errors, necessitating careful consideration of error propagation, layer sensitivity, and timestep-dependent distortions. The balance between model efficiency and generative fidelity thus becomes a central challenge, one that requires a nuanced understanding of both algorithmic and system-level dynamics [111].

This survey has highlighted the multifaceted nature of quantization in diffusion models, beginning with a rigorous exposition of the mathematical foundations of forward and reverse diffusion processes, and proceeding to formalize quantization operators and error propagation mechanisms. We have explored a comprehensive taxonomy of quantization strategies, including post-training quantization, quantization-aware training, mixed-precision methods, and adaptive dynamic schemes. Each approach entails distinct trade-offs in terms of memory savings, computational acceleration, hardware compatibility, and generative quality, and the choice of technique must consider both the model architecture and the characteristics of the target dataset. Granularity of quantization—whether per-tensor, per-channel, or per-element—further shapes the performance-efficiency trade-off, particularly in large-scale models where memory alignment, kernel efficiency, and system throughput are critical factors. Through tables, diagrams, and formal equations, we have endeavored to capture the breadth of these considerations, providing a structured framework for understanding the interplay between quantization design and diffusion model performance.

The challenges and open research directions in this field are substantial and span multiple dimensions. Error accumulation across iterative timesteps, sensitivity to heterogeneous and out-of-distribution data, hardware-specific numerical behavior, and interactions with complementary model compression techniques all pose intricate problems that require interdisciplinary solutions. Dynamic and layer-adaptive quantization strategies, hybrid precision pipelines, joint optimization frameworks integrating pruning or low-rank decomposition, and robust evaluation metrics are among the most promising avenues for future work. Additionally, big data considerations—such as high throughput, large batch processing, distributed computation, and sustainable energy consumption—necessitate system-level co-design, where algorithmic choices, hardware capabilities, and

pipeline architectures are optimized jointly to achieve practical, scalable, and reliable generative performance.

The applications and impact of quantized diffusion models in real-world big data environments further underscore their transformative potential. In healthcare, efficient generative models enable high-resolution imaging, modality translation, and data augmentation, while maintaining the fidelity necessary for clinical decision support. In autonomous systems and robotics, quantized models allow real-time scenario simulation and anomaly detection. Multimedia content generation benefits from reduced latency and increased throughput, supporting industrial-scale personalized content creation. In scientific research and industrial analytics, quantization permits the handling of petabyte-scale datasets, enabling high-fidelity interpolation, forecasting, and scenario exploration. Across these domains, the deployment of quantized models not only improves computational and energy efficiency but also democratizes access to high-performance generative AI, allowing smaller organizations and resource-constrained environments to leverage state-of-the-art models.

Evaluation metrics and benchmarking frameworks remain critical enablers of progress, providing standardized means to quantify generative fidelity, diversity, robustness, and system-level performance. Traditional metrics such as FID, IS, and KID, while informative, must be complemented with perceptual similarity measures, robustness analyses, diversity assessments, and computational profiling to fully capture the effects of quantization. Emerging trends in benchmarking emphasize hardware-aware evaluation, cross-platform reproducibility, and composite metrics that jointly assess quality and efficiency. Such frameworks will be indispensable for guiding research, informing deployment decisions, and establishing best practices for quantized diffusion models in diverse big data contexts.

Looking ahead, the future of quantized generative diffusion models lies at the convergence of algorithmic sophistication, hardware-aware engineering, adaptive and dynamic precision management, robustness and fairness assurance, and energy-efficient design. The field is poised for significant breakthroughs as researchers explore adaptive quantization policies, integrate compression techniques, develop cross-platform co-optimization strategies, and design sustainable AI systems capable of generating high-fidelity outputs at scale. The realization of these advances promises not only to make generative AI accessible and efficient in big data environments but also to redefine the boundaries of what is computationally feasible, transforming the landscape of AI applications across science, industry, and society.

In conclusion, quantization is not merely a tool for compression or acceleration; it is a foundational component for the practical deployment of large-scale generative diffusion models. By systematically reducing precision while preserving generative fidelity, quantization enables these models to scale to unprecedented sizes and datasets, unlocking applications that were previously impractical. The interplay between mathematical rigor, system-level optimization, hardware compatibility, and big data considerations defines the frontier of this research area. Continued innovation in quantization methods, evaluation frame-

works, and hardware-software co-design will be essential for realizing the full potential of generative diffusion models, ensuring that they can operate efficiently, reliably, and sustainably across the increasingly diverse and demanding applications of the modern era.

References

1. Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
2. Kexun Zhang, Xianjun Yang, William Yang Wang, and Lei Li. Redi: efficient learning-free diffusion inference via trajectory retrieval. In *International Conference on Machine Learning*, pages 41770–41785. PMLR, 2023.
3. Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7827–7839, 2024.
4. Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion through temporal attention decomposition. *Transactions on Machine Learning Research*.
5. Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.
6. Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.
7. Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.
8. Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Kai Li, and Song Han. Distrifusion: Distributed parallel inference for high-resolution diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7193, 2024.
9. Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1):19–41, 2024.
10. Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18456–18466, 2023.
11. Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.
12. Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36, 2024.
13. Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9892–9902, 2024.
14. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

15. Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models, 2022.
16. Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8775–8784, 2024.
17. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
18. Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17535–17545, October 2023.
19. Hengyuan Ma, Li Zhang, Xiatian Zhu, and Jianfeng Feng. Accelerating score-based generative models with preconditioned diffusion sampling. In *European Conference on Computer Vision*, pages 1–16. Springer, 2022.
20. Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. δ-dit: Accelerating diffusion transformers without training via denoising property alignment.
21. Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. Fora: Fast-forward caching in diffusion transformer acceleration. *arXiv preprint arXiv:2407.01425*, 2024.
22. Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems, 2024.
23. Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
24. Yonghao Yu, Shunan Zhu, Huai Qin, and Haorui Li. Boostdream: Efficient refining for high-quality text-to-3d generation from multi-view diffusion. *arXiv preprint arXiv:2401.16764*, 2024.
25. Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α-(de) blending: A minimalist deterministic diffusion model. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–8, 2023.
26. Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation.
27. Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
28. Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper, 2018.
29. Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds, 2022.
30. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
31. Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. Tfmq-dm: Temporal feature maintenance quantization for diffusion models, 2024.

32. Jiarui Fang, Jinzhe Pan, Jiannan Wang, Aoyu Li, and Xibo Sun. Pipefusion: Patch-level pipeline parallelism for diffusion transformers inference. *arXiv preprint arXiv:2405.14430*, 2024.
33. Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.
34. Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems*, 2023.
35. Enshu Liu, Xuefei Ning, Huazhong Yang, and Yu Wang. A unified sampling framework for solver searching of diffusion probabilistic models. In *The Twelfth International Conference on Learning Representations*, 2023.
36. Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.
37. Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.
38. Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.
39. Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.
40. Mingxing Peng, Kehua Chen, Xusen Guo, Qiming Zhang, Hongliang Lu, Hui Zhong, Di Chen, Meixin Zhu, and Hai Yang. Diffusion models for intelligent transportation systems: A survey. *arXiv preprint arXiv:2409.15816*, 2024.
41. Mohsen Zand, Ali Etemad, and Michael Greenspan. Diffusion models with deterministic normalizing flow priors. *arXiv preprint arXiv:2309.01274*, 2023.
42. Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W. Battaglia. Generating images with sparse representations, 2021.
43. Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5517–5526, 2023.
44. Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.
45. Zhicong Tang, Shuyang Gu, Chunyu Wang, Ting Zhang, Jianmin Bao, Dong Chen, and Baining Guo. Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. *arXiv preprint arXiv:2312.11459*, 2023.
46. Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023.
47. Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, et al. Cache me if you can: Accelerating diffusion models through block caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6211–6220, 2024.
48. Fu-Yun Wang, Zhaoyang Huang, Xiaoyu Shi, Weikang Bian, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning. *arXiv preprint arXiv:2402.00769*, 2024.
49. Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, December 2023.

50. Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A simple early exiting framework for accelerated sampling in diffusion models. *arXiv preprint arXiv:2408.05927*, 2024.
51. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
52. Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403, 2021.
53. Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.
54. Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems, 2023.
55. Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.
56. Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023.
57. Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A simple early exiting framework for accelerated sampling in diffusion models. In *Forty-first International Conference on Machine Learning*, 2024.
58. Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
59. William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
60. Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.
61. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
62. Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
63. Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.
64. Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
65. Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024.

66. Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. *arXiv preprint arXiv:2502.21079*, 2025.
67. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.
68. Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models, 2023.
69. Yuxiang Fu, Qi Yan, Lele Wang, Ke Li, and Renjie Liao. Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. *arXiv preprint arXiv:2503.09950*, 2025.
70. Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.
71. Yuanzhi Zhu, Xingchao Liu, and Qiang Liu. Slimflow: Training smaller one-step diffusion models with rectified flow. *arXiv preprint arXiv:2407.12718*, 2024.
72. Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization framework for vision transformers with twin uniform quantization, 2022.
73. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
74. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
75. Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36, 2024.
76. Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-cache: Accelerating diffusion transformer via layer caching. *arXiv preprint arXiv:2406.01733*, 2024.
77. Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. *arXiv preprint arXiv:2310.03270*, 2023.
78. Yuewei Yang, Xiaoliang Dai, Jialiang Wang, Peizhao Zhang, and Hongbo Zhang. Efficient quantization strategies for latent diffusion models, 2023.
79. chengzeyi. Stable fast. https://github.com/chengzeyi/stable-fast, 2024.
80. Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.
81. Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
82. Shane Barratt and Rishi Sharma. A note on the inception score, 2018.
83. Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025.

84. Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024.
85. Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024.
86. Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.
87. Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation. *arXiv preprint arXiv:2312.12491*, 2023.
88. Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.
89. Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
90. Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pages 7197–7206. PMLR, 2020.
91. Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.
92. Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
93. Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.
94. Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 22–39, 2024.
95. Tiankai Hang and Shuyang Gu. Improved noise schedule for diffusion training. *arXiv preprint arXiv:2407.03297*, 2024.
96. Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
97. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
98. Geng Yang, Yanyue Xie, Zhong Jia Xue, Sung-En Chang, Yanyu Li, Peiyan Dong, Jie Lei, Weiying Xie, Yanzhi Wang, Xue Lin, et al. Sda: Low-bit stable diffusion acceleration on edge fpgas. 2023.
99. Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao

Tian, Hua Wu, and Haifeng Wang. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10135–10145, June 2023.

100. Chang Chen, Fei Deng, Kenji Kawaguchi, Caglar Gulcehre, and Sungjin Ahn. Simple hierarchical planning with diffusion. *arXiv preprint arXiv:2401.02644*, 2024.
101. Ratko Pilipović, Patricio Bulić, and Vladimir Risojević. Compression of convolutional neural networks: A short survey. In *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–6. IEEE, 2018.
102. Hui Zhang, Zuxuan Wu, Zhen Xing, Jie Shao, and Yu-Gang Jiang. Adadiff: Adaptive step selection for fast diffusion. *arXiv preprint arXiv:2311.14768*, 2023.
103. Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025.
104. Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. Grid diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8734–8743, 2024.
105. Kaixuan Huang, Yukang Yang, Kaidi Fu, Yanyi Chu, Le Cong, and Mengdi Wang. Latent diffusion models for controllable rna sequence generation. *arXiv preprint arXiv:2409.09828*, 2024.
106. Yassine Zniyed, Thanh Phuong Nguyen, et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4358–4370, 2024.
107. Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models, 2025.
108. Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In *International Conference on Machine Learning*, pages 9847–9856. PMLR, 2020.
109. Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021.
110. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
111. Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.