Q-DRIFT: QUANTIZATION-AWARE DRIFT CORRECTION FOR DIFFUSION MODEL SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) is a practical path to deploy large diffusion models, but quantization noise can accumulate over the denoising trajectory and degrade generation quality. We propose Q-Drift, a sampler-side correction that aims to preserve the intended sampling marginals through a deterministic drift adjustment motivated by generalized marginal-preserving SDEs. Q-Drift uses the calibrated conditional residual variance of quantization error to determine a correction factor at each step. Our SDXL study shows that calibration with as few as 10 paired full-precision/quantized runs remains effective. The resulting sampler correction is plug-and-play with common samplers, diffusion models, and PTQ methods, while incurring negligible overhead at inference. Across six diverse text-to-image models (spanning DiT and U-Net), three samplers (Euler, flow-matching, DPM-Solver++), and two PTQ methods (SVDQuant, MixDQ), Q-Drift improves FID over the corresponding quantized baseline in most settings, with up to 4.79 FID reduction on PixArt-Σ (SVDQuant W3A4), while preserving CLIP scores. Code is available at `https://anonymous.4open.science/r/Q-Drift`.

1 INTRODUCTION

Diffusion models have become a dominant paradigm for high-quality image generation (Ho et al., 2020; Song et al., 2021b), and modern text-to-image systems such as Stable Diffusion have shown their practical effectiveness in real-world generation (Rombach et al., 2022; Podell et al., 2024). Their iterative denoising and growing scale, however, incur substantial computational and memory costs, which pose significant challenges for deployment in resource-constrained or latency-critical settings (Shen et al., 2025; Karras et al., 2022).

Post-training quantization (PTQ) converts a pretrained model into a low-bit representation using only a small calibration set, avoiding the training cost of quantization-aware training (QAT), and it has become a practical and widely adopted way to compress diffusion models. Nevertheless, quantization-induced degradation remains a persistent challenge, especially under aggressive low-bit settings or with fast, few-step samplers (Shang et al., 2023; Li et al., 2023b; He et al., 2023b). Recent work mitigates it by improving the fidelity of the quantized model to its full-precision counterpart: SVDQuant absorbs extreme outliers into low-rank components (Li et al., 2025), and MixDQ uses metric-decoupled sensitivity analysis with integer-programming-based mixed-precision allocation (Zhao et al., 2024). These approaches substantially reduce model-level quantization error, yet noticeable degradation in generation quality often persists in practice (He et al., 2023b; Yao et al., 2024).

Many PTQ methods optimize layer-wise reconstruction at individual sampling steps. These objectives do not directly control the accumulated sampling error. Generation, however, is iterative: the denoiser output at one step determines the input to the next, so small local errors accumulate along the trajectory (Li & van der Schaar, 2024), and quantization error therefore interacts with the sampling process rather than remaining purely step-local.

Motivated by this, we introduce *Quantization-Aware Drift Correction (Q-Drift)*, a sampler-side method that approaches quantized diffusion sampling from a distributional perspective and aims to preserve the intended sampling marginals. The key idea is to use the conditional residual variance of quantization error to determine a deterministic drift adjustment, drawing on the drift–diffusion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pairing in generalized marginal-preserving SDEs (Song et al., 2021b; Karras et al., 2022). This connects a statistic available from calibration to a correction of the sampling dynamics.

Q-Drift translates this design into a lightweight correction: one precomputed scalar factor per sampling step. It complements model-level PTQ methods such as SVDQuant and MixDQ (Li et al., 2025; Zhao et al., 2024), requires no retraining or weight changes, and incurs negligible inference overhead. The same rescaling principle supports adaptations to Euler, flow-matching, and DPM-Solver++ samplers. Our SDXL study further shows that as few as 10 paired full-precision/quantized calibration runs suffice to retain the correction's effectiveness.

We evaluate Q-Drift on six text-to-image models spanning DiT and U-Net architectures, three samplers, and two PTQ methods. It improves FID in all seven settings of our main evaluation at comparable CLIP scores, with a reduction of up to 4.79 FID on PixArt-Σ. Our SDXL ablations support conditional residual variance and a single scalar factor per step as practical design choices. Together, these results demonstrate that a lightweight sampler-side adjustment can improve distributional fidelity across diverse quantized diffusion models.

2 Related Work

2.1 Diffusion Sampling and SDE Perspectives

Diffusion models generate images by iteratively denoising from a simple Gaussian and have become a central paradigm for high-fidelity generation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021). A continuous-time formulation connects sampling to reverse-time SDEs and the probability-flow ODE (Anderson, 1982; Haussmann & Pardoux, 1986; Song & Ermon, 2019; Song et al., 2021b), providing a principled lens on discretization, schedules, and the accuracy–efficiency tradeoff at inference (Karras et al., 2022; Shen et al., 2025), and it has driven extensive work on retraining-free acceleration through better discretization and higher-order ODE/SDE solvers (Song et al., 2021a; Liu et al., 2022; Lu et al., 2022a; Zhang & Chen, 2023; Zhao et al., 2023; Lu et al., 2023; Xue et al., 2023) as well as distillation into few-step generators and consistency mappings (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Sauer et al., 2024; Yin et al., 2024). While these lines of work substantially reduce the number of function evaluations (NFEs), deployment still faces significant memory and compute costs for large diffusion backbones (Rombach et al., 2022; Peebles & Xie, 2023), motivating the complementary compression techniques discussed below.

2.2 Quantizing Diffusion Models

PTQ is a standard approach for compressing large generative models and is particularly attractive for diffusion models because it avoids retraining (Nagel et al., 2020; Shang et al., 2023; Li et al., 2023b). A key challenge is that activation and output statistics shift substantially across denoising timesteps, and sequential generation can amplify small approximation errors, making calibration and distribution alignment central to maintaining generation quality (Shang et al., 2023; Li et al., 2023b; Liu et al., 2024; Wang et al., 2024a; Li et al., 2023a; He et al., 2023b; Huang et al., 2024; Chu et al., 2024; Li & van der Schaar, 2024). Accordingly, many PTQ methods adapt reconstruction-based quantization to the multi-timestep setting and combine it with diffusion-specific calibration or mixed-precision choices (Li et al., 2021; Shang et al., 2023; Li et al., 2023b; Wang et al., 2024a; Sun et al., 2024), while others target temporal and conditioning pathways or modality-specific backbones (Huang et al., 2024; Tian et al., 2024). For text-to-image diffusion, additional sensitivity in cross-attention and conditioning modules motivates progressive calibration, activation relaxation, metric-driven mixed precision, and distribution-aware grouping (Pandey et al., 2023; Tang et al., 2024; Zhao et al., 2024; Ryu et al., 2025), and when PTQ quality drops at aggressive bitwidths, QAT or selective fine-tuning becomes increasingly important (He et al., 2023a; Wang et al., 2024b; Ryu et al., 2024; Guo et al., 2024). As backbones have moved from U-Nets to transformers (DiTs), quantization research has extended to transformer-specific sensitivity analyses and PTQ pipelines built for timestep-dependent behaviors and structured outliers (Peebles & Xie, 2023; Bao et al., 2023; Yang et al., 2024; Wu et al., 2024; Chen et al., 2025; Dong & Zhang, 2024; Liu & Zhang, 2024), with aggressive settings (e.g., 4-bit) increasingly emphasizing outlier- and distribution-aware quantization (Li et al., 2025; Lee et al., 2025; Zhao et al., 2025).

2.3 Quantization Noise Compensation During Sampling

Since diffusion sampling is a sequential procedure where each denoiser output becomes the next-step input, small approximation errors can propagate and amplify across timesteps (Li & van der Schaar, 2024). For quantized diffusion models, a growing body of work explicitly models quantization noise during sampling and proposes step- or timestep-dependent correction mechanisms. PTQD (He et al., 2023b) decomposes quantization noise into correlated and residual components and compensates the induced mean/variance mismatch in the denoising process. TAC-Diffusion (Yao et al., 2024) proposes a timestep-aware correction that dynamically adjusts quantization errors during inference. QNCD (Chu et al., 2024) estimates the accumulated quantization noise at run time and filters it out during sampling. D^2-DPM (Zeng et al., 2025) introduces a dual denoising mechanism to mitigate quantization-induced noise accumulation during iterative sampling. We compare these methods and our method in Sec. 5.4.

Our work is most closely related to sampler-time compensation approaches, but differs in objective and mechanism. Our objective is to preserve the intended sampling marginals. We use calibrated conditional residual variance to determine a deterministic drift rescaling motivated by generalized marginal-preserving SDEs.

3 Preliminaries

3.1 Diffusion Sampling as a Generalized SDE

Diffusion models generate samples by simulating a reverse-time stochastic process. This process is designed so that its marginal distribution $p_t(\mathbf{x})$ at each noise level matches that of a fixed forward noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b). Following the noise-level parameterization of Karras et al. (2022), we write $p_t(\mathbf{x}) \equiv p(\mathbf{x}; \sigma(t))$ with $\dot{\sigma}(t) > 0$; sampling runs backward in t, so noise decreases.

Following Karras et al. (2022), we use the generalized reverse-time SDEs below, which share the same target marginals under the exact score in continuous time:

$$d\mathbf{x} = \underbrace{-\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))\, dt}_{\text{probability-flow ODE}} \underbrace{\underbrace{-\beta(t)\sigma(t)^2 \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))\, dt}_{\text{deterministic noise decay}} \underbrace{+\sqrt{2\beta(t)}\, \sigma(t)\, d\mathbf{w}_t}_{\text{noise injection}}}_{\text{Langevin diffusion SDE}}, \quad (1)$$

where $\mathbf{w}_t$ denotes a reverse-time Wiener process with increment covariance $|dt|\mathbf{I}$. In the Langevin diffusion SDE term, $\beta(t) \geq 0$ freely controls the amount of stochasticity by pairing a deterministic noise decay (Langevin drift) with fresh noise injection (Langevin diffusion), while preserving the same marginals. This pairing motivates the deterministic correction used by Q-Drift.

3.2 Euler–Maruyama Discretization

For an Itô SDE $d\mathbf{x} = \mathbf{a}(\mathbf{x}, t)\, dt + \mathbf{b}(t)\, d\mathbf{w}_t$ on a time grid $\{t_i\}$ with $\Delta t_i \triangleq t_{i+1} - t_i$, the Euler–Maruyama discretization (Kloeden & Platen, 1992) gives

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{a}(\mathbf{x}_i, t_i)\, \Delta t_i + \mathbf{b}(t_i)\sqrt{|\Delta t_i|}\, \mathbf{z}_i, \qquad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2)$$

so the drift scales with Δt_i and the diffusion with $\sqrt{|\Delta t_i|}$.

3.3 Empirical Joint Gaussianity of Quantization Noise and Output

We follow the empirical characterization of quantization effects reported in D^2-DPM (Zeng et al., 2025). At a fixed timestep (equivalently, a fixed noise level), let $\epsilon_\theta^{(t)}$ denote the full-precision noise prediction and $\hat{\epsilon}_\theta^{(t)}$ its quantized counterpart, and define the quantization noise as $\Delta\epsilon_\theta^{(t)} \triangleq \hat{\epsilon}_\theta^{(t)} - \epsilon_\theta^{(t)}$. When the timestep is clear we drop the superscript and write $\Delta\epsilon$, or $\Delta\epsilon_i$ at step i. D^2-DPM reports that at each timestep both $\Delta\epsilon_\theta^{(t)}$ and $\hat{\epsilon}_\theta^{(t)}$ are well approximated by Gaussians, that the two are clearly correlated, and that their joint density is approximately elliptical. Accordingly, D^2-DPM models the

two jointly as

$$\begin{bmatrix} \hat{\epsilon}_\theta^{(t)} \\ \Delta\epsilon_\theta^{(t)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\hat{\epsilon}}^{(t)} \\ \mu_\Delta^{(t)} \end{bmatrix}, \begin{bmatrix} \Sigma_{\hat{\epsilon}\hat{\epsilon}}^{(t)} & \Sigma_{\hat{\epsilon}\Delta}^{(t)} \\ \Sigma_{\Delta\hat{\epsilon}}^{(t)} & \Sigma_{\Delta\Delta}^{(t)} \end{bmatrix} \right). \tag{3}$$

Under the joint Gaussian model in Eq. 3, the conditional distribution $\Delta\epsilon_\theta^{(t)} \mid \hat{\epsilon}_\theta^{(t)} = \hat{\epsilon}$ is Gaussian, i.e.,

$$\Delta\epsilon_\theta^{(t)} \mid \hat{\epsilon}_\theta^{(t)} = \hat{\epsilon} \sim \mathcal{N}\big(\mu_{\text{cond}}^{(t)}(\hat{\epsilon}), \Sigma_{\text{cond}}^{(t)}\big), \tag{4}$$

where the conditional mean and covariance admit closed-form expressions:

$$\mu_{\text{cond}}^{(t)}(\hat{\epsilon}) = \mu_\Delta^{(t)} + \Sigma_{\Delta\hat{\epsilon}}^{(t)} \big(\Sigma_{\hat{\epsilon}\hat{\epsilon}}^{(t)}\big)^{-1} \big(\hat{\epsilon} - \mu_{\hat{\epsilon}}^{(t)}\big), \tag{5}$$

$$\Sigma_{\text{cond}}^{(t)} = \Sigma_{\Delta\Delta}^{(t)} - \Sigma_{\Delta\hat{\epsilon}}^{(t)} \big(\Sigma_{\hat{\epsilon}\hat{\epsilon}}^{(t)}\big)^{-1} \Sigma_{\hat{\epsilon}\Delta}^{(t)}. \tag{6}$$

3.4 Practical Simplifications for Joint Gaussian Modeling

Estimating the full covariance blocks in Eq. 3 is impractical at this dimensionality, so, following D^2-DPM (Zeng et al., 2025), we use diagonal covariance blocks (retaining the per-element correlation between $\hat{\epsilon}_\theta^{(t)}$ and $\Delta\epsilon_\theta^{(t)}$ through a diagonal cross-covariance), an isotropic parameterization of each block, and separate statistics per timestep, since quantization effects vary across noise levels. These choices make the model estimable from a small calibration set. Appendix D.2 examines selected distributional and covariance properties underlying these approximations. For the local variance-matching construction in Sec. 4.2, we use independent auxiliary perturbations across steps.

4 Method

4.1 Overview

Q-Drift aims to preserve the intended sampling marginals under quantization. We use the empirical Gaussian modeling of quantization error and its correlation with the quantized output (Sec. 3.3–3.4) to estimate a timestep-wise conditional residual variance. This statistic determines the correction scale through the drift–diffusion pairing in Eq. 1. The resulting method rescales the deterministic update with one calibrated factor per step, without retraining.

4.2 Derivation for a First-Order Euler Sampler

We derive the correction scale by (i) identifying the quantization perturbation in the Euler update, (ii) representing its calibrated residual variability with an auxiliary Gaussian perturbation, and (iii) matching the auxiliary variance to the SDE diffusion term and using the paired drift.

σ-parameterized generalized marginal-preserving SDEs. Starting from Eq. 1 and using the noise-level parameterization $\sigma = \sigma(t)$ with $d\sigma = \dot{\sigma}(t)\, dt$, we can reparameterize the reverse-time dynamics in σ as

$$d\mathbf{x} = -\sigma\, \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)\, d\sigma - \tilde{\beta}_\sigma(\sigma)\, \sigma^2\, \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)\, d\sigma + \sigma \sqrt{2\tilde{\beta}_\sigma(\sigma)}\, d\mathbf{w}_\sigma, \tag{7}$$

where $\tilde{\beta}_\sigma(\sigma) \triangleq \beta(t)/\dot{\sigma}(t) \geq 0$, and $\mathbf{w}_\sigma$ is a Wiener process indexed by σ (so $\mathrm{Var}(d\mathbf{w}_\sigma) = |d\sigma|$). As in Eq. 1, $\tilde{\beta}_\sigma$ is a free parameter that simultaneously controls the injected noise (Langevin diffusion) and the paired deterministic noise decay (Langevin drift), preserving the same marginals (Karras et al., 2022).

Euler sampler from the probability-flow ODE. Setting $\tilde{\beta}_\sigma(\sigma) = 0$ in Eq. 7 yields the probability-flow ODE in σ:

$$d\mathbf{x} = -\sigma\, \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)\, d\sigma. \tag{8}$$

Using the noise-prediction parameterization $s_\theta(\mathbf{x}, \sigma, c) = -\epsilon_\theta(\mathbf{x}, \sigma, c)/\sigma$ to approximate the score gives $d\mathbf{x} = \epsilon_\theta(\mathbf{x}, \sigma, c)\, d\sigma$. Applying the Euler–Maruyama discretization (Eq. 2) to Eq. 8 gives

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\sigma_i\, \epsilon_\theta(\mathbf{x}_i, \sigma_i, c), \tag{9}$$

where $\Delta\sigma_i \triangleq \sigma_{i+1} - \sigma_i < 0$ is the step on the decreasing noise grid $\{\sigma_i\}$. This matches standard Euler samplers under ϵ-prediction (Karras et al., 2022).

Quantization perturbation and its residual variability. With post-training quantization, the sampler uses $\hat{\epsilon}_\theta = \epsilon_\theta + \Delta\epsilon$. Substituting this into Eq. 9 gives

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\sigma_i\, \epsilon_\theta(\mathbf{x}_i, \sigma_i, c) + \Delta\sigma_i\, \Delta\epsilon_i, \tag{10}$$

where $\Delta\epsilon_i$ denotes the quantization noise term. Let V_{σ_i} denote the calibrated conditional residual-variance statistic defined in Sec. 4.3. To translate this statistic into a correction scale, we introduce an auxiliary Gaussian perturbation $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, V_{\sigma_i}\mathbf{I})$, independent of the current state and of the auxiliary perturbations at other steps. The increment $\Delta\sigma_i\boldsymbol{\eta}_i$ has covariance $(\Delta\sigma_i)^2 V_{\sigma_i}\mathbf{I}$. We use this stochastic surrogate to design the correction.

Matching the auxiliary perturbation variance. Consider an Euler–Maruyama step of the auxiliary SDE in Eq. 7, whose diffusion coefficient is $\sigma_i\sqrt{2\tilde{\beta}_{\sigma,i}}$:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + (\cdots)\Delta\sigma_i + \sigma_i\sqrt{2\tilde{\beta}_{\sigma,i}}\,\sqrt{|\Delta\sigma_i|}\,\mathbf{z}_i, \tag{11}$$

where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $(\cdots)$ collects the drift terms. Matching the per-coordinate variance $2\tilde{\beta}_{\sigma,i}\sigma_i^2|\Delta\sigma_i|$ of its diffusion increment to the auxiliary variance $(\Delta\sigma_i)^2 V_{\sigma_i}$ yields

$$\tilde{\beta}_{\sigma,i} = \frac{|\Delta\sigma_i|}{2\sigma_i^2}\, V_{\sigma_i}. \tag{12}$$

Correction scale from the paired drift. In Eq. 7, the matched diffusion strength also sets the paired noise-decay drift. Both drift terms are proportional to the score, so their sum is $-\sigma_i(1 + \tilde{\beta}_{\sigma,i}\sigma_i)\nabla_{\mathbf{x}} \log p\, d\sigma$. This pairing motivates rescaling the deterministic Euler update by $(1 + c_i)$, with

$$c_i \;\triangleq\; \tilde{\beta}_{\sigma,i}\,\sigma_i = \frac{|\Delta\sigma_i|}{2\sigma_i}\, V_{\sigma_i}. \tag{13}$$

Final Q-Drift update. The final Q-Drift step is

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\sigma_i\,(1 + c_i)\,\hat{\epsilon}_\theta(\mathbf{x}_i, \sigma_i, c), \quad \text{with } c_i = \frac{|\Delta\sigma_i|}{2\sigma_i}\, V_{\sigma_i}. \tag{14}$$

Q-Drift preserves the direction of the quantized Euler update while adjusting its magnitude. The correction increases with the calibrated residual variance and the relative step size $|\Delta\sigma_i|/\sigma_i$, and vanishes when $V_{\sigma_i} = 0$. Q-Drift applies the resulting multiplier to the quantized prediction, requiring only the precomputed V_{σ_i} and no additional noise injection.

Extension to various samplers. We adapt this rescaling principle to flow matching and DPM-Solver++, as detailed in Appendix C.1 and Appendix C.2, respectively.

4.3 Calibration: Estimating the Conditional Variance Statistic

Offline, on a calibration set, we run a full-precision sampling trajectory and record at every step σ_i the paired outputs $\epsilon_\theta(\mathbf{x}, \sigma_i, c)$ and $\hat{\epsilon}_\theta(\mathbf{x}, \sigma_i, c)$, whose difference is the quantization noise $\Delta\epsilon_i$. The statistic V_{σ_i} measures the residual variability of quantization error after conditioning on the quantized output:

$$V_{\sigma_i} \;\triangleq\; \mathbb{E}\Big[\mathrm{Var}\big(\Delta\epsilon_i \mid \hat{\epsilon}_\theta(\mathbf{x}, \sigma_i, c)\big)\Big]. \tag{15}$$

Here, variance denotes the average per-coordinate conditional variance under the covariance model in Sec. 3.4, and the expectation is over the calibration distribution. The conditional mean minimizes squared prediction error given the quantized output; its residual variance quantifies the variability remaining after that prediction. We estimate this statistic using the fitted joint Gaussian model in Eq. 6.

Q-Drift uses this variance statistic without subtracting the conditional mean, yielding one correction factor per step. We compare this approach with prior corrections in Sec. 5.4 and examine the choice of variance statistic in Sec. 5.5.

4.4 ALGORITHM

Algorithm 1 in Appendix A summarizes the overall Q-Drift procedure: an offline pass that records paired full-precision and quantized outputs along a sampling trajectory and fits V_{σ_i} per step, and an online pass that applies Eq. 14.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and samplers. Following SVDQuant (Li et al., 2025), we evaluate Q-Drift on six text-to-image models spanning DiT and U-Net backbones: FLUX.1-dev (20 steps), its distilled variant FLUX.1-schnell (4 steps), PixArt-Σ (Chen et al., 2024) (20 steps) and Sana (Xie et al., 2024) (20 steps) for DiT, and SDXL (Podell et al., 2024) (30 steps) with its distilled variant SDXL-Turbo (Sauer et al., 2024) (4 steps) for U-Net. Each model keeps its default inference configuration and its standard sampler: flow matching for the two FLUX models, first-order Euler for SDXL and SDXL-Turbo, and DPM-Solver++(2M, midpoint) for PixArt-Σ and Sana, following their official pipelines (Lu et al., 2022b).

Datasets. Following the SVDQuant protocol (Li et al., 2025), we sample 5K prompts for evaluation and a disjoint set of 1K prompts for calibration from MJHQ-30K (PlaygroundAI, 2024). The two subsets use different random seeds, and all methods are evaluated on the same evaluation subset for fair comparison.

Baselines and compared variants. Our quantized baselines are SVDQuant (Li et al., 2025) and MixDQ (Zhao et al., 2024), representative recent PTQ methods for text-to-image diffusion with publicly available implementations. For each model we report FP (full precision), Quantized (the quantized model named in the model column), and Q-Drift (ours, applied on top of that same quantized model at sampling time). Q-Drift involves no fine-tuning and does not modify network weights; it only corrects the sampler dynamics.

Metrics. We measure both image fidelity and prompt alignment. For fidelity, we report FID (lower is better) (Heusel et al., 2017), computed on images generated from the MJHQ prompt set. For prompt alignment, we report CLIP Score (higher is better) (Hessel et al., 2021; Radford et al., 2021). We additionally report PSNR, LPIPS, and SSIM for similarity between each method's outputs and the FP outputs.

5.2 MAIN RESULTS

Table 1 summarizes results across model families and quantization settings. Q-Drift improves FID over the corresponding quantized baseline in all seven settings of Table 1, across the evaluated diffusion backbones, PTQ methods, and samplers, and the effect is clearest when read as recovery from quantization-induced degradation. For SDXL-Turbo under SVDQuant W3A4 it reduces FID from 29.17 to 27.64 against a full-precision reference of 24.77; on PixArt-Σ, where aggressive quantization degrades quality the most, it gives the largest absolute gain, from 48.85 to 44.06; and even in a comparatively robust setting such as Sana it improves FID from 17.32 to 16.54, recovering roughly 58% of the gap to full precision. Additional results under milder quantization settings are provided in Appendix D.1. CLIP scores remain essentially unchanged, suggesting that Q-Drift preserves prompt alignment while improving distributional fidelity, and since it applies only a lightweight step-wise drift rescaling, it incurs negligible inference overhead. As reported in Appendix E, Q-Drift also lowers KID in all seven settings of Table 1. Paired 95% bootstrap intervals for $\Delta\mathrm{FID} = \mathrm{FID}(\text{Q-Drift}) - \mathrm{FID}(\text{Quantized})$ lie entirely below zero in six settings; the exception is FLUX.1-dev, whose interval includes zero. Q-Drift is designed with marginal preservation as its objective, and we evaluate its effect on distributional fidelity using FID and KID. Similarity to individual FP outputs, measured by PSNR, LPIPS, and SSIM, is not the design objective.

Table 1: Main results on MJHQ-30K under aggressive quantization settings. SVDQuant rows use W3A4 quantization, and MixDQ row uses W4A8 quantization. Boldface in the FID column indicates the lower value between the quantized baseline and Q-Drift for each setting.

Model	Method	FID↓	CLIP↑	PSNR↑	LPIPS↓	SSIM↑
FLUX.1-dev	FP	20.68	25.80	–	–	–
(SVDQuant W3A4)	Quantized	24.14	24.66	16.13	0.491	0.621
	Q-Drift	**24.06**	24.69	16.15	0.488	0.623
FLUX.1-schnell	FP	19.18	26.55	–	–	–
(SVDQuant W3A4)	Quantized	23.10	25.43	13.75	0.529	0.519
	Q-Drift	**22.13**	25.57	13.75	0.520	0.528
SDXL	FP	17.20	27.40	–	–	–
(SVDQuant W3A4)	Quantized	31.73	26.39	14.88	0.549	0.552
	Q-Drift	**30.69**	26.40	14.88	0.547	0.551
SDXL-Turbo	FP	24.77	26.39	–	–	–
(SVDQuant W3A4)	Quantized	29.17	26.38	14.30	0.455	0.456
	Q-Drift	**27.64**	26.38	14.35	0.457	0.457
SDXL-Turbo	FP	24.77	26.39	–	–	–
(MixDQ W4A8)	Quantized	28.37	25.90	15.68	0.402	0.516
	Q-Drift	**27.38**	25.87	12.62	0.587	0.356
PixArt-Sigma	FP	16.52	26.71	–	–	–
(SVDQuant W3A4)	Quantized	48.85	24.41	13.53	0.547	0.503
	Q-Drift	**44.06**	24.56	13.61	0.544	0.504
Sana	FP	15.98	27.30	–	–	–
(SVDQuant W3A4)	Quantized	17.32	27.19	16.71	0.284	0.616
	Q-Drift	**16.54**	27.20	16.69	0.286	0.612

Table 2: **Comparison with prior sampler-side corrections** on SDXL (SVDQuant W3A4, Euler).

Sampler-side method	FID↓	CLIP↑
Quantized baseline	31.73	26.39
PTQD	31.51	26.35
QNCD	32.01	26.37
D^2-DPM, deterministic	74.28	25.65
D^2-DPM, stochastic	82.87	25.32
Q-Drift	**30.69**	26.40

5.3 Qualitative Visual Comparisons

Fig. 1 compares SDXL (SVDQuant W3A4) outputs under identical prompts and initial noise. Q-Drift yields subtle but consistent improvements relative to the quantized baseline: it sharpens local detail in the first three rows and resolves an anatomically implausible leg overlap in the fourth. In the fifth row it produces a structure that differs noticeably from both FP16 and the quantized sample while preserving prompt alignment and perceptual quality, a reminder that the correction reshapes the trajectory rather than tracking the FP sample.

5.4 Comparison with Prior Sampler-Side Correction Methods

Empirical comparison. Table 2 compares prior sampler-side corrections with Q-Drift on SDXL (SVDQuant W3A4, Euler), adapting each method to this sampler as faithfully as we can. We found no public implementation of TAC-Diffusion, and for QNCD we omit the intra-step feature smoothing, since SVDQuant already applies it. Q-Drift gives the lowest FID, at a CLIP score comparable to the quantized baseline, while every prior correction lands above the quantized baseline it was meant

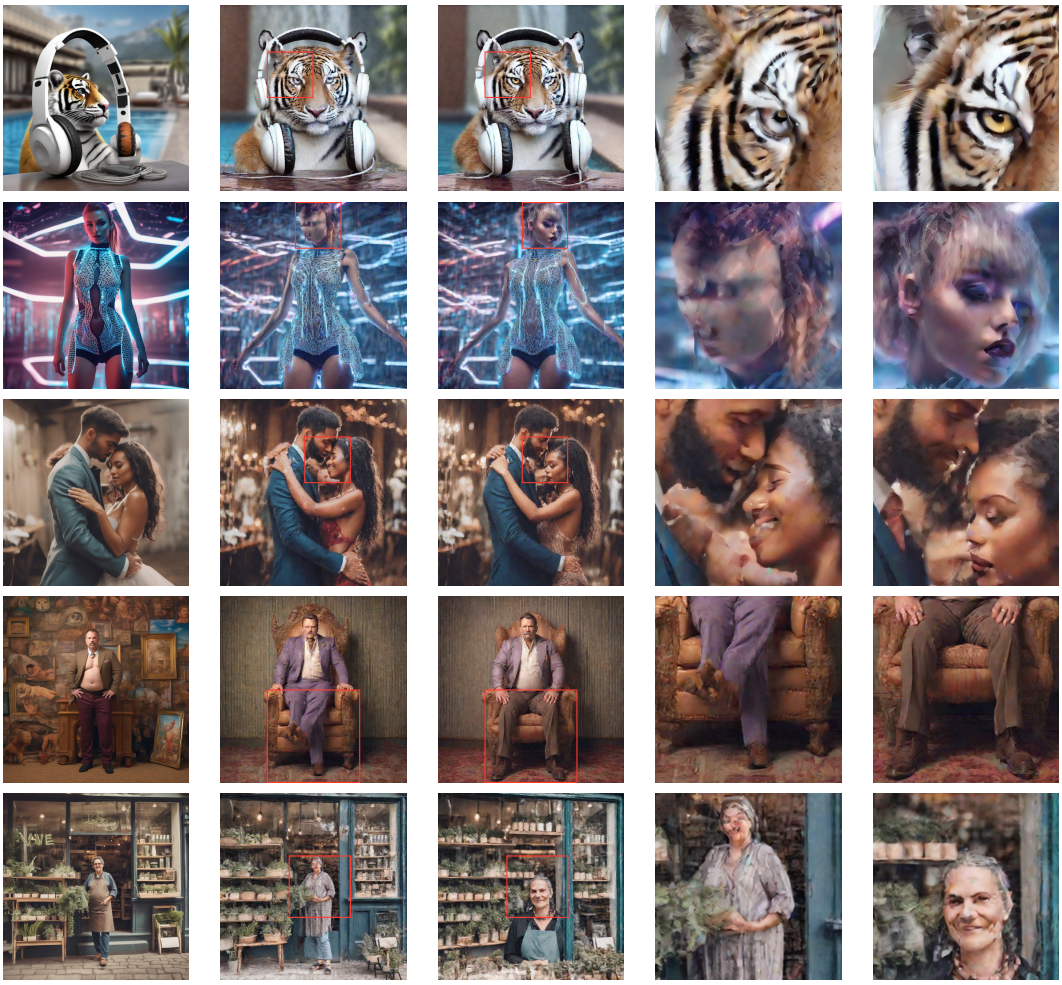


Figure 1: **Visual comparisons on SDXL (SVDQuant W3A4).** For readability, prompt texts are provided in Appendix F. Red boxes indicate the regions shown in the zoomed panels.

to improve, except PTQD, which is close to it. We next examine the correction mechanisms and their behavior in this setting.

Prediction variance under conditional-mean correction. Prior corrections aim to predict and remove the quantization error through different mechanisms (Table 4, Appendix B). Under squared error, the optimal correction based on the quantized output is the conditional mean $\mathbb{E}[\Delta\epsilon \mid \hat{\epsilon}]$. Let $\tilde{\epsilon} = \hat{\epsilon} - \mathbb{E}[\Delta\epsilon \mid \hat{\epsilon}] = \mathbb{E}[\epsilon_\theta \mid \hat{\epsilon}]$. The law of total variance gives

$$\mathrm{Var}(\epsilon_\theta) = \mathrm{Var}(\tilde{\epsilon}) + \mathbb{E}\big[\mathrm{Var}(\Delta\epsilon \mid \hat{\epsilon})\big], \tag{16}$$

Thus, conditional-mean correction reduces prediction variance relative to the full-precision output.

D^2-DPM addresses residual variability in two ways. Its stochastic variant subtracts a random sample from $\mathcal{N}(\mu_{\text{cond}}, \Sigma_{\text{cond}})$ rather than μ_{cond} itself, restoring prediction variance under the conditional Gaussian model. Its deterministic variant subtracts μ_{cond} and adjusts the sampler's diffusion term, as PTQD's variance schedule calibration does for the injected noise. Such diffusion adjustments are unavailable in the deterministic Euler, flow-matching, and DPM-Solver++ samplers used here, whereas stochastic error subtraction introduces randomness into the update.

D^2-DPM yields 74.28 FID in its deterministic variant and 82.87 in its stochastic variant. In the stochastic variant, the injected noise amounts to 10.5%–54.0% of the update magnitude across steps.

Table 3: **Calibration size and design ablations** on SDXL (SVDQuant W3A4).

Calibration / variant	FID↓	CLIP↑
Quantized baseline (no correction)	31.73	26.39
Q-Drift, 1K-prompt reference	30.69	26.40
$K = 10$, max $\sum_i \lvert\Delta c_i\rvert$	30.85	26.36
$K = 10$, min $\sum_i \Delta c_i$	30.67	26.36
$K = 10$, max $\sum_i \Delta c_i$	30.52	26.37
scalar unconditional variance	31.95	26.35
channel-wise conditional variance	30.82	26.36

Neither variant improves FID in this setting. PTQD remains close to the baseline. Its scaling mechanism is switched off here by its own rule that resets a negative coefficient to zero, so all that remains is a fixed per-channel unconditional mean, and subtracting a constant shifts the prediction without shrinking its variance.

Calibration-based error predictors must also generalize across conditioning inputs. Their effectiveness depends on the calibration distribution and the information used to predict the error. TAC-Diffusion (Yao et al., 2024) specifically reports that prompt-conditioned quantization noise is too diverse to cover with a calibration set of practical size, and responds by correcting only the unconditional branch of the prediction.

Q-Drift instead uses residual variance to rescale the deterministic update, without subtracting a predicted error or injecting noise.

5.5 Calibration Efficiency and Ablation Studies

Calibration Efficiency. Q-Drift calibrates only the per-step scalar factors c_i, whereas prior methods fit sample-level error corrections. In our comparison, PTQD and D^2-DPM use $1{,}000$ paired runs; TAC-Diffusion reports using 256 samples for text-to-image generation. We test smaller calibration sets on SDXL (SVDQuant W3A4), re-estimating $\{c_i\}$ from each of 200 subsets of $K = 10$ paired runs drawn from the same 1K-pair calibration pool. Taking the 1K-prompt factors $\{c_i^{(\mathrm{ref})}\}$ as reference, we define $\Delta c_i \triangleq c_i - c_i^{(\mathrm{ref})}$ and select three stress cases: the subset maximizing $\sum_i |\Delta c_i|$ and those maximizing and minimizing $\sum_i \Delta c_i$. The $K = 10$ rows in Table 3 report these stress cases rather than typical subsets. All three yield FID within 0.17 of the reference, with similar CLIP scores, supporting the use of $K = 10$ in this setting.

Ablation Study. Both design ablations use the same 1K-prompt calibration set as the scalar conditional-variance reference. Unconditional variance includes variability of the conditional mean, increasing the correction scale. Under the same calibration and evaluation setup, it yields 31.95 FID versus 30.69 with conditional variance (Table 3), supporting the conditional statistic here. The channel-wise variant estimates conditional variance separately for each latent channel and applies the corresponding correction channel-wise, yielding 30.82 FID versus 30.69 for the scalar form. The finer granularity provides no FID gain here, supporting one factor per step.

6 Conclusion

We presented Q-Drift, a quantization-aware drift correction that revisits quantized diffusion sampling from a distributional perspective. Across the models, samplers, and PTQ methods, Q-Drift improves FID at comparable CLIP scores, with negligible inference overhead. Our SDXL study further demonstrates effective calibration with as few as 10 paired runs. Relaxing its calibration-friendly assumptions, such as the simplified covariance structure, is left to future work.

AI USE STATEMENT

We used generative AI tools to assist with implementing our methods, help formulate mathematical claims, assist in writing proofs, and provide feedback on research methodology and experimental design. Specifically, we used code assistance tools to implement our methods. For mathematical claims, we first developed the proofs ourselves, then used generative AI tools to refine them and improve their presentation. We have not used generative AI tools to develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Generating synthetic datasets is not applicable to this work. Additionally, we used generative AI tools to summarize existing literature, draft parts of the paper, edit the manuscript to improve readability. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-
$sigma$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. URL https://arxiv.org/abs/2403.04692.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Huanpeng Chu, Wei Wu, Chengjie Zang, and Kun Yuan. Qncd: Quantization noise correction for diffusion models. In *ACM International Conference on Multimedia (ACM MM)*, 2024.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhenyuan Dong and Sai Qian Zhang. Ditas: Quantizing diffusion transformers via enhanced activation smoothing. *arXiv preprint arXiv:2409.07756*, 2024. URL https://arxiv.org/abs/2409.07756.

Hang Guo, Yawei Li, Tao Dai, Shu-Tao Xia, and Luca Benini. Intlora: Integral low-rank adaptation of quantized diffusion models. *arXiv preprint arXiv:2410.21759*, 2024. URL https://arxiv.org/abs/2410.21759.

Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986. doi: 10.1214/aop/1176992362.

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*, 2023a. URL https://arxiv.org/abs/2310.03270.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. In *Advances in Neural Information Processing Systems*, 2023b.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning, 2021. URL https://arxiv.org/abs/2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yushi Huang, Ruihao Gong, Kaidi Wu, Ding Liang, Lingxi Wang, and Chao Lin. Tfmq-dm: Temporal feature maintenance quantization for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer, 1992.

Donghyeon Lee et al. Dmq: Dissecting outliers of diffusion models for post-training quantization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Yangming Li and Mihaela van der Schaar. On error propagation of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Yanjing Li, Sheng Xu, Xianbin Cao, Xiao Sun, and Baochang Zhang. Q-dm: An efficient low-bit quantized diffusion model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yiming Li, Tong Zhang, and Jianfei Chen. Q-diffusion: Quantizing diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023b.

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. *arXiv preprint arXiv:2102.05426*, 2021. URL https://arxiv.org/abs/2102.05426.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022.

Wenxuan Liu and Sai Qian Zhang. Hq-dit: Efficient diffusion transformer with fp4 hybrid quantization. *arXiv preprint arXiv:2405.19751*, 2024. URL https://arxiv.org/abs/2405.19751.

Xuewen Liu, Zhikai Li, Junrui Xiao, and Qingyi Gu. Eda-dm: Enhanced distribution alignment for post-training quantization of diffusion models. *arXiv preprint arXiv:2401.04585*, 2024. URL https://arxiv.org/abs/2401.04585.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling. In *Advances in Neural Information Processing Systems*, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with phase-aware parameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. URL https://arxiv.org/abs/2310.04378. OpenReview: duBCwjb68o.

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 Learning Representations (ICLR)*, 2020.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

Nilesh Prasad Pandey, Marios Fournarakis, Chirag Patel, and Markus Nagel. Softmax bias correction for quantized generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

PlaygroundAI. MJHQ-30K dataset. `https://huggingface.co/datasets/playgroundai/MJHQ-30K`, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022.

Hyogon Ryu, Seohyun Lim, and Hyunjung Shim. Memory-efficient fine-tuning for quantized diffusion model. In *European Conference on Computer Vision (ECCV)*, 2024.

Hyogon Ryu, Nahyeon Park, and Hyunjung Shim. Dgq: Distribution-aware group quantization for text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, 2024.

Yuheng Shang, Hao Wang, Zhe Li, and Jianfei Chen. Ptq4dm: Post-training quantization for diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hui Shen, Jingxuan Zhang, Boning Xiong, Rui Hu, Shoufa Chen, Zhongwei Wan, Xin Wang, Yu Zhang, Zixuan Gong, Guangyin Bao, Chaofan Tao, Yongfeng Huang, Ye Yuan, and Mi Zhang. Efficient diffusion models: A survey. *arXiv preprint arXiv:2502.06805*, 2025. URL `https://arxiv.org/abs/2502.06805`.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Yang Song, Chenlin Meng, and Stefano Ermon. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Haojun Sun, Chen Tang, Zhi Wang, Yuan Meng, Xinzhu Ma, Wenwu Zhu, et al. Tmpq-dm: Joint timestep reduction and quantization precision selection for efficient diffusion models. *arXiv preprint arXiv:2404.09532*, 2024. URL https://arxiv.org/abs/2404.09532.

Siao Tang, Xin Wang, Hong Chen, Chaoyu Guan, Zewen Wu, Yansong Tang, and Wenwu Zhu. Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Shilong Tian, Hong Chen, Chengtao Lv, Yu Liu, Jinyang Guo, Xianglong Liu, Shengxi Li, Hao Yang, and Tao Xie. Qvd: Post-training quantization for video diffusion models. In *ACM International Conference on Multimedia (ACM MM)*, 2024.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate post-training quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Haoxuan Wang, Yuzhang Shang, Zhihang Yuan, Junyi Wu, Junchi Yan, and Yan Yan. Quest: Low-bit diffusion model quantization via efficient selective finetuning. *arXiv preprint arXiv:2402.03666*, 2024b. URL https://arxiv.org/abs/2402.03666.

Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Ptq4dit: Post-training quantization for diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. URL https://arxiv.org/abs/2410.10629.

Shuchen Xue, Mingyang Yi, Weijian Luo, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Yuewei Yang, Jialiang Wang, Xiaoliang Dai, Peizhao Zhang, and Hongbo Zhang. An analysis on quantizing diffusion transformers. *arXiv preprint arXiv:2406.11100*, 2024. URL https://arxiv.org/abs/2406.11100.

Yuzhe Yao, Feng Tian, Jun Chen, Haonan Lin, Guang Dai, Yong Liu, and Jingdong Wang. Timestep-aware correction for quantized diffusion models. In *European Conference on Computer Vision*, 2024.

Tianwei Yin et al. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Qian Zeng, Jie Song, Han Zheng, Hao Jiang, and Mingli Song. D^2-dpm: Dual denoising for quantized diffusion probabilistic models. In *AAAI Conference on Artificial Intelligence*, 2025.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations (ICLR)*, 2023.

Tian Zhao et al. Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Tianchen Zhao, Xuefei Ning, Tongcheng Fang, Enshu Liu, Guyue Huang, Zinan Lin, Shengen Yan, Guohao Dai, and Yu Wang. Mixdq: Memory-efficient few-step text-to-image diffusion models with metric-decoupled mixed precision quantization. In *European Conference on Computer Vision (ECCV)*, 2024.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

APPENDIX

A Q-DRIFT ALGORITHM

Algorithm 1 summarizes Q-Drift's offline calibration and online sampling procedure for a first-order Euler sampler.

Algorithm 1 Q-Drift: Quantization-Aware Drift Correction (Calibration + Sampling)

Input: full-precision denoiser ϵ_θ, quantized denoiser $\hat{\epsilon}_\theta$, conditioning c, schedule $\{\sigma_i\}_{i=0}^{M}$, calibration set $\mathcal{D}_{\text{cal}}$, number of samples N.
Output: $\{\mathbf{x}_{\sigma_M}^{(n)}\}_{n=1}^{N}$.
Offline calibration.
for each conditioning $c_{\text{cal}} \in \mathcal{D}_{\text{cal}}$ **do**
  Sample $\mathbf{x}_{\sigma_0} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$.
  for $i = 0$ **to** $M - 1$ **do**
    $\epsilon \leftarrow \epsilon_\theta(\mathbf{x}_{\sigma_i}, \sigma_i, c_{\text{cal}})$
    $\hat{\epsilon} \leftarrow \hat{\epsilon}_\theta(\mathbf{x}_{\sigma_i}, \sigma_i, c_{\text{cal}})$
    $\Delta\epsilon \leftarrow \hat{\epsilon} - \epsilon$
    Store paired outputs $(\hat{\epsilon}, \Delta\epsilon)$ at σ_i
    $\Delta\sigma_i \leftarrow \sigma_{i+1} - \sigma_i$
    $\mathbf{x}_{\sigma_{i+1}} \leftarrow \mathbf{x}_{\sigma_i} + \Delta\sigma_i\, \epsilon$
  end for
end for
for $i = 0$ **to** $M - 1$ **do**
  Fit the Gaussian model to stored pairs at σ_i (Sec. 3.3–3.4)
  Compute V_{σ_i} (Eq. 15)
end for
Online sampling (first-order Euler sampler).
for $n = 1$ **to** N **do**
  Sample $\mathbf{x}_{\sigma_0}^{(n)} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$.
  for $i = 0$ **to** $M - 1$ **do**
    $\Delta\sigma_i \leftarrow \sigma_{i+1} - \sigma_i$
    $\hat{\epsilon} \leftarrow \hat{\epsilon}_\theta(\mathbf{x}_{\sigma_i}^{(n)}, \sigma_i, c)$
    $c_i \leftarrow \dfrac{|\Delta\sigma_i|}{2\sigma_i}\, V_{\sigma_i}$
    $\mathbf{x}_{\sigma_{i+1}}^{(n)} \leftarrow \mathbf{x}_{\sigma_i}^{(n)} + \Delta\sigma_i\, (1 + c_i)\, \hat{\epsilon}$
  end for
end for
return $\{\mathbf{x}_{\sigma_M}^{(n)}\}_{n=1}^{N}$.

B MECHANISMS OF PRIOR SAMPLER-SIDE CORRECTIONS

Table 4 summarizes the statistics and correction mechanisms used by prior sampler-side methods and Q-Drift.

C ADDITIONAL DERIVATIONS AND EXTENSIONS

C.1 EXTENSION TO FLOW-MATCHING SAMPLER

Many flow-matching and rectified-flow samplers use a first-order update of the form

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\sigma_i\, v_\theta(\mathbf{x}_i, \sigma_i, c), \tag{17}$$

where σ denotes the sampler's interpolation coordinate and v_θ is its velocity field. The update has the same numerical form as the Euler update in Eq. 9.

Table 4: **Mechanisms of sampler-side corrections for quantized diffusion.** Methods that combine several mechanisms are listed one row per mechanism. QNCD's intra-step feature smoothing is omitted, as it acts inside the network rather than on the sampler. "stochastic" marks mechanisms that require $\sigma_t > 0$; "studied" refers to the deterministic Euler, flow-matching, and DPM-Solver++ samplers evaluated here.

Method (mechanism)	Statistic estimated	How it enters the sampler	Samplers
PTQD (CNC)	correlation coeff. k	divides the output by $1+k$	any
PTQD (BC)	channel-wise mean of $\Delta\epsilon$	subtracted from the output	any
PTQD (VSC)	residual variance σ_q^2	shrinks the injected noise σ_t	stochastic
TAC-Diff. (NER)	channel-wise coeff. K_t	rescales the output	any
TAC-Diff. (IBC)	input bias	corrects the sampler input x_t	any
QNCD (inter)	per-sample noise estimate	subtracted from the output	any
D^2-DPM (TSQNM)	conditional mean μ_{cond}	subtracted from the output	any
D^2-DPM (S-D^2)	conditional cov. Σ_{cond}	sampled, then subtracted	any
D^2-DPM (D-D^2)	conditional cov. Σ_{cond}	shrinks the diffusion term Σ_t	stochastic
Q-Drift	residual variance V_{σ_i}	rescales the deterministic update	studied

Q-Drift for Flow-matching. With quantization, the sampler uses $\hat{v}_\theta = v_\theta + \Delta v$. We adapt the Euler rescaling rule by using the conditional residual variance of Δv_i given $\hat{v}_\theta$ as V_{σ_i} in Eq. 15. With $c_i = \frac{|\Delta\sigma_i|}{2\sigma_i} V_{\sigma_i}$, the update is

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\sigma_i \,(1 + c_i)\, \hat{v}_\theta(\mathbf{x}_i, \sigma_i, c). \tag{18}$$

Scope of the extension. A general flow-matching velocity does not satisfy the score–velocity relation used in the VE derivation. We therefore treat this rescaling as an empirical extension to the flow-matching sampler.

C.2 Q-Drift for DPM-Solver++ in Log-SNR Coordinates

We adapt Q-Drift to DPM-Solver++ (Lu et al., 2022b) by propagating calibrated residual-variance statistics through the solver's model-output coefficients and rescaling the baseline update in normalized coordinates. The generalized SDE in log-SNR coordinates motivates the correction scale.

Log-SNR notation and VE reparameterization. We write the forward marginal as

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{19}$$

and define the log-SNR coordinate

$$\lambda_t \triangleq \log \frac{\alpha_t}{\sigma_t}, \qquad \frac{\sigma_t}{\alpha_t} = e^{-\lambda_t}. \tag{20}$$

In normalized coordinates $\mathbf{y}_t \triangleq \mathbf{x}_t/\alpha_t$, this becomes

$$\mathbf{y}_t = \mathbf{x}_0 + \bar{\sigma}_t \boldsymbol{\epsilon}, \qquad \bar{\sigma}_t \triangleq \frac{\sigma_t}{\alpha_t} = e^{-\lambda_t}. \tag{21}$$

The normalized forward marginals thus follow a VE-type path with unit signal scale and noise level $\bar{\sigma}$ (Song et al., 2021b). Sampling proceeds toward increasing λ.

Generalized SDEs in log-SNR coordinates. Applying Eq. 7 to this normalized marginal path gives

$$d\mathbf{y} = -\bar{\sigma}\big(1 + \tilde{\beta}_{\bar{\sigma}}\bar{\sigma}\big)\nabla_{\mathbf{y}} \log p(\mathbf{y}; \bar{\sigma})\, d\bar{\sigma} + \bar{\sigma}\sqrt{2\tilde{\beta}_{\bar{\sigma}}}\, d\mathbf{w}_{\bar{\sigma}}, \tag{22}$$

where $\mathbf{w}_{\bar{\sigma}}$ is a reverse-time Wiener process with increment covariance $|d\bar{\sigma}|\mathbf{I}$. Using $d\bar{\sigma} = -\bar{\sigma}\, d\lambda$, define $\tilde{\beta}_\lambda \triangleq \bar{\sigma}\tilde{\beta}_{\bar{\sigma}} \geq 0$ and the exact score-based quantity $\epsilon_*(\mathbf{y}, \lambda) \triangleq -\bar{\sigma}\nabla_{\mathbf{y}} \log p(\mathbf{y}; \bar{\sigma})$. Then

$$d\mathbf{y} = -(1 + \tilde{\beta}_\lambda)\bar{\sigma}\epsilon_*(\mathbf{y}, \lambda)\, d\lambda + \bar{\sigma}\sqrt{2\tilde{\beta}_\lambda}\, d\mathbf{w}_\lambda, \tag{23}$$

where $\mathbf{w}_\lambda$ is a Wiener process along increasing λ. With the exact score, these continuous-time SDEs share the prescribed marginals: the diffusion and its paired score drift vary together (Karras et al., 2022).

Probability-flow ODE in log-SNR coordinates. Setting $\tilde{\beta}_\lambda = 0$ and replacing ϵ_* with the learned prediction gives

$$d\mathbf{y} = -e^{-\lambda}\epsilon_\theta(\mathbf{y}, \lambda, c)\, d\lambda. \tag{24}$$

Here $\epsilon_\theta(\mathbf{y}, \lambda, c)$ is shorthand for the original denoiser evaluated at $\mathbf{x}_t = \alpha_t \mathbf{y}$ and the corresponding timestep t. Integrating the learned ODE from s to t gives

$$\mathbf{y}_t = \mathbf{y}_s - \int_{\lambda_s}^{\lambda_t} e^{-\lambda}\epsilon_\theta(\mathbf{y}_\lambda, \lambda, c)\, d\lambda. \tag{25}$$

This is the noise-prediction representation of the diffusion ODE. DPM-Solver++ discretizes an equivalent continuous-time data-prediction formulation (Lu et al., 2022b).

Local propagation of auxiliary residual variances. At fixed current and stored states, the direct contribution of model-output perturbations to a solver step is described by coefficients $w_{i,j}$, with $j \in \mathcal{J}_i$ indexing the predictions used in that step. As in Sec. 4.2, we introduce independent auxiliary perturbations $\boldsymbol{\eta}_j \sim \mathcal{N}(\mathbf{0}, V_j\mathbf{I})$, where V_j is the calibrated statistic in Sec. 4.3. Their propagated increment is

$$\delta\mathbf{y}_i^{\text{aux}} = \sum_{j \in \mathcal{J}_i} w_{i,j}\boldsymbol{\eta}_j, \tag{26}$$

with covariance

$$\text{Cov}(\delta\mathbf{y}_i^{\text{aux}}) = \left(\sum_{j \in \mathcal{J}_i} w_{i,j}^2 V_j\right)\mathbf{I}. \tag{27}$$

This local construction holds the state history fixed and omits dependence between successive solver increments, which can share stored predictions. For a diffusion strength $\tilde{\beta}_{\lambda,i}$ constant over the step, the diffusion term in Eq. 23 has increment covariance $2\tilde{\beta}_{\lambda,i}\int_{\lambda_{i-1}}^{\lambda_i} e^{-2\lambda}\, d\lambda\, \mathbf{I} = \tilde{\beta}_{\lambda,i}|\bar{\sigma}_{i-1}^2 - \bar{\sigma}_i^2|\mathbf{I}$. Matching this covariance to Eq. 27 defines

$$\tilde{\beta}_{\lambda,i} = \frac{\sum_{j \in \mathcal{J}_i} w_{i,j}^2 V_j}{|\bar{\sigma}_{i-1}^2 - \bar{\sigma}_i^2|}. \tag{28}$$

Here and below, α_i, σ_i, and λ_i abbreviate their values at t_i.

Correction scale and normalized update. The paired drift multiplier in Eq. 23 motivates the choice

$$c_i \triangleq \tilde{\beta}_{\lambda,i}. \tag{29}$$

We first compute the baseline quantized solver step $\mathbf{x}_i^{\text{base}}$ from the current state and stored predictions. We then rescale its full increment in normalized coordinates:

$$\begin{aligned} \mathbf{y}_i^{\text{base}} &= \mathbf{x}_i^{\text{base}}/\alpha_i, \\ \mathbf{y}_i &= \mathbf{y}_{i-1} + (1 + c_i)(\mathbf{y}_i^{\text{base}} - \mathbf{y}_{i-1}), \qquad \mathbf{x}_i = \alpha_i\mathbf{y}_i. \end{aligned} \tag{30}$$

This adapts the drift-rescaling principle to the discrete solver while retaining its baseline step computation.

Instantiation: DPM-Solver++(2M, midpoint). For noise prediction, DPM-Solver++ converts the quantized output to a data prediction:

$$\hat{\mathbf{x}}_{\theta,i} = \frac{\mathbf{x}_i - \sigma_i \hat{\epsilon}_{\theta,i}}{\alpha_i} = \mathbf{y}_i - \bar{\sigma}_i \hat{\epsilon}_{\theta,i}. \tag{31}$$

Let $h_i \triangleq \lambda_i - \lambda_{i-1}$ and $r_i \triangleq h_{i-1}/h_i$. The second-order multistep midpoint update (Algorithm 2 of Lu et al. (2022b)) is

$$\begin{aligned} \mathbf{x}_i^{\text{base}} &= \frac{\sigma_i}{\sigma_{i-1}} \mathbf{x}_{i-1} - \alpha_i (e^{-h_i} - 1) \hat{\mathbf{D}}_i, \\ \hat{\mathbf{D}}_i &= \left(1 + \frac{1}{2r_i}\right) \hat{\mathbf{x}}_{\theta,i-1} - \frac{1}{2r_i} \hat{\mathbf{x}}_{\theta,i-2}. \end{aligned} \tag{32}$$

At fixed states, a perturbation $\Delta\epsilon_k$ changes the data prediction by $-\bar{\sigma}_k \Delta\epsilon_k$. Its direct contribution to the normalized baseline step is therefore

$$\delta \mathbf{y}_i^{\text{local}} = (e^{-h_i} - 1) \left[\left(1 + \frac{1}{2r_i}\right) \bar{\sigma}_{i-1} \Delta\epsilon_{i-1} - \frac{1}{2r_i} \bar{\sigma}_{i-2} \Delta\epsilon_{i-2} \right]. \tag{33}$$

The model-output perturbation coefficients used in Eq. 27 are thus

$$w_{i,i-1} = (e^{-h_i} - 1) \left(1 + \frac{1}{2r_i}\right) \bar{\sigma}_{i-1}, \qquad w_{i,i-2} = -(e^{-h_i} - 1) \frac{\bar{\sigma}_{i-2}}{2r_i}. \tag{34}$$

The resulting correction scale is

$$c_i = \tilde{\beta}_{\lambda,i} = \frac{(e^{-h_i} - 1)^2 \left[\left(1 + \frac{1}{2r_i}\right)^2 \bar{\sigma}_{i-1}^2 V_{i-1} + \left(\frac{1}{2r_i}\right)^2 \bar{\sigma}_{i-2}^2 V_{i-2} \right]}{|\bar{\sigma}_{i-1}^2 - \bar{\sigma}_i^2|}. \tag{35}$$

For first-order steps, including warmup, the same construction uses the single coefficient $w_{i,i-1} = \bar{\sigma}_i - \bar{\sigma}_{i-1}$. Finally, Eq. 30 can be expressed in the original coordinates as

$$\mathbf{x}_i = (1 + c_i) \mathbf{x}_i^{\text{base}} - c_i \frac{\alpha_i}{\alpha_{i-1}} \mathbf{x}_{i-1}, \tag{36}$$

where $\mathbf{x}_i^{\text{base}}$ is given by Eq. 32 for a second-order step.

D Additional Experiments

Unless otherwise noted, all supplementary experiments follow the same setup as in the main experiments.

D.1 Additional Results (Milder Quantization Settings)

Table 5 reports additional results under milder quantization settings for SVDQuant and includes an additional PTQ method, ViDiT-Q (Zhao et al., 2025). Under these milder settings, the quantized baseline can already match or outperform full precision in FID, leaving less headroom for improvement. Accordingly, the gains from Q-Drift are mixed across settings.

D.2 Empirical Examination of the Modeling Assumptions

We examine selected distributional and covariance properties underlying the modeling assumptions in Sec. 3.3–Sec. 3.4 on SDXL with SVDQuant W3A4, using the 5K paired full-precision and quantized calibration set collected for SDXL (SVDQuant W3A4). We examine three representative timesteps, chosen as the nearest discrete indices to $0.9T$, $0.5T$, and $0.1T$.

Marginal and joint Gaussianity. This experiment directly probes the joint Gaussian modeling assumption introduced in Sec. 3.3. Fig. 2 shows that the marginal distributions of $\hat{\epsilon}_\theta^{(t)}$ and $\Delta\epsilon_\theta^{(t)}$ at the selected coordinate are well-approximated by fitted Gaussians across timesteps. The corresponding joint densities are approximately elliptical, with the fitted Gaussian ellipse capturing the dominant orientation and spread of the samples. These observations support the timestep-wise Gaussian approximation at the examined coordinate.

Table 5: Additional results on MJHQ-30K (milder quantization settings). Rows labeled W4A4 use W4A4 quantization for SVDQuant; ViDiT-Q rows use W8A8 quantization.

Model	Method	FID↓	CLIP↑	PSNR↑	LPIPS↓	SSIM↑
FLUX.1-dev	FP	20.68	25.80	–	–	–
(SVDQuant W4A4)	Quantized	**20.49**	25.68	22.95	0.194	0.830
	Q-Drift	20.55	25.68	22.98	0.193	0.830
FLUX.1-schnell	FP	19.18	26.55	–	–	–
(SVDQuant W4A4)	Quantized	**18.62**	26.41	18.44	0.251	0.728
	Q-Drift	18.69	26.44	18.42	0.251	0.727
SDXL	FP	17.20	27.40	–	–	–
(SVDQuant W4A4)	Quantized	**17.05**	27.27	20.41	0.270	0.728
	Q-Drift	17.17	27.29	20.41	0.270	0.727
SDXL-Turbo	FP	24.77	26.39	–	–	–
(SVDQuant W4A4)	Quantized	24.75	26.37	18.44	0.242	0.656
	Q-Drift	**24.51**	26.42	12.85	0.549	0.380
PixArt-Sigma	FP	16.52	26.71	–	–	–
(SVDQuant W4A4)	Quantized	**19.61**	26.37	16.75	0.360	0.629
	Q-Drift	19.65	26.36	16.71	0.368	0.619
PixArt-Sigma	FP	16.52	26.71	–	–	–
(ViDiT-Q W8A8)	Quantized	15.67	26.65	21.11	0.176	0.774
	Q-Drift	**15.62**	26.67	21.08	0.176	0.775
Sana	FP	15.98	27.30	–	–	–
(SVDQuant W4A4)	Quantized	15.74	27.23	17.27	0.260	0.637
	Q-Drift	**15.68**	27.24	17.16	0.270	0.626

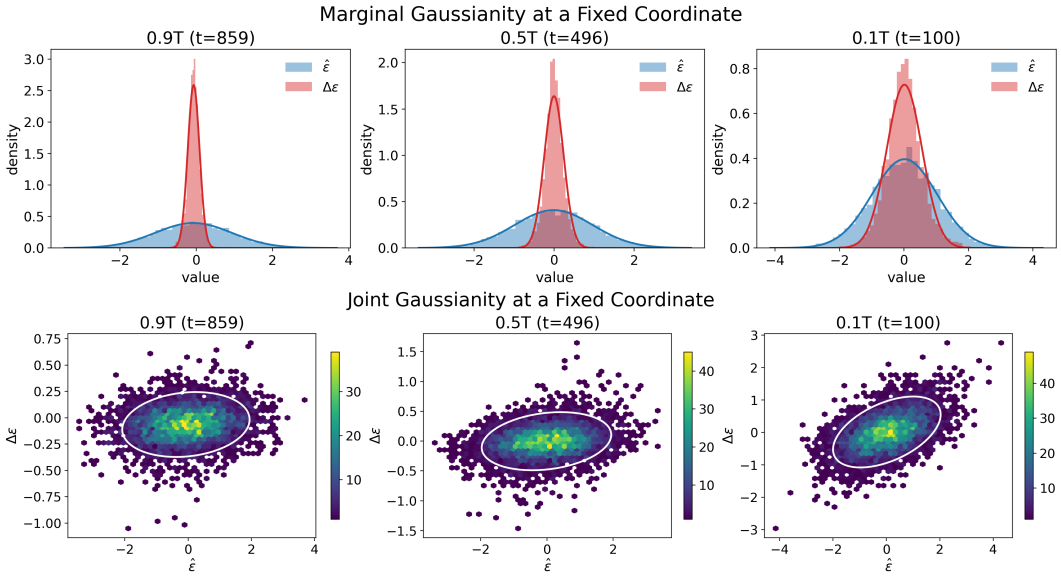


Figure 2: Empirical validation of marginal and joint Gaussianity on SDXL (SVDQuant W3A4). Top: Histograms of $\hat{\epsilon}_\theta^{(t)}$ and $\Delta\epsilon_\theta^{(t)}$ at a latent coordinate, overlaid with fitted Gaussian curves. Bottom: The corresponding 2D joint densities with fitted Gaussian ellipses. All 1D and 2D visualizations are shown at the same latent coordinate, $(c, h, w) = (0, 64, 64)$.

Diagonal covariance as a tractable approximation. The more restrictive assumption in Sec. 3.4 is the diagonal approximation to the covariance blocks $\Sigma_{\hat{\epsilon}\hat{\epsilon}}^{(t)}$, $\Sigma_{\Delta\Delta}^{(t)}$, and $\Sigma_{\hat{\epsilon}\Delta}^{(t)}$. This simplification is

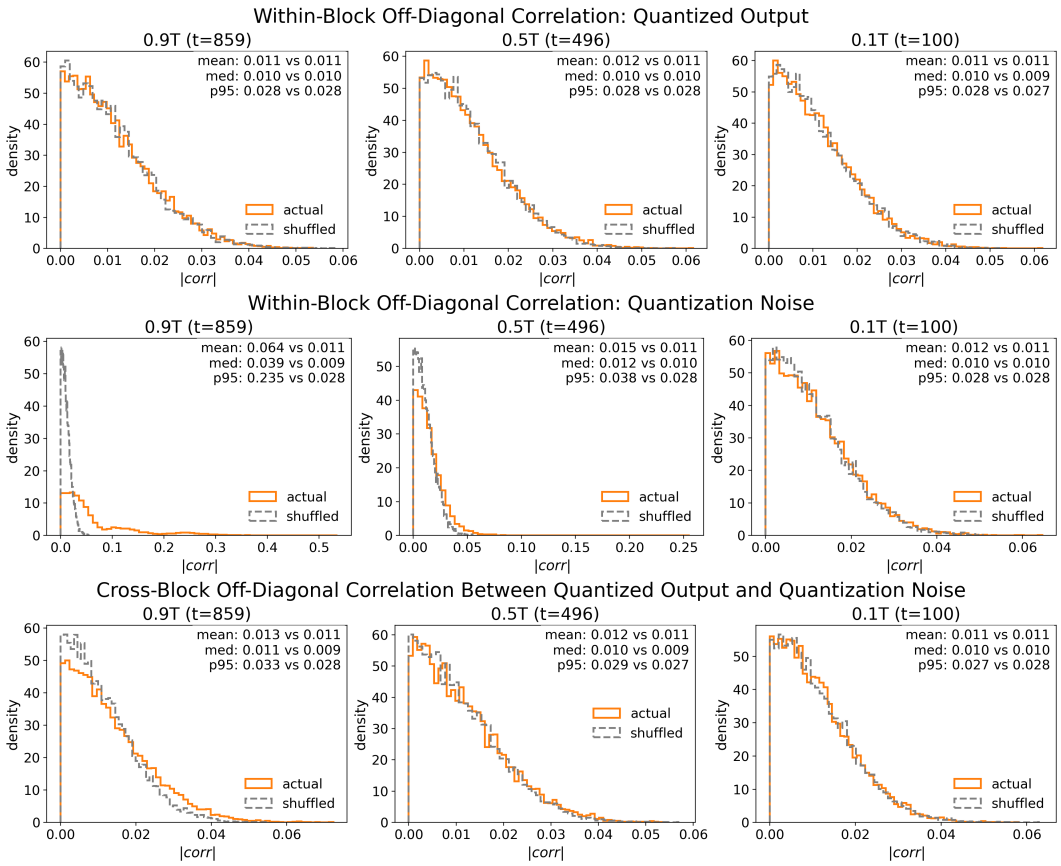


Figure 3: Empirical validation of the diagonal-covariance simplification. For each selected timestep, we compare the distribution of absolute correlations from 10,000 random off-diagonal pairs, with each correlation estimated over 5,000 calibration samples, against a shuffled baseline. The shuffled baseline is obtained by randomly permuting a variable in each pair. In each panel, the inset reports the mean, median (med), and 95th percentile (p95) of the absolute-correlation distribution, listed as actual vs shuffled. Top: off-diagonal entries of $\Sigma^{(t)}_{\hat{\epsilon}\hat{\epsilon}}$. Middle: off-diagonal entries of $\Sigma^{(t)}_{\Delta\Delta}$. Bottom: off-diagonal cross-block entries of $\Sigma^{(t)}_{\hat{\epsilon}\Delta}$ for $i \neq j$.

practically necessary since with only 5K calibration samples, directly estimating a covariance block over a latent tensor of size $C \times H \times W = 4 \times 128 \times 128 = 65{,}536$ would be statistically unstable. In this regime, the ambient dimensionality already exceeds the number of available samples, so the resulting sample covariance would be severely underdetermined and its off-diagonal structure would be poorly estimated.

To assess whether the off-diagonal entries are negligible in practice, we compare the empirical distribution of absolute correlations from random off-diagonal pairs against a shuffled baseline. The shuffled baseline is obtained by randomly permuting a variable in each pair, which preserves the marginal distribution of each element while removing pairwise dependence. This comparison is necessary because finite-sample estimation produces nonzero empirical correlations even under independence. Accordingly, the shuffled baseline estimates the finite-sample noise floor associated with the 5K calibration set. Agreement with this baseline indicates little resolved pairwise dependence among the sampled entries at the available sample size. Fig. 3 reports three complementary checks: off-diagonal entries of $\Sigma^{(t)}_{\hat{\epsilon}\hat{\epsilon}}$, off-diagonal entries of $\Sigma^{(t)}_{\Delta\Delta}$, and off-diagonal cross-block entries of $\Sigma^{(t)}_{\hat{\epsilon}\Delta}$ for $i \neq j$.

Two observations are important. First, the off-diagonal entries of $\Sigma^{(t)}_{\hat{\epsilon}\hat{\epsilon}}$ and the off-diagonal cross-block entries of $\Sigma^{(t)}_{\hat{\epsilon}\Delta}$ are generally close to the shuffled baseline. Second, the off-diagonal entries

Table 6: **Distributional metrics and statistical uncertainty** on MJHQ-30K, computed on the same generated image sets used in Table 1. KID is in $\times 10^3$ scale and brackets denote 95% paired bootstrap intervals. A Δ interval lying entirely below zero indicates an improvement supported by the sample.

Model	Method	FID↓	ΔFID 95% CI	KID↓
FLUX.1-dev	FP	20.68	–	3.44
(SVDQuant W3A4)	Quantized	24.14	–	5.67
	Q-Drift	24.06	[−0.30, 0.09]	5.59
FLUX.1-schnell	FP	19.18	–	2.66
(SVDQuant W3A4)	Quantized	23.10	–	5.03
	Q-Drift	22.13	[−1.18, −0.84]	4.52
SDXL	FP	17.20	–	1.49
(SVDQuant W3A4)	Quantized	31.73	–	9.36
	Q-Drift	30.69	[−1.38, −0.70]	8.65
SDXL-Turbo	FP	24.77	–	6.76
(SVDQuant W3A4)	Quantized	29.17	–	8.75
	Q-Drift	27.64	[−1.83, −1.24]	7.80
SDXL-Turbo	FP	24.77	–	6.76
(MixDQ W4A8)	Quantized	28.37	–	8.32
	Q-Drift	27.38	[−1.49, −0.50]	7.86
PixArt-Σ	FP	16.52	–	1.86
(SVDQuant W3A4)	Quantized	48.85	–	17.47
	Q-Drift	44.06	[−5.24, −4.26]	14.47
Sana	FP	15.98	–	1.51
(SVDQuant W3A4)	Quantized	17.32	–	2.07
	Q-Drift	16.54	[−0.96, −0.57]	1.71

of $\Sigma^{(t)}_{\Delta\Delta}$ exhibit non-negligible dependence at the earliest, highest-noise timestep, while the gap to the shuffled baseline becomes much smaller at later timesteps. Therefore, the diagonal covariance model should be viewed as a calibration-friendly approximation rather than an exact description.

E Distributional Metrics and Statistical Uncertainty

FID is a point estimate computed on a finite sample, so a raw FID gap does not by itself show that one sampler produces a closer distribution than another. Table 6 therefore reports two further quantities alongside FID.

The first is KID, an unbiased estimator of squared polynomial-kernel MMD (Bińkowski et al., 2018), which provides an additional measure of distributional fidelity. Its expectation is independent of sample size, although its sampling variability depends on it. We compute KID from 100 random subsets of $1{,}000$ images and report it in $\times 10^3$ scale.

The second is a paired 95% percentile bootstrap interval for $\Delta\mathrm{FID} = \mathrm{FID}(\text{Q-Drift}) - \mathrm{FID}(\text{Quantized})$. We use $1{,}000$ resamples, sharing the sampled indices of aligned generated images between the two methods and using a common resampled reference set within each replicate. The interval endpoints are the 2.5th and 97.5th percentiles of the bootstrap differences. Pairing retains the dependence between the two FID estimates, but does not guarantee cancellation of their finite-sample biases.

Q-Drift lowers KID in every setting. Six of the seven $\Delta\mathrm{FID}$ intervals lie entirely below zero; the exception is FLUX.1-dev, whose FID gain is the smallest of all settings and is not resolved at this sample size. No interval in the table lies above zero.

F PROMPTS FOR THE QUALITATIVE VISUAL COMPARISONS

The prompts corresponding to the five rows in Fig. 1 are listed below in row order.

Row 1: "pingüino con auriculares puestos subiendo las aletas en la piscina del hotel monge tibetano poniendole unos auriculares a un tigre blanco ultra realistic 3d photo v5."

Row 2: "Futuristic Lace dress, Fashion Concept, led lights. Sexy attractive pose. Hotpants. Beautiful, cinematic lighting, 8K, sharp details. Crystal clear feel, High quality, Fineart print, HQ. Tron vibes. Futuristic background. colourful. Hypermaximalist, hyperdetailed, awardwinning fashion photography, professional colour grading, clean sharp focus. Clean details."

Row 3: "a moment in time of a stunning, cute couple, grinding on each other."

Row 4: "wellbuilt, beefy, stout, heavyset, mature man, wide shot, Realistic editorial photography session featuring a stocky very handsome in his fourties with light brown hair and fair skin, handsome, attractive, interesting background, art by Jean Baptiste Monge, Lisa frank, Boris Vallejo, shot photography by Wes Anderson, Nikon Photography Symmetrical Frontal Portrait, High Definition, 8K, 200mm."

Row 5: "an image of a shop owner for a sustainable business."