Q-SCHED: PUSHING THE BOUNDARIES OF FEW-STEP DIFFUSION MODELS WITH QUANTIZATION-AWARE SCHEDULING

Anonymous authors
Paper under double-blind review

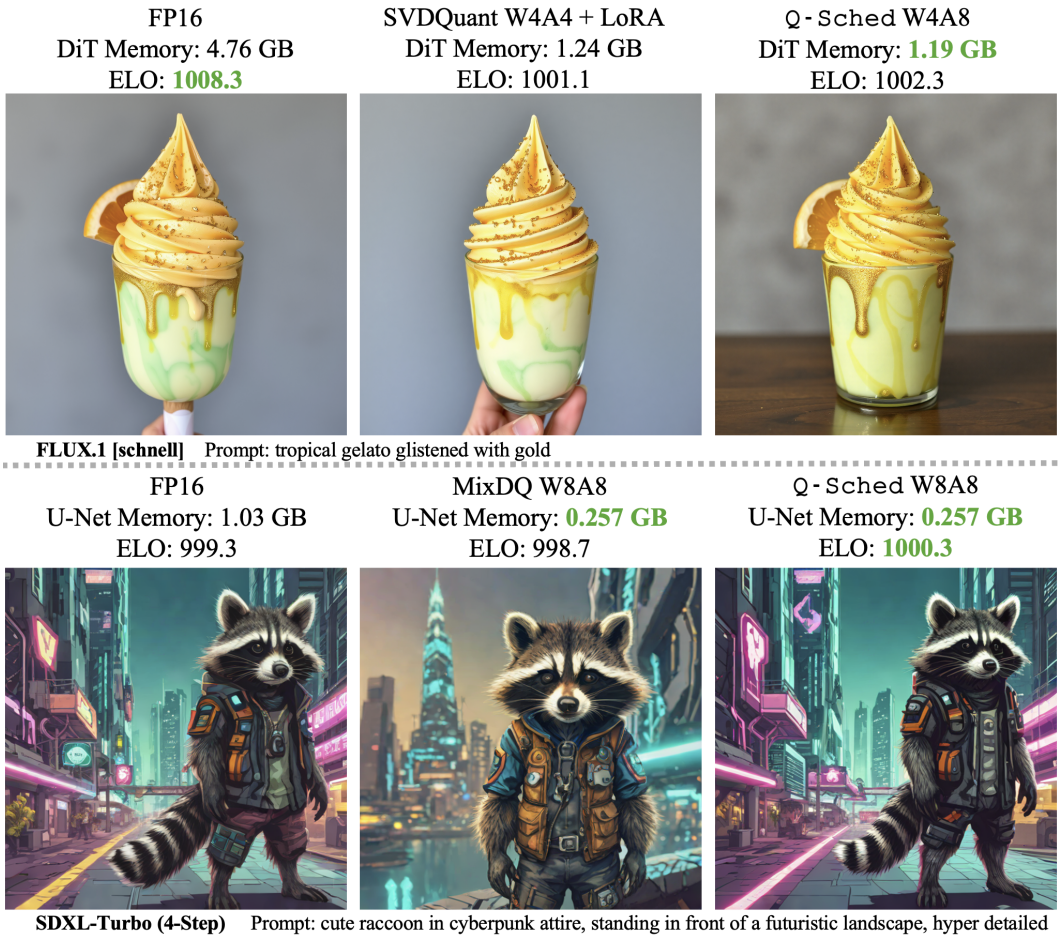


Figure 1: When large diffusion models are reduced to W8A8 or W4A8 for deployment, image fidelity drops. Q-Sched applies scheduler-level tuning, just two coefficients per step, to steer the sampler back to FP16-like quality, with no new checkpoints, no finetuning, and no extra FLOPs.

ABSTRACT

Text-to-image diffusion models remain computationally intensive: generating a single image typically requires dozens of passes through large transformer backbones (*e.g.* , SDXL uses ~ 50 evaluations of a 2.6B-parameter model). Few-step variants reduce the step count to 2–8, but still rely on large, full-precision U-Net/DiT backbones, making inference impractical on resource-constrained platforms, both on-device (latency/energy) and in data centers with multi-instance GPU (MIG) style GPU partitioning (limited memory/throughput per slice). Existing post-training quantization (PTQ) methods are further hampered by dependence on full-precision calibration.

We introduce Q-Sched, a scheduler-level PTQ approach that adapts the diffusion sampler rather than the model weights. By adjusting the few-step sampling trajectory with quantization-aware preconditioning coefficients, Q-Sched matches or surpasses full-precision quality while delivering a $4\times$ reduction in model size and preserving a single reusable checkpoint across bit-widths. To learn these coefficients, we propose a reference-free Joint Alignment–Quality (JAQ) loss, which combines text–image compatibility with an image-quality objective for fine-grained control; JAQ requires only a handful of calibration prompts and avoids any full-precision inference during calibration.

Empirically, Q-Sched yields substantial gains: a **15.5%** FID improvement over the FP16 4-step Latent Consistency Model and a **16.6%** improvement over the FP16 8-step Phased Consistency Model, demonstrating that quantization and few-step distillation are complementary for high-fidelity generation. A large-scale user study with $\mathbf{80{,}000}+$ annotations further validates these results on both FLUX.1[schnell] and SDXL-Turbo. Code will be released.

1 INTRODUCTION

Diffusion models have achieved state-of-the-art generative quality across vision (Amit et al., 2021; Baranchuk et al., 2021; Brempong et al., 2022; Ho et al., 2022; Meng et al., 2021; Yang et al., 2022a), language (Austin et al., 2021; Li et al., 2022b), multimodal modeling (Avrahami et al., 2022; Ramesh et al., 2022), and scientific domains (Anand & Achim, 2022; Cao et al., 2022). Yet systems such as Stable Diffusion XL (Podell et al., 2023; Meng et al., 2021) and CogVideoX (Yang et al., 2024) remain costly at inference time: denoising typically requires tens to hundreds of steps, each invoking a large U-Net or Diffusion transformer (DiT) (Peebles & Xie, 2023).

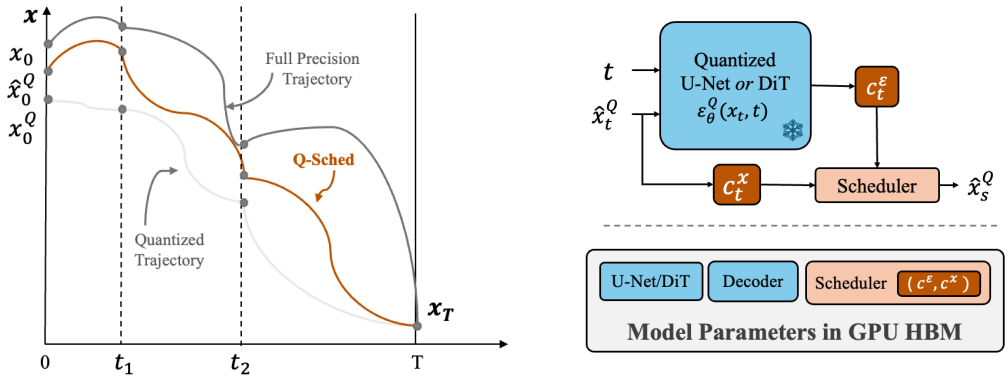
Practical deployment therefore hinges on two levers: (1) reducing the number of function evaluations (few-step sampling), and (2) lowering the cost per evaluation (compression via quantization (He et al., 2024; Guo et al., 2022), pruning (Fang et al., 2024), or distillation (Huang et al., 2024)). These levers are particularly important in two widely used settings. *On-device*, memory and compute budgets are tight, latency and energy constraints are strict, and privacy/offline use cases preclude server offloading (Zhao et al., 2024b). *In data centers with MIG partitioning*, a single GPU is sliced into multiple smaller instances to increase concurrency and predictability; each slice has limited memory/throughput, making model footprint and per-step cost decisive (Zhang et al., 2023; Li et al., 2022a). In both cases, few-step sampling and quantization are natural, complementary choices.

However, few-step acceleration is sensitive to the accuracy of the underlying probability-flow ordinary differential equation (ODE) or variance-preserving stochastic differential equation (SDE) that links the noise-estimation network to the final sample (Song et al., 2021). Quantization perturbs that network, inducing a mismatch that alters the ODE/SDE trajectory and amplifies artifacts, an effect that becomes more pronounced as the number of steps shrinks. Simply reusing full-precision schedulers on quantized backbones will inevitably induce quality degradation.

To bridge this gap, we introduce **Q-Sched**, a quantization-aware noise scheduler that adapts the few-step trajectory to the compressed model *without modifying any weights*. Q-Sched inserts lightweight coefficients $(\mathbf{c}^{\mathbf{x}}, \mathbf{c}^{\epsilon})$ into the scheduler (Figures 2a and 2b), correcting quantization-induced drift while keeping a single U-Net/DiT checkpoint reusable across FP16, W8A8, and W4A8 deployments. This design directly targets the constraints above: it preserves the latency benefits of few-step sampling, fits within on-device and MIG memory budgets, and avoids checkpoint sprawl in production.

Our contributions are summarized as follows:

1. In this work, we introduce **Q-Sched**, a quantization-aware scheduler that integrates seamlessly with few-step diffusion models. It achieves up to a **15.5% FID improvement** over a 4-step latent consistency model (LCM) (Luo et al., 2023) baseline and, as shown in Figure 1, can match or surpass full-precision arena scores *while simultaneously reducing model size* on SDXL-Turbo (4-Step) (Sauer et al., 2024) and FLUX.1[schnell] (Black Forest Labs, 2024).



(a) Quantization shifts the diffusion sampling trajectory, reducing fidelity. `Q-Sched` corrects this drift by adapting the scheduler.

(b) At timestep $t \to s$, `Q-Sched` applies lightweight coefficients (c_t^x, c_t^ϵ) within the scheduler, enabling deployment of quantized models from a single U-Net/DiT checkpoint.

Figure 2: `Q-Sched` directly optimizes the few-step diffusion scheduler (see Figures 2a and 2b), addressing quantization-induced trajectory drift without modifying model weights. Unlike prior approaches that alter the transformer or U-Net backbone through retraining or post-training adjustments, `Q-Sched` leaves weights fixed, allowing seamless reuse of one pretrained checkpoint across FP16, W8A8, and W4A8 deployments. This simplifies model management and reduces storage overhead while maintaining high image fidelity.

2. `Q-Sched`'s novel **preconditioning coefficients** enable quantized models to deliberately deviate from potentially overfit few-step baselines (Figure 2a), alleviating oversmoothing and texture artifacts from distillation and quantization while improving the balance between fidelity and artifact severity.
3. To optimize these coefficients, we propose the **Joint Alignment–Quality (JAQ) loss** which balances perceptual fidelity with text–image alignment. Being reference-free, JAQ also enables precise control over visual properties (*e.g.*, texture, detail, saturation) without requiring access to a full-precision model.
4. We establish a **theoretical existence guarantee** (Theorem 1), proving that `Q-Sched` coefficients always exist which reduce expected sampling error relative to the original quantized scheduler. This provides a principled explanation for `Q-Sched` 's systematic improvements.
5. Finally, a large-scale **human preference study** with over 80,000 annotations demonstrates that `Q-Sched` outperforms MixDQ (Zhao et al., 2024a) on SDXL-Turbo and SVDQuant (Li et al., 2025) on FLUX.1[schnell] in terms of perceived image quality.

As illustrated in Figure 1, `Q-Sched` attains the highest ELO rating in pairwise image-quality comparisons among evaluated methods. Furthermore, Figure 3 shows that `Q-Sched` is Pareto-optimal with respect to both ELO and model size, underscoring its ability to balance perceptual quality and efficiency more effectively than competing approaches.

2 Background and Related Work

Diffusion models generate samples by denoising corrupted data across a trajectory of timesteps $t \in [0, T]$, where T is typically large (≥ 25). Each step applies a denoising network $\mathcal{E}_\theta$, conditioned on both t and its noisy input x_t. While this iterative scheme yields high-fidelity samples, invoking a large U-Net or DiT backbone at every step makes inference prohibitively slow in deployment.

Recent few-step models highlight this bottleneck. **SDXL-Turbo** leverages Adversarial Diffusion Distillation (ADD), combining score distillation with an adversarial loss, to reduce sampling to just 1–4 steps, enabling real-time generation on commodity GPUs (Sauer et al., 2024). **FLUX.1[schnell]** introduces a 12B-parameter rectified-flow transformer with open weights, optimized for 1-4 step

inference, making it attractive for latency-constrained serving (Black Forest Labs, 2024). Most recently, **FLUX.1[kontext]** extends the family beyond text-to-image toward *in-context* generation and editing, accepting text and images jointly and unifying both tasks in a flow-matching framework (Labs et al., 2025). These advances exemplify the field's shift toward *deployment-ready* diffusion systems that meet strict latency and memory budgets.

Few-step diffusion and distillation. Few-step methods compress the teacher's long trajectory into a handful of evaluations, preserving most of the fidelity at a fraction of the cost. Distillation is the primary approach: early demonstrations distilled long-run teachers into 1–8 step students, such as Instaflow (Liu et al., 2023), rectified-flow straightening (Liu et al., 2022), and adversarially guided ADD (Sauer et al., 2024). Consistency Models (CMs) (Song et al., 2023) frame generation as a self-consistency mapping from any noisy state to the clean sample, yielding efficient few-step samplers. Variants include Latent Consistency Models (LCMs) (Luo et al., 2023) with Stable Diffusion (Rombach et al., 2022) backbones, Trajectory Consistency Distillation (TCD) (Zheng et al., 2024) with trajectory-aware schedules, and Phased Consistency Models (PCMs) (Wang et al., 2024) with improved guidance and stability. Across these designs, the *scheduler* plays a critical role in determining quality in the few-step regime. The update rule for few-step diffusion models using quantized backbone $\mathcal{E}_\theta^Q$ is:

$$x_s = \Phi(t, x_t, \mathcal{E}_\theta^Q), \tag{1}$$

where x_s denotes the intermediate sample at timestep $s \in [0, t]$ and $\Phi(\cdot)$ is a few-step scheduler. In Section 3, we illustrate our approach using the TCD scheduler (Zheng et al., 2024) as a running example. However, `Q-Sched` is fully general and can be applied on top of any few-step scheduler that fits the abstraction in Equation (1).

Quantization for diffusion models. Post-training quantization (PTQ) has largely targeted $\mathcal{E}_\theta$ and its activations across timesteps. Timestep-aware calibration approaches (PTQ4DM (Shang et al., 2022), ADP-DM (Wang et al., 2023a), Q-Diffusion (Li et al., 2023)), dynamic schemes such as TDQ (So et al., 2024), and error-compensation methods (Q-DM (Li et al., 2024c)) all operate by modifying weights or activations and require full-precision calibration. MixDQ (Zhao et al., 2024a) extends to few-step models with a mixed-precision allocation strategy guided by beggining-of-sentence(BOS)-aware quantization and layer sensitivity analysis. SVDQuant (Li et al., 2025) targets 4-bit weights and activations by absorbing outliers into a high-precision low-rank branch via SVD, shifting variance from activations into weights before fusing the branch back into low-bit kernels.

We posit that in the few-step setting, quantization bias additionally manifests as a *scheduler mismatch*: a fixed full-precision schedule can systematically over- or under-correct, amplifying artifacts. One method that avoids modifying network weights is PTQD (He et al., 2024), which models the quantization-induced shift as an affine perturbation of the full-precision denoiser, $\mathcal{E}_\theta^Q(x_t, t) = (1 + \gamma)\mathcal{E}_\theta + \delta$, and compensates it via variance scaling and a bias term applied directly to the sampler update on x_t. In practice, γ is estimated via standard-deviation matching while δ is treated as uncorrelated Gaussian noise. We adapt PTQD-style bias correction, originally developed for *un-distilled* diffusion models, into TCD (see Section G) and generalize the principle to other few-step samplers as a baseline for our approach.

`Q-Sched` reframes quantized few-step generation as scheduler adaptation. It learns quantization-aware preconditioning coefficients to correct trajectory drift with negligible overhead, while leaving the backbone frozen. The approach integrates seamlessly with few-step schedulers, needs only lightweight calibration, and preserves a single checkpoint across FP16, W8A8, and W4A8. Unlike prior PTQ methods that adjust weights or activations, `Q-Sched` adapts the scheduler itself, complementing existing PTQ and distillation techniques to recover full-precision quality at reduced footprints while retaining the latency benefits of few-step sampling.

3 Quantization-Aware Scheduling

To prepare the TCD scheduler for optimization with `Q-Sched`, let us consider sampling with a quantized network. TCD's Strategic Stochastic Sampling (SSS) (Zheng et al., 2024) using a quantized network $\mathcal{E}_\theta^Q(x_t, t)$ is given by:

$$\mathbf{x_s} = \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t \mathcal{E}_\theta^Q(x_t, t)}{\alpha_t} + \sigma_{s'}\mathcal{E}_\theta^Q(x_t, t)\Big) + \eta\mathbf{z} \tag{2}$$

where the noise schedule is given by σ, α and the sampler injects stochastic noise sampled from a distribution $\mathbf{z} \sim N(0, I)$. The sampler relies on an intermediary timestep, $s' \in [s, t]$, where stochastic noise is added. The degree of randomness is controlled by the stochastic control parameter η:

$$\eta = \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}} \quad . \tag{3}$$

which can be adjusted at sampling time to vary image randomness. The TCD sampler in Equation (2), used in Phased Consistency Models, is a state-of-the-art few-step diffusion method that depends on two inputs from the previous step—x_t and $\mathcal{E}_\theta^Q(x_t, t)$—which are central to applying `Q-Sched`.

`Q-Sched`: A Learnable Schedule Pre-Conditioner We introduce `Q-Sched`, a lightweight post-training method that adapts the noise schedule of few-step diffusion models using two learnable scalar preconditioning coefficients, c_t^x and c_t^ϵ, applied respectively to x_t and $\mathcal{E}_\theta^Q(x_t, t)$ at time t. As illustrated in Figure 2b, `Q-Sched` operates independently of the model backbone (U-Net or transformer), making it broadly compatible with any few-step scheduler resembling TCD.

Under `Q-Sched`, the TCD sampling update becomes:

$$\mathbf{x_s} = \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{c_t^x\mathbf{x_t} - \sigma_t c_t^\epsilon \mathcal{E}_\theta^Q(x_t, t)}{\alpha_t} + \sigma_{s'}c_t^\epsilon\mathcal{E}_\theta^Q(x_t, t)\Big) + \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z} \tag{4}$$

To learn the preconditioning coefficients $(\mathbf{c^x}, \mathbf{c}^\epsilon) := (c_t^x, c_t^\epsilon)_{t=0}^T$, we perform hyperparameter search as outlined in Algorithm 1. In practice, we find grid search is sufficient, as each model involves only two coefficients per timestep across 2–8 timesteps. Even small adjustments to these coefficients yield noticeably crisper images with fewer quantization artifacts.

A natural question arises: **why are two coefficients sufficient to improve image quality?** It turns our that the reconstruction error between full precision and quantized images (at timestep $t = 0$), denoted by Δx_0, can be strictly improved using scheduler coefficients:

Theorem 1 (Strict Existence Guarantees)**.** *There exists Q-Sched coefficients* $(\mathbf{c^x}, \mathbf{c}^\epsilon) \neq 0$ *such that* $E[||\Delta\tilde{x_0}||] < E[||\Delta x_0||]$.

As shown in Appendix H, Δx_0 is a linear combination of per-step denoising errors $\Delta E_\theta(t)$ with coefficients k_t, m_t. Since the error is homogeneous in these terms, rescaling via $\tilde{k}_t = c_t^x k_t$ and $\tilde{m}_t = c_t^\epsilon m_t$ strictly reduces the expected error over naïve quantization. Thus, re-weighting the sampler, without modifying network weights, guarantees a reduction in error with respect to the full precision images. Next, we will discuss our new reference-free loss function, JAQ, and its advantages over existing image assessment tools.

JAQ: A Joint Alignment Quality Loss Function Reference-free metrics such as CLIPScore (Hessel et al., 2021) have become essential for quick evaluation of text-to-image generation models. Unlike FID (Heusel et al., 2017), SSIM (Wang et al., 2004), and other *comparative* metrics, reference-free metrics do not rely on a ground truth reference image and therefore are very useful in generative tasks when a ground truth is not available. When quantizing these generative models, the resultant images, $\hat{x}_0^Q$, are generated by an altered sampling trajectory as evidenced in Figure 7, where $\hat{x}_0^Q$ is a different, sometimes cleaner image than a those derived from the full precision backbone. In short, the quantized model's sampling trajectory coarsely follows the full precision model yet generates sufficient differences that reference-based metrics do not capture the image's detail.

Our Joint Alignment Quality loss combines a text-to-image compatibility score with a pure image quality score to achieve better results than simply optimizing with respect to metrics such as CLIP-Score or CLIP-IQA (Wang et al., 2023b) independently. We design the JAQ loss so that it can better

Algorithm 1 Search for `Q-Sched` Coefficients

Input: search range $[c_{min}, c_{max}]$, search points n, number of diffusion steps ω
loss function JAQ, calibration set $\mathcal{C}$, search optimizer `opt`

Initialize $S^* \leftarrow \infty$
▷ initialize each parameter in uniformly distributed range (c_{min}, c_{max})
$(c^x_{start}, c^x_{end}, c^\epsilon_{start}, c^\epsilon_{end}) \leftarrow \texttt{opt.init}(c_{min}, c_{max})$
for $i \in [0, n]$ **do**
  $\mathbf{c^x} \leftarrow \texttt{linspace}(c^x_{start}, c^x_{end}, \omega)$
  $\mathbf{c^\epsilon} \leftarrow \texttt{linspace}(c^\epsilon_{start}, c^\epsilon_{end}, \omega)$
  $S \leftarrow [\,]$
  for $x \in \mathcal{C}$ **do**
    $S_x \leftarrow \text{JAQ}(x; \mathbf{c^x}, \mathbf{c^\epsilon})$
    $S = S \cup S_x$
  end for
  ▷ $\bar{S}$ is the arithmetic mean of S
  if $\bar{S} < S^*$ **then**
    $S^* \leftarrow \bar{S}$, $\mathbf{c^x_\star} \leftarrow \mathbf{c^x}$, $\mathbf{c^\epsilon_\star} \leftarrow \mathbf{c^\epsilon}$
  end if
  $(c^x_{start}, c^x_{end}, c^\epsilon_{start}, c^\epsilon_{end}) \leftarrow \texttt{opt.step}(\bar{S})$
end for
return $\mathbf{c^x_\star}, \mathbf{c^\epsilon_\star}$

differentiate between images that are highly similar to one another, whereas standard image quality metrics are designed to rank images that come from a much larger distribution. Given a text-to-image compatibility metric, $\texttt{TC}(x)$, and a pure image quality metric, $\texttt{IQ}(x)$, JAQ combines them as follows:

$$\texttt{JAQ}(x) = \texttt{TC}(x) + k \cdot \texttt{IQ}(x) \tag{5}$$

Optimizing solely for text–image compatibility (e.g., CLIPScore) sacrifices visual detail and fails to capture quantization artifacts (Figures 6 and 8). Conversely, relying only on image quality can generate extraneous details. JAQ balances these objectives through a linear combination, with k controlling the tradeoff between prompt fidelity and image detail.

4 Experiments

Experimental Setup We apply `Q-Sched` across diverse few-step diffusion models, including U-Net (Ronneberger et al., 2015) and DiT (Peebles & Xie, 2023) backbones, and across different distillation strategies: consistency-based (LCM (Luo et al., 2023), PCM (Wang et al., 2024)) and flow-matching approaches (SDXL-Turbo (Sauer et al., 2024) and FLUX.1[schnell] (Black Forest Labs, 2024)). We quantize models in both 4-bit weights, 8-bit activations (W4A8) and 8-bit weights, 8-bit activations (W8A8). Only the U-Net or DiT backbone is quantized, as it dominates model size (see Table 5).

LCM and PCM are tested at 2, 4, and 8 steps on COCO-30k (Lin et al., 2014), using FID (vs. real), `CLIPScore` (prompt alignment), and FID-SD (vs. Stable Diffusion). FLUX.1 and SDXL-Turbo are evaluated on the SVDQuant (Li et al., 2025; 2024b) subset of MJHQ-30k (5,000 high-quality Midjourney prompts in 10 categories), using FID and human preference studies to capture perceptual quality.

We employ two variants of the Joint Alignment Quality (JAQ) loss: one derived from CLIP-based metrics and another from human preference scores. In the CLIP-based variant, we set `TC(x)` = `CLIPScore(x)` and `IQ(x)` = `CLIP-IQA(x)`. For SDXL-Turbo and FLUX.1, we instead adopt a preference-based variant, with `TC(x)` = `AQ-MAP(x)` and `IQ(x)` = `HPSV2(x)`. Here, AQ-MAP (Li et al., 2024a) provides a spatial alignment score, while HPSV2 (Wu et al., 2023) is fine-tuned on real human judgments. In both cases, we fix $k = 2$.

Table 1: Comparison of different schedulers on Phased Consistency Models and Latent Consistency Models using a Stable Diffusion v1-5 backbone. The original schedule is TCD (Zheng et al., 2024) for Phased Consistency Models and the Multi-step Consistency Sampling (Luo et al., 2023) for Latent Consistency Models. The FID and CLIPScore are calculated with respect to the COCO-30k dataset. NFEs stands for *number of function evaluations* referring to the number of passes through the network $\mathcal{E}_\theta^Q(x_t, t)$.

NFEs	Precision	Schedule	PCMs		LCMs	
			FID	CLIPScore	FID	CLIPScore
2	FP16	Original	24.17	25.489	38.74	25.155
	W4A8	Original	28.70	25.343	40.93	24.886
	W4A8	PTQD	23.33	25.265	37.59	24.919
	W4A8	Q-Sched	**22.24**	**25.543**	**32.50**	**25.152**
4	FP16	Original	23.29	25.482	31.94	**25.969**
	W4A8	Original	23.08	25.557	38.41	25.456
	W4A8	PTQD	19.42	25.639	39.72	24.678
	W4A8	Q-Sched	**17.39**	**25.715**	**26.98**	25.336
8	FP16	Original	20.15	**25.714**	27.34	**26.052**
	W4A8	Original	18.48	25.664	27.55	25.397
	W4A8	PTQD	**15.85**	25.770	28.06	25.241
	W4A8	Q-Sched	16.83	25.698	**25.82**	25.214

Results: Latent and Phased Consistency Models In Table 1, we evaluate three schedulers across two consistency model families and show that Q-Sched learns a new few-step trajectory that mitigates artifacts and can even surpass both FP16 and W4A8 in detail. It achieves strong FID scores and outperforms PTQD in 4/6 consistency variants on Stable Diffusion v1-5, while using only a fraction of the calibration set. We compare with PTQD (He et al., 2024), the only other quantization-aware scheduler for few-step diffusion. Unlike PTQD, which relies on a 1,024-image full-precision calibration set, Q-Sched requires only 20 representative sDCI prompts (Li et al., 2025), reused across evaluations. Unlike PTQD, which requires full-precision references, Q-Sched operates with just twenty prompts and can exceed a full precision few-step model by **16.1%, 15.5%, and 5.6%** at 2, 4, and 8 steps, respectively. This highlights that quantization and few-step distillation act as complementary compression strategies.

Scheduler	Precision	FID	FID-SD	CLIPScore
TCD	FP16	18.65	10.45	26.531
TCD	W4A8	22.70	12.51	26.241
PTQD	W4A8	161.96	176.29	25.910
Q-Sched	W4A8	**18.89**	**12.17**	**26.513**

(a) Comparison on a 2-step Phased Consistency model using the Stable Diffusion XL backbone. FID-SD is computed relative to images generated by Stable Diffusion XL using corresponding COCO-30k prompts.

Method	Precision	FID
-	FP16	25.48
Naive	W4A8	25.75
MixDQ	W4A8	25.36
Q-Sched	W4A8	**21.41**
Naive	W8A8	25.49
MixDQ	W8A8	**25.16**
Q-Sched	W8A8	26.34

(b) Quantized model comparison on SDXL-Turbo under varying bitwidths. FID is computed on the MJHQ dataset.

Table 2: Quantitative evaluation of Large-scale few-step diffusion models with a Stable Diffusion XL backbone. W4A8 and W8A8 are a $4\times$ and $8\times$ model size reduction in comparison to FP16, yet our method improves over baseline. As FID, FID-SD, and CLIPScore may exhibit reduced reliability at large model scales, we complement these metrics with user preference studies in Figure 3.

In Table 2a, we evaluate a large-scale 2-step Phased Consistency Model on the Stable Diffusion XL backbone. Q-Sched incurs only a 1.2% FID drop in W4A8, showing that quantization-aware preconditioning preserves quality even under aggressive compression. By contrast, PTQD degrades

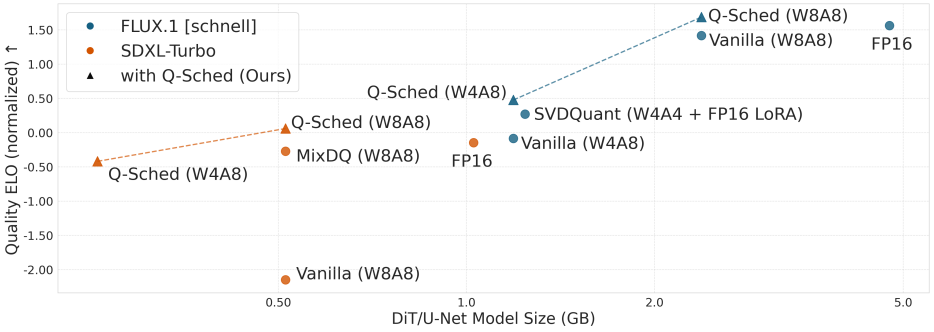


Figure 3: ELO Score *vs.* Model Size for various quantization methods on FLUX.1[schnell] (Black Forest Labs, 2024) and SDXL-Turbo (Sauer et al., 2024).

sharply, as its Gaussian noise assumption breaks down in few-step diffusion—particularly for large models where each step approximates an ODE segment rather than a Gaussian denoising step.

Results: SDXL-Turbo and FLUX.1[schnell] In Table 2b, we compare quantization strategies on SDXL-Turbo (4-step inference) using the FID metric on the MJHQ dataset, evaluating two bitwidth settings: W4A8 and W8A8. Under W4A8, `Q-Sched` achieves a FID of 21.41, significantly outperforming MixDQ (Zhao et al., 2024a) (25.36) and Naive (25.75), demonstrating strong robustness to aggressive quantization. However, at W8A8, `Q-Sched` shows a higher FID (26.34) than both MixDQ (25.16) and Naive (25.49), suggesting that its advantages are most pronounced in lower-bit regimes, where other methods degrade more severely.

In Figure 3, we present user preference results for `Q-Sched` applied to both SDXL-Turbo and FLUX.1 [schnell], showing that it outperforms MixDQ (Zhao et al., 2024a) and SVDQuant (Li et al., 2025), respectively, at similar model sizes (see Section A for details). We compute an ELO rating, a relative quality ranking inspired by chess scoring, by aggregating all pairwise 1v1 image comparisons across models, where a higher score reflects consistent user preference.

In Figure 4, we compare `Q-Sched` across bit-widths using a user study. W4A4 proved too aggressive, but W4A5 and W4A6 produced images comparable to full precision. 1v1 comparisons with full-precision FLUX.1 (Black Forest Labs, 2024) follow the protocol in Appendix A.

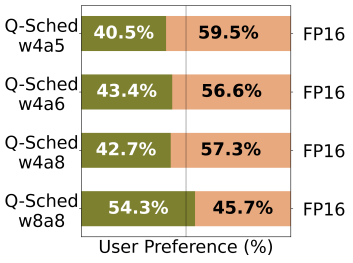


Figure 4: Comparing `Q-Sched` across various bit-widths.

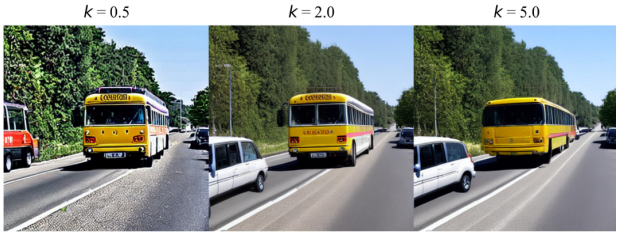
Comparison with Image Quality Metrics We also evaluate `Q-Sched` on FLUX.1[schnell] and SDXL-Turbo using human preference metrics, showing that JAQ effectively captures image fidelity. Beyond the metrics we've already mentioned, we compare with PickScore (Kirstain et al., 2023) which predicts human preferences from large-scale image–text comparisons, and MANIQA (Yang et al., 2022b) which uses multi-dimensional attention to assess perceptual quality without references. As seen in Table 3, JAQ aligns closely with established metrics while uniquely balancing fine-grained details often degraded by quantization.

Ablation on Pre-Conditioning Coefficients and Loss Function Choice We ablate the choice of pre-conditioning coefficients in the Phased Consistency Model by comparing performance when optimizing only the model-side coefficient $\mathbf{c}^{\epsilon}$, the sample-side coefficient $\mathbf{c}^{\mathbf{x}}$, or both jointly. As shown in Figure 5b, jointly optimizing both $\mathbf{c}^{\epsilon}$ and $\mathbf{c}^{\mathbf{x}}$ consistently yields the best results across all three metrics: PickScore, HPSv2, and JAQ Loss. These findings highlight the importance of treating both denoising and reconstruction terms as tunable components rather than fixing one a priori. All metrics are averaged over 1024 images generated with the SDXL backbone.

Table 3: Comparison across image quality metrics. "MixDQ" refers to the W8A8 MixDQ (Zhao et al., 2024a) variant and "SVDQ" refers to LoRA-based W4A4 SVDQuant Li et al. (2025).

	SDXL-Turbo (4-Step)				FLUX.1 [schnell]			
	FP16	W8A8	MixDQ	W8A8 Q-Sched	FP16	W4A8	SVDQ	W4A8 Q-Sched
CLIP Score ↑	**25.62**	**25.62**	25.38	25.36	**25.61**	25.17	25.52	25.27
CLIP IQA ↑	0.725	0.727	0.727	**0.731**	**0.716**	0.712	0.714	0.707
HPV2 ↑	0.276	0.276	0.275	**0.278**	**0.275**	0.274	**0.275**	0.272
AQ-MAP ↑	0.693	0.694	0.693	**0.696**	**0.700**	**0.700**	0.697	**0.700**
Pick Score ↑	18.48	18.49	18.48	**18.51**	18.43	18.42	18.40	**18.46**
MANIQA ↑	0.508	0.513	0.502	**0.511**	**0.528**	0.500	0.514	0.506
JAQ (ours) ↑	1.663	1.665	1.659	**1.669**	**1.676**	1.675	1.669	1.673

How Do We Choose *k* For The JAQ Loss? We optimize the `Q-Sched` preconditioners using the JAQ loss, which balances image quality and text-image consistency via a tradeoff hyperparameter, k. As shown in Figure 5a, small k values can lead to color distortion, while larger values (*e.g.*, $k = 5$) cause outputs to drift from the true data distribution. In such cases, the JAQ loss behaves similarly to CLIP-IQA-Q, which lacks sensitivity to concept alignment. We find that a hand-tuned value of k is sufficient for producing a high-quality noise schedule, and the final results are not highly sensitive to its exact choice. Throughout our experiments, we use $k = 2$.



(a) Choice of k for the JAQ loss. k balances the contribution of $\texttt{TC}(x)$ *vs.* $\texttt{IQ}(x)$. Prompt: "a car and a bus on a french highway".

	c^ϵ	c^x	(c^ϵ, c^x)
PickScore	21.83	22.25	**22.30**
HPSV2	**0.288**	0.262	**0.288**
JAQ Loss	3.367	3.383	**3.392**

(b) Ablation on choice of pre-conditioning coefficients. We find that optimizing both model and sample coefficients jointly yields optimal image quality. Image quality metrics are averaged over 1024 images generated from the Phased Consistency Models with the SDXL backbone.

Figure 5: Ablation studies on various design choices for `Q-Sched`.

5 Conclusion

Few-step diffusion models dramatically reduce inference cost by distilling large generative models, such as Stable Diffusion XL, into versions requiring only 2–8 denoising steps, achieving a 5–25× speedup. However, these models typically reduce runtime without addressing model size. Our method, `Q-Sched`, pushes this efficiency frontier further by introducing quantization into the few-step regime. Through noise-aware preconditioning coefficients, `Q-Sched` enables effective quantization with minimal performance loss. We report **8.0% and 16.1% FID improvements** over full-precision baselines for PCMs and LCMs, respectively. A user preference study also shows that **`Q-Sched` outperforms existing quantization methods on FLUX.1[schnell] and SDXL-Turbo in perceived image quality**. These results demonstrate that quantization and few-step distillation are complementary, enabling substantial efficiency gains without compromising generation quality.

6 Ethics Statement

Model compression broadens the accessibility of AI by enabling large foundation models to run on resource-constrained GPUs. The potential societal consequences of our work are similar to those of prior approaches, as both quantization and few-step diffusion serve as compression methods for text-to-image generative models. Such models can produce synthetic images that may mislead, misrepresent, or cause social harm. We conduct a user preference study on a crowdsourcing platform

in which participants worldwide are shown generated content, which, like all synthetic media, carries inherent potential for misuse and harm.

7 LLM USAGE

We made use of large language models (LLMs) to assist in the preparation of this manuscript. LLMs were employed for language polishing, formatting support (e.g., LaTeX macros, algorithm pseudocode, figure/table captions), and iterative feedback on clarity and conciseness of explanations.

REFERENCES

Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Black Forest Labs. Flux.1-schnell. `https://huggingface.co/black-forest-labs/FLUX.1-schnell`, 2024. Accessed: 2025-05-14.

Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4175–4186, 2022.

Chentao Cao, Zhuo-Xu Cui, Shaonan Liu, Dong Liang, and Yanjie Zhu. High-frequency space diffusion models for accelerated mri. *arXiv preprint arXiv:2208.05481*, 2022.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *Advances in neural information processing systems*, 36, 2024.

Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. Squant: On-the-fly data-free quantization via diagonal hessian approximation. *arXiv preprint arXiv:2202.07471*, 2022.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.

Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Baolin Li, Tirthak Patel, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. Miso: exploiting multi-instance gpu capability on multi-tenant gpu clusters. In *Proceedings of the 13th Symposium on Cloud Computing*, pp. 173–189, 2022a.

Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Q-refine: A perceptual quality refiner for ai-generated image. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024a.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024b.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Junxian Guo, Xiuyu Li, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank component for 4-bit diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343, 2022b.

Xiuyu Li, Long Lian, Yijiang Liu, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. *arXiv preprint arXiv:2302.04304*, 2023.

Yanjing Li, Sheng Xu, Xianbin Cao, Xiao Sun, and Baochang Zhang. Q-dm: An efficient low-bit quantized diffusion model. *Advances in Neural Information Processing Systems*, 36, 2024c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

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.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 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.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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. *arXiv preprint arXiv:2307.01952*, 2023.

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.

Rapidata. Rapidata: An api that provides fast access to large-scale human evaluations, 2025. URL `https://www.rapidata.ai/`. Accessed: 2025-05-16.

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*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. *arXiv preprint arXiv:2211.15736*, 2022.

Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate data-free quantization for diffusion models. *arXiv preprint arXiv:2305.18723*, 2(5), 2023a.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency model. *arXiv preprint arXiv:2405.18407*, 2024.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *CoRR*, 2023.

Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022a.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Huaizheng Zhang, Yuanming Li, Wencong Xiao, Yizheng Huang, Xing Di, Jianxiong Yin, Simon See, Yong Luo, Chiew Tong Lau, and Yang You. Migperf: A comprehensive benchmark for deep learning training and inference workloads on multi-instance gpus. *arXiv preprint arXiv:2301.00407*, 2023.

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*, pp. 285–302. Springer, 2024a.

Yang Zhao, Yanwu Xu, Zhisheng Xiao, Haolin Jia, and Tingbo Hou. Mobilediffusion: Instant text-to-image generation on mobile devices. In *European Conference on Computer Vision*, pp. 225–242. Springer, 2024b.

Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation. *arXiv preprint arXiv:2402.19159*, 2024.

A Details on User Preference Assessment

We design our evaluation setup following the user preference study methodology from SDXL-Turbo (Sauer et al., 2024), with several improvements. For each model pair in this study, we perform 1-vs-1 comparisons based on shared prompts. Human responses, collected via Rapidata (Rapidata, 2025), come from evaluators who are presented with two images, each generated by a different model for the same prompt, and are asked: "Which image is of higher quality and more aesthetically pleasing?"

Evaluators are globally sourced and must pass a set of validation questions designed to assess annotation quality. Only those who successfully complete this qualification step are allowed to rate the models.

ELO scores are computed using the same approach as SDXL-Turbo (Sauer et al., 2024), with K = 32 K=32. We find that this value of K enables more noticeable ranking adjustments, especially when models have similar performance levels.

All models in our study are evaluated using 1,000 prompts sampled from the MJHQ-30k dataset. We release this subset, which we call the `Q-Sched` split, to enable consistent benchmarking of future quantization methods. Each prompt is evaluated by four unique annotators. Therefore, each 1-vs-1 comparison results in 4,000 total human annotations.

B Compute Resources & Statistical Significance

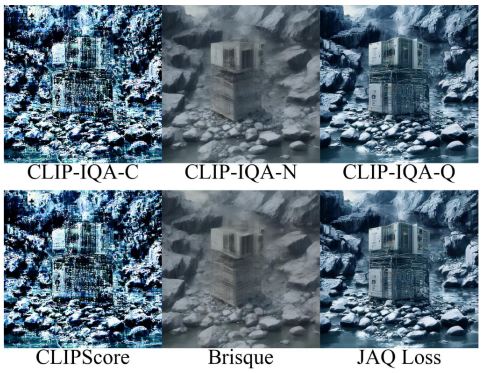
We conduct all our experiments on a high-end AI server with eight Nvidia A6000s. Each model can be run independently on one A6000 and `Q-Sched` takes approximately twenty minutes to run the full grid search.

Our main experiments are averaged over two-three runs but we do not report error bars at this time.

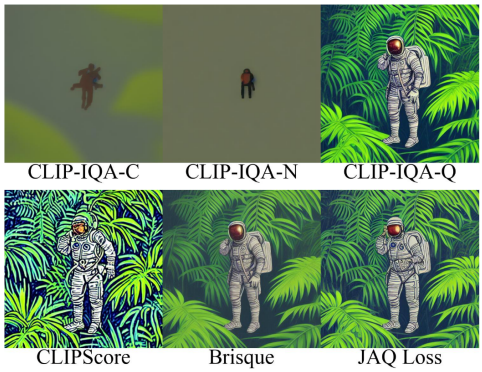
C Ablation Studies

C.1 Comparing Loss Functions for Q-Sched

To evaluate the overall image quality for text-image generative modeling, CLIPScore (Hessel et al., 2021) is specifically designed to capture text-image compatibility and does not consider overall image quality. In Figure 6, we illustrate that `Q-Sched` optimized with CLIPScore produces an updated noise schedule that is over saturated and lacks image depth. In contrast, Brisque (Mittal et al., 2012) is often used as a standard reference-free image quality metric, but when used in `Q-Sched` it creates images with smoother and less detailed features. We consider three variants of CLIP-IQA (Wang et al., 2023b) and find that CLIP-IQA using the predefined quality prompt (we denote this version by CLIP-IQA-Q) achieves a noise schedule with high-fidelity images. However, CLIP-IQA-Q has a



(a) Prompt: "An Eniac computer balanced on top of a stack of rocks over a river"



(b) Prompt: "Astronaut in a jungle, cold color palette, muted colors, detailed, 8k"

Figure 6: Optimizing `Q-Sched` with various reference-less image quality metrics. Our loss function, JAQ, is a linear combination of CLIPScore and CLIP-IQA-Q. We compare against three CLIP-IQA prompts: Complexity, Noisiness, and Quality denoted as -C, -N, -Q respectively.

Table 4: Adding stochasticity and its effect on W4A8 quantization for PCM using a Stable Diffusion v1-5 backbone. We report FID on COCO-30k. The stochasticity term, η, controls the amount of added Gaussian noise. $\eta = 0$ is deterministic sampling.

Method	$\eta =$					
	0	0.1	0.3	0.5	0.7	0.9
TCD	28.70	24.06	**23.44**	22.97	26.74	22.40
PTQD	23.33	25.59	24.95	25.69	24.53	26.71
`Q-Sched`	**22.24**	**19.29**	**23.44**	**19.67**	**19.46**	**17.87**

significant weakness: it cannot properly score images with hallucinations because it does not have an understanding of the underlying image prompt or concept. Therefore, we combine the benefits of CLIPScore and CLIP-IQA-Q into the JAQ loss and find that the resulting schedule fares extremely well with respect to raw image quality as well as to concept adherence.

C.2 Adding Stochasticity

Phased Consistency Model's implementation of the original sampler, TCD, is deterministic, meaning that there is no additive noise during sampling. The controllable noise parameter, η, allows a practitioner to adjust the additive noise during the sampling process and is defined in Equation (2). In order to compare PTQD's correction to our method, we ablate across different levels of stochasticity and report performance for six stochasticity levels in Table 4. $\eta = 0$ refers to deterministic sampling and PTQD's uncorrelated noise correction is not used since it adds stochastic noise by construction. Please see the appendix for more details on PTQD's implementation in both deterministic and stochastic sampling regimes.

We find that `Q-Sched` outperforms PTQD for all stochasticity regimes on the 2-step phased consistency model. With a simple grid search using our JAQ loss, we can outperform PTQD and the original TCD scheduler in different sampling regimes.

D Quantization-Induced Artifacts

As shown in Figure 7, `Q-Sched` is able to generate images that differ sufficiently from the full precision model. We ground our quantized diffusion model with image quality metrics, rather than it's error with respect to full preicision.

In our preliminary analysis using a two-step Consistency Model, we observed several characteristic ways in which quantization degrades image quality. As shown in Figure 8, quantized models tend

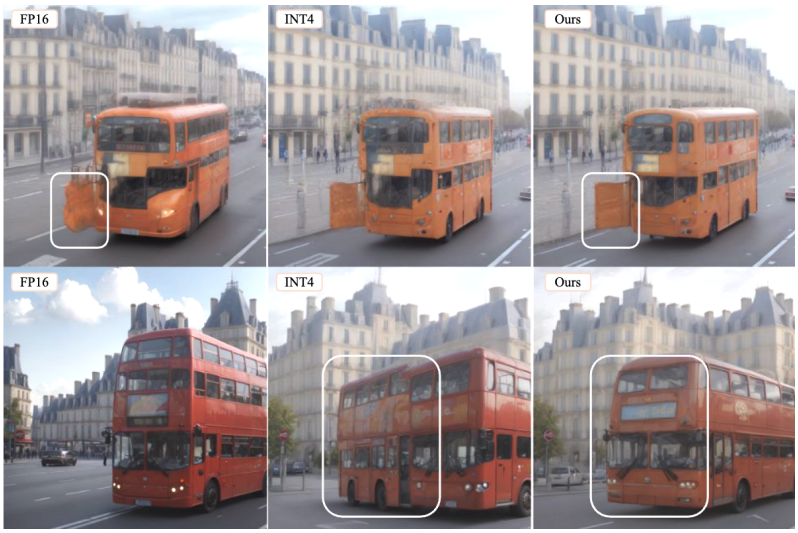


Figure 7: 4-Step (top row) and 8-Step (bottom row) LCMs. Prompt: "a car and a bus on a french highway". `Q-Sched` is capable of avoiding artifacts present in the FP16 or INT4 generative images. `Q-Sched` is close to the original schedule since it generates similar images yet our optimized schedule allows for `Q-Sched` to avoid some artifacts generated from the original schedule.

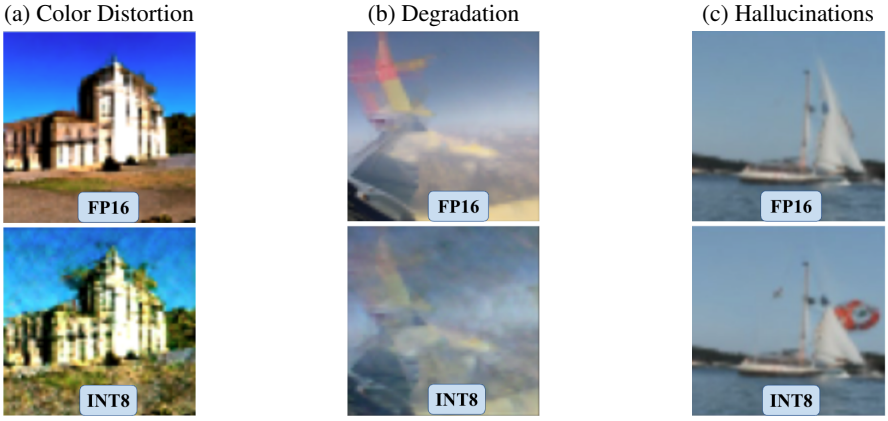


Figure 8: Three types of image artifacts that occur when quantizing image generation models. Images are unconditionally generated from a Two-Step Consistency Model (Song et al., 2023).

to exhibit three prominent types of artifacts: color distortion, image degradation, and hallucinated structures. These issues are especially pronounced in low-bit settings and appear consistently across a variety of models and prompts.

E Model Size Analysis

In Table 5, we show the full model size breakdown for the diffusion model backbone, text encoders, and the VAE decoder. During inference, either one or both text encoders are used, and we do not need the VAE encoder, since this is for training exclusively.

For our ELO *vs.* Model Size Pareto front in Figure 3, we consider the DiT memory and compute model size by taking the parameter count and multiplying it by the number of bytes required per parameter. For W4A4 + LoRA 64, the setup used for SVDQuant (Li et al., 2025), we compute the number of LoRA parameters using the back-of-the-envelope calculation provided in SVDQuant and add it to this calculation. We provide raw data for clarity in Table 6.

Table 5: FP16 Diffusion Model Size Breakdown (in GB)

	LCM	PCM	SDXL-Turbo	FLUX.1[schnell]
UNet/DiT	1.72	4.84	1.03	4.76
Text Encoder(s)	0.25	0.29	0.33	1.95
VAE Decoder	0.07	0.13	0.02	0.02
Total	2.04 GB	5.26 GB	1.37 GB	6.73 GB

Table 6: DiT Memory (in GB) for various bitwidths.

Precision	SDXL-Turbo	FLUX.1[schnell]
FP16	1.03	4.76
W8A8	0.51	2.38
W4A4 + LoRA 64	0.28	1.24
W4A8	0.26	1.19

F Additional Analysis on COCO-30k

This result reinforces the core finding of our paper: quantization, when paired with a scheduler designed to account for noise sensitivity (as in `Q-Sched`), can be synergistic with few-step diffusion rather than detrimental. Notably, our quantized model achieves a lower FID than the original full-precision model, suggesting that `Q-Sched` helps overcome limitations introduced by both step reduction and bit-level compression.

These findings complement the results on SDXL-Turbo and FLUX.1[schnell] discussed in the main paper, and further establish `Q-Sched` as a general-purpose solution for high-fidelity, compressed diffusion generation.

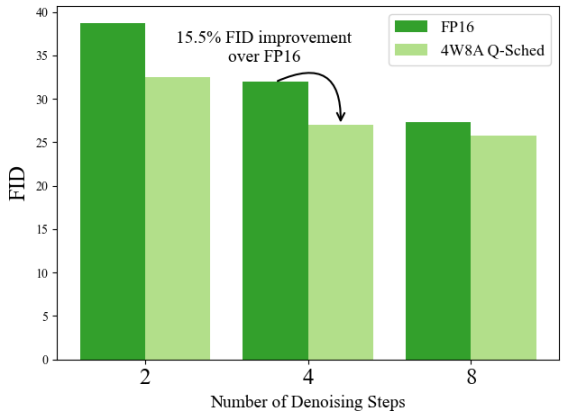


Figure 9: FID on COCO-30k. A W4A8 compressed model with our `Q-Sched` scheduler outperforms its FP16 counterpart with a $4\times$ reduction in model size.

G Applying PTQD to the TCD Scheduler

Using PTQD's linear parameterization for the quantization error, we substitute $\mathcal{E}_\theta^Q(x_t, t) = (1+\gamma)\cdot \mathcal{E}_\theta + \delta$ into Equation (2):

$$\mathbf{x_s} = \frac{\alpha_s}{\alpha_{s'}}\left(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t \mathcal{E}_\theta(x_t,t)}{\alpha_t} + \sigma_{s'}\mathcal{E}_\theta(x_t,t)\right) + \frac{\alpha_s}{\alpha_{s'}(1+\gamma)}(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t})\delta + \sqrt{1-\frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z}. \quad (6)$$

PTQD assumes the uncorrelated noise is sampled from a normal distribution $\delta \sim N(\mu_\delta, \sigma_\delta)$. This method applies bias correction to handle the mean deviation, μ_δ, and analytically compute standard deviation, σ_δ. We adapt PTQD's approach to the TCD schedule and use the new standard deviation, σ_δ for sampling δ:

$$\sigma_\delta^2 = 1 - \frac{\alpha_s^2}{\alpha_{s'}^2}(1 - (\frac{\delta(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t})}{(1+\gamma)})^2). \tag{7}$$

For the edge case, where $\sigma_\delta < 0$, the deviation is set to zero ($\sigma_\delta = 0$). The proof for extending PTQD to the TCD scheduler is in the appendix.

PTQD attempts to model the distribution shift from a full precision to quantized model using two assumptions:

1. The quantized model's distribution shift can be modeled through a linear correction term.
2. The uncorrelated quantization noise is normally distributed.

While these assumptions are similar to prior work on diffusion models, they are likely to break down on the few-step diffusions where the denoising process is distilled from many steps and is not expected to be linear nor follow a Gaussian distribution.

Quantization Noise Correction using PTQD Based on the PTQD quantization noise assumption, the quantization error is linearly parametrized as $\Delta\mathcal{E}_\theta = \gamma \cdot \mathcal{E}_\theta + \delta$ where γ, δ are learnable parameters corresponding to the correlated noise w.r.t. full precision and the uncorrelated noise respectively. PTQD models the uncorrelated noise as Gaussian (*i.e.*, $\delta \sim \mathcal{N}(\mu_q, \sigma_q)$).

Variance Schedule Calibration for Trajectory Consistency Distillation (TCD) TCD's Strategic Stochastic Sampling (SSS) using a quantized network $\mathcal{E}_\theta^Q(x_t, t)$ is given by:

$$\mathbf{x_s} = \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t\mathcal{E}_\theta^Q(x_t,t)}{\alpha_t} + \sigma_{s'}\mathcal{E}_\theta^Q(x_t,t)\Big) + \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z} \tag{8}$$

Using PTQD's linear parametrization for the quantization error, we substitute $\mathcal{E}_\theta^Q(x_t,t) = (1+\gamma)\cdot\mathcal{E}_\theta + \delta$:

$$\mathbf{x_s} = \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t((1+\gamma)\cdot\mathcal{E}_\theta(x_t,t) + \delta)}{\alpha_t} + \sigma_{s'}((1+\gamma)\cdot\mathcal{E}_\theta(x_t,t) + \delta)\Big) + \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z} \tag{9}$$

$$= \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t(1+\gamma)\mathcal{E}_\theta(x_t,t)}{\alpha_t} + \sigma_{s'}(1+\gamma)\mathcal{E}_\theta(x_t,t) - \frac{\alpha_{s'}\sigma_t\delta}{\alpha_t} + \sigma_{s'}\delta\Big) + \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z} \tag{10}$$

$$= \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t(1+\gamma)\mathcal{E}_\theta(x_t,t)}{\alpha_t} + \sigma_{s'}(1+\gamma)\mathcal{E}_\theta(x_t,t)\Big) + \frac{\alpha_s}{\alpha_{s'}}(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t})\delta + \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z} \tag{11}$$

$$\tag{12}$$

The correlated noise can be corrected by applying:

$$\frac{\mathcal{E}_\theta^Q(x_t,t)}{1+\gamma} = \frac{(1+\gamma)\mathcal{E}_\theta(x_t,t) + \delta}{1+\gamma} \tag{13}$$

$$= \mathcal{E}_\theta(x_t,t) + \frac{\delta}{1+\gamma} \tag{14}$$

The resultant SSS sampling step becomes:

$$\mathbf{x_s} = \frac{\alpha_s}{\alpha_{s'}}\Big(\alpha_{s'}\frac{\mathbf{x_t} - \sigma_t\mathcal{E}_\theta(x_t,t)}{\alpha_t} + \sigma_{s'}\mathcal{E}_\theta(x_t,t)\Big) + \frac{\alpha_s}{\alpha_{s'}(1+\gamma)}(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t})\delta + \sqrt{1-\frac{\alpha_s^2}{\alpha_{s'}^2}}\mathbf{z} \tag{15}$$

$$\tag{16}$$

The variance schedule becomes:

$$\sigma_\delta^2 = 1 - \frac{\alpha_s^2}{\alpha_{s'}^2} - (\frac{\alpha_s}{\alpha_{s'}(1+\gamma)}(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t}))^2\delta^2 \tag{17}$$

$$= 1 - \frac{\alpha_s^2}{\alpha_{s'}^2}(1 - \frac{(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t})^2}{(1+\gamma)^2}\delta^2) \tag{18}$$

$$= 1 - \frac{\alpha_s^2}{\alpha_{s'}^2}(1 - (\frac{\delta(\sigma_{s'} - \frac{\alpha_{s'}\sigma_t}{\alpha_t})}{(1+\gamma)})^2) \tag{19}$$

Since $\mathbf{z} \sim N(\mu_\delta, \sigma_\delta)$, we must handle the edge case when $\sigma_\delta < 0$. If the variance is negative, we simply set $\sigma_\delta = 0$.

Upon comparing `Q-Sched` to PTQD you may ask "*Why is `Q-Sched` able to learn a better noise schedule when it is also a linear correction?*" `Q-Sched` learns scalar coefficients on x_t and $\mathcal{E}_\theta$ that are optimized with respect to the reference-free JAQ loss. This allows us to learn a new schedule with linear corrections to improve our overall noise schedule, rather than matching the existing full precision schedule. This is an important distinction from PTQD, which tries to learn a linear correction with respect to full precision, which may not be possible since quantization produces a nonlinear distortion on the diffusion model. In short, PTQD attempts to match the full precision sampling trajectory, whereas `Q-Sched` aims to learn a new sampling trajectory given a compressed $\mathcal{E}_\theta$.

H Proof of Theorem 1: Strict Existence Guarantees for Quantization-Aware Scheduling

Theorem 1 (Strict Existence Guarantees)**.** *There exists Q-Sched coefficients* $(\mathbf{c}^\epsilon, \mathbf{c}^\mathbf{x}) \neq 0$ *such that* $E[||\Delta\tilde{x_0}||] < E[||\Delta x_0||]$.

Proof. Let us consider the few-step sampling trajectories for the pre-trained and quantized models, parametrized by $\mathcal{E}_\theta(t)$ and $\mathcal{E}_\theta^Q(t)$ respectively. These two few-step diffusion models sample at the same time-steps, $0 = t_0 < t_1, t_2 \cdots t_N = T$, where N represents the number of steps in the few-step model. For ease of notation, we will use the time-step 0 to refer to t_0 and 1 to refer to t_1, *etc.* A denoising step going from time $t+1 \to t$, produces a partially denoised image, x_t, and its quantized counterpart, x_t^Q. Following directly from Equation 9, the denoising error, $\Delta x_t = x_t - x_t^Q$, can be explicitly computed as:

$$\Delta x_t = \frac{\alpha_t}{\alpha_{t'}}\Big(\alpha_{t'}\frac{\Delta x_{t+1} - \sigma_{t+1}(\mathcal{E}_\theta(t+1) - \mathcal{E}_\theta^Q(t))}{\alpha_{t+1}} + \sigma_{t'}(\mathcal{E}_\theta(t+1) - \mathcal{E}_\theta^Q(t+1))\Big) \tag{20}$$

$$= \frac{\alpha_t}{\alpha_{t+1}}\Delta_{t+1} + \frac{\alpha_t}{\alpha_{t'}}(\sigma_{t'} - \frac{\sigma_{t+1}}{\alpha_{t+1}})(\mathcal{E}_\theta(t+1) - \mathcal{E}_\theta^Q(t+1)) \tag{21}$$

$$= k_t\Delta x_{t+1} + m_t\Delta\mathcal{E}_\theta(t+1)) \tag{22}$$

where we define the sampler coefficients as $k_t = \frac{\alpha_t}{\alpha_{t+1}}$, $m_t = \frac{\alpha_t}{\alpha_{t'}}(\sigma_{t'} - \frac{\sigma_{t+1}}{\alpha_{t+1}})$ and denote the change in the network as $\Delta\mathcal{E}_\theta(t) = \mathcal{E}_\theta(t) - \mathcal{E}_\theta^Q(t)$. Assuming the initial denoised image is the same ($x_N = x_N^Q$), the error in the final denoised image, Δx_0, is given by:

$$\Delta x_0 = k_0 \Delta x_1 + m_0 \Delta \mathcal{E}_\theta(1) \tag{23}$$

$$= k_0 k_1 k_2 ...(k_N \Delta x_N + m_{N-1} \Delta \mathcal{E}_\theta(N)) + \cdots + k_0 k_1 m_2 \Delta \mathcal{E}_\theta(3) + k_0 m_1 \Delta \mathcal{E}_\theta(2) + m_0 \Delta \mathcal{E}_\theta(1) \tag{24}$$

$$= \sum_{s=1}^{S} \Big(\Pi_{v=0}^{s-2} k_v \Big) m_{s-1} \Delta \mathcal{E}_\theta(s) \tag{25}$$

The average expected error over all images in a given dataset, $x_0 \in \mathcal{D}$, is given by:

$$E[||\Delta x_0||] = \sum_{s=1}^{S} \Big(\Pi_{v=0}^{S-2} k_v \Big) m_{s-1} E[||\Delta \mathcal{E}_\theta(s)||] \tag{26}$$

since $E[||\Delta x_0||]$ is a homogeneous function.

In Q-Sched, we apply our quantization-aware pre-conditioning on every noise coefficient: $\tilde{m}_t = c_t^\epsilon \cdot m_t$ and $\tilde{k}_t = c_t^x \cdot k_t$. Let us denote the expected error induced by Q-Sched with respect to the pre-trained model's x_0 as $E[||\Delta \tilde{x_0}||]$.

We empirically show in Tables 1 and 2 that $E[||\Delta x_0||] \neq 0$ since the images produced by the naive quantization method produce a different FID from the original pre-trained model's image distribution. Since Equation 26 is a linear function of $k_t, m_t, \forall t \in 1 \dots N$, and there is a global minimum at $E[||x_0 - x_0||] = 0$, it must be that $\exists \tilde{m}_t^*, \tilde{k}_t^* \forall t$ such that $E[||\Delta \tilde{x_0}||] < E[||\Delta x_0||]$. In short, we guarantee that there exists quantization-aware coefficients that strictly improve our expected quantization error over naive quantization. □

H.1 Aside: Positive Sampler Coefficients

The TCD Scheduler has $\beta_0 = 0.0085, \beta_N = 0.012$, $\alpha_t = 1 - \beta_t, \sigma_t = \Pi_{i=0}^{t} \alpha_i$ with a scaled linear schedule:

$$\beta_t = \Big(\sqrt{\beta_0} + t \cdot (\sqrt{\beta_N} - \sqrt{\beta_0}) \Big)^2 \tag{27}$$

Therefore: $1 > \alpha_0 > \alpha_1 > \cdots > \alpha_N > 0$ and $1 > \sigma_0 > \sigma_1 > \dots \sigma_N > 0$. We note the $t' = (1 - \gamma)t$ where $\gamma \in [0, 1]$, so $t' \leq t$. This implies that $\sigma_{t'} > \sigma_{t+1}$ so:

$$k_t > 0 \qquad , \qquad m_t = \frac{\alpha_t}{\alpha_{t'}} (\sigma_{t'} - \frac{\sigma_{t+1}}{\alpha_{t+1}}) > 0 \tag{28}$$

This illustrates that $k_t, m_t \in \mathbb{R}^+$.