TELEQUANT: TEMPORAL NOISE SHAPING FOR POST-TRAINING QUANTIZED DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review



Figure 1: TeleQuant is a training-free temporal noise-shaping technique for quantized diffusion sampling. Wrapped around frozen ViDiT-Q and SVDQuant on PixArt-Σ and SDXL, it improves visual fidelity at $1.03\times$–$1.05\times$ end-to-end cost over the frozen host, with only a marginal rise in peak memory.

ABSTRACT

Post-training quantization lowers the cost of the repeated denoiser evaluations that diffusion Transformers use for image and video generation. Because those evaluations are chained, the rounding errors they leave behind interact across steps. Correlated rounding errors can reinforce, accumulate into structured trajectory drift, and persist in the returned sample. This drift can degrade image and video fidelity. We propose TeleQuant, a training-free temporal noise-shaping framework that feeds a discarded activation-rounding residual into the next evaluation, with a gain matched to adjacent solver responses and activation scales. We derive a first-order telescoping identity: under ideal state storage, accumulated error separates into a terminal residual and a sensitivity-variation term, motivating selective layer placement and damped feedback. INT8 residual state and a fused activation operator add the compensation on the host's path. The feedback lowers accumulated latent drift and improves paired fidelity to matched FP16 outputs. Evaluations on DiT-XL/2, PixArt-Σ, SDXL, and Open-Sora 1.2 show improved image fidelity and video consistency over frozen ViDiT-Q and SVDQuant hosts, with matched rounding ablations supporting the complete gain. The profiled image pipelines retain most of the host speedup, at $1.03\times$–$1.05\times$ latency. Code is available at https://anonymous.4open.science/r/TeleQuant-8217/.

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.

1 INTRODUCTION

Diffusion Transformers generate images and videos through repeated denoiser evaluations (Peebles & Xie, 2023). Fast solvers reduce their number (Song et al., 2021; Lu et al., 2025); post-training quantization reduces the cost of each evaluation. Recent methods balance channels, preserve temporal features, or isolate outliers in a high-precision branch (Wu et al., 2024; Huang et al., 2024a; Zhao et al., 2025; Li et al., 2025). Channel reordering and structured transforms further improve low-bit representations (Cheng et al., 2026; Lee et al., 2026; Bouquet et al., 2026). We call a quantized model together with its quantization rules the host.

Local reconstruction accuracy and trajectory-level error are separate concerns, because each perturbed prediction changes the inputs to later evaluations. Correlated errors can reinforce and produce persistent trajectory drift. Temporal calibration and precision allocation adapt the host to changing timestep statistics (Ding et al., 2025; Liu et al., 2026b; Zhang et al., 2026b; Lewandowski et al., 2026). Other work already addresses accumulation through compensated feature differences or solver-level correction (Gao et al., 2025; Liu et al., 2026a; Ding et al., 2026). Our focus is narrower: coordinating the rounding decisions of a frozen host according to their effect on the final sample.

We pursue this with TeleQuant (Telescoping Quantization). A residual discarded at one evaluation is fed into the next pre-rounding activation. Activation scales and the solver's remaining response change over time, so the reuse is matched to them with an analytical gain. Under a first-order surrogate, the error separates into a terminal residual and a sensitivity-variation term, tying the observed structure to a concrete rounding rule. The recurrence lowers accumulated latent drift during sampling and reduces trajectory error on the evaluated hosts.

The same analysis guides deployment. Calibration selects sites where sensitivity varies slowly and feedback improves measured error. Damping limits the injected perturbation; INT8 storage and fused activation quantization control its cost. The host's weights, precision policy, and matrix multiplication path remain fixed. Figure 1 shows matched-input samples from the profiled pipelines. Our main contributions are summarized as follows:

- We show how cross-step correlation among quantization errors contributes to trajectory drift, and characterize the low-frequency structure of the evaluated trajectories.
- We propose TeleQuant, a training-free temporal noise-shaping framework with solver-response-aware and scale-aware feedback. We derive a first-order telescoping identity and bound the remainder through sensitivity variation.
- We introduce a deployment pipeline that pairs calibration-guided site screening with INT8 residual state, built around a plug-and-play fused operator that carries temporal compensation inside the host's quantized path.
- Image and video evaluations show consistent fidelity recovery and improved temporal consistency over the tested frozen hosts. Controlled ablations examine the mechanism, with $1.03\times$–$1.05\times$ end-to-end latency relative to the frozen hosts on the profiled image pipelines.

2 RELATED WORK

Calibration and quantization representations. Early diffusion PTQ calibrates denoiser reconstruction and corrects quantization-induced distortion (Shang et al., 2023; Li et al., 2023; He et al., 2023). Subsequent work improves calibration coverage, distribution alignment, and temporal-feature reconstruction (Wang et al., 2024; Liu et al., 2024; Huang et al., 2024a). For Transformer denoisers, channel balancing and smoothing address activation outliers (Wu et al., 2024; Zhao et al., 2025; Dong & Zhang, 2025); low-rank decomposition, grouped quantization, and semantic segmentation further reduce local distortion (Li et al., 2025; Chen et al., 2025; Zhang et al., 2026a). Recent methods refine the representation through rotations, channel permutations, and structured transforms (Yang et al., 2026; Cheng et al., 2026; Lee et al., 2026; Bouquet et al., 2026). These approaches improve what is rounded. TeleQuant coordinates the residual sequence produced after the host is fixed.

Temporal calibration and precision allocation. Timestep grouping, variance equalization, and time-aware reconstruction adapt quantization to evolving activation distributions (Ding et al., 2025; Xu et al., 2025; Liu et al., 2026b). Mixed precision assigns bits according to layer or timestep sensitivity (Zhao et al., 2024; Li et al., 2026; Zhang et al., 2026b; Lewandowski et al., 2026; Han

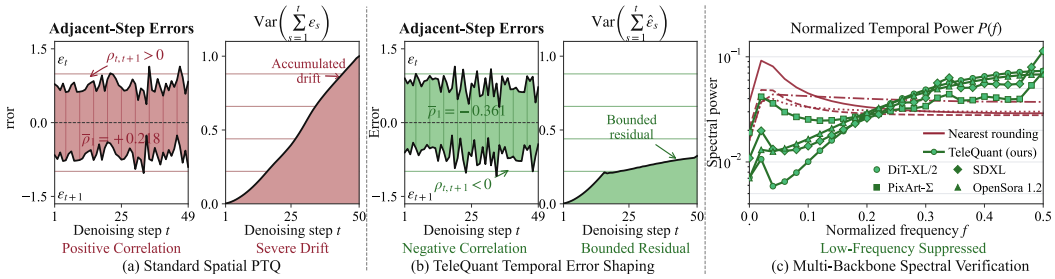


Figure 2: Empirical temporal error dynamics and multi-backbone spectral verification. (a) Coherent step errors under spatial PTQ accumulate into trajectory drift in the displayed runs. (b) TeleQuant reshapes lag correlation and reduces accumulated residual drift in the displayed runs. (c) Multi-backbone verification across four diffusion architectures, with suppressed low-frequency spectral power.

et al., 2026). These methods can alter temporal error structure through the host configuration. Our feedback instead acts on the realized rounding residual at each evaluation, retaining the host's precision policy.

Video quantization and temporal reuse. Video compression must also preserve information across frames. QVD maintains discriminative temporal features; Q-VDiT uses token-aware estimation and temporal distillation (Tian et al., 2024; Feng et al., 2025). QuantSparse combines attention distillation with temporal residual correction for sparse attention (Feng et al., 2026). Caching introduces another source of approximation: MoDiff quantizes compensated activation differences, Q&C calibrates and compensates joint quantization–cache error, and 6Bit-Diffusion combines mixed precision with temporal delta caching (Gao et al., 2025; Ding et al., 2026; Su et al., 2026). Thus temporal reuse and compensation are established ideas. Our state is a bounded rounding residual, and our gain matches its adjacent solver-weighted contributions without reconstructing a layer output from cached differences.

Trajectory compensation and noise shaping. Timestep-aware cumulative error compensation (TCEC) corrects accumulated error at the solver level (Liu et al., 2026a). TeleQuant acts before internal activation rounding and derives its gain from the remaining solver response and activation scale. The distinction is this matching rule and its telescoping identity, rather than residual feedback alone, which is classical (Schreier & Temes, 2005). Table 3 compares the two intervention points under matched hosts; Appendix F.5 examines compatibility with the methods above.

3 Preliminary Analysis

Reverse diffusion repeatedly evaluates a denoiser and updates a latent state with a numerical solver (Song et al., 2021; Lu et al., 2022; Zhao et al., 2023). We index evaluations in execution order, $t = 0, \dots, T$, with final state x_{T+1}. The solver consumes a prediction $f_\theta(x_t, t)$ expressed as noise, data, or velocity. An error inside one evaluation can therefore affect every later state. Our feedback follows this denoising axis; consistency across video frames is an evaluation outcome.

At layer $l \in \{1, \dots, L\}$, write the activation as $x_{t,l} = s_{t,l} z_{t,l}$ with step size $s_{t,l} > 0$. A uniform quantizer produces $q_{t,l} = \mathrm{clip}(\lfloor z_{t,l} \rceil, q_{\min}, q_{\max})$ on the host's integer grid. The normalized error is $\bar{e}_{t,l} = q_{t,l} - z_{t,l}$, and the physical perturbation is $e_{t,l} = s_{t,l} \bar{e}_{t,l}$. Calibration methods reduce local reconstruction error (Liu et al., 2024; Wu et al., 2024; Liu et al., 2026b; Zhang et al., 2026a), often using objectives such as $\sum_{t,l} \mathbb{E}\|e_{t,l}\|_2^2$. We use scalar scales for notation; Appendix A.4 gives the tokenwise generalization. To relate local distortion to the final sample, we must also account for propagation.

Linearizing the denoiser, an activation error changes its prediction by $J_{t,l} e_{t,l}$, where $J_{t,l} = \partial f_\theta(x_t, t)/\partial x_{t,l}$. We approximate the remaining solver transfer by a scalar response c_t, using the same prediction parameterization as $J_{t,l}$. Thus define $\delta_l^{(1)} = \sum_{t=0}^{T} c_t J_{t,l} e_{t,l}$ and $\delta^{(1)} = \sum_l \delta_l^{(1)}$. This tractable surrogate separates the schedule response from local sensitivity; Appendix A.1 derives its coefficients and bounds its difference from the full matrix-valued transfer.

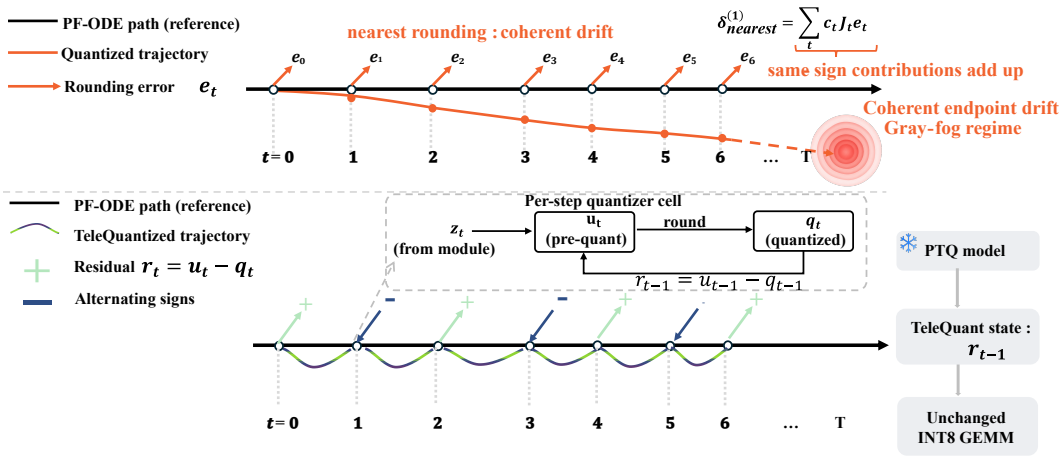


Figure 3: Temporal feedback within the frozen host. **Top:** conventional nearest rounding leaves same-sign residuals, whose solver-weighted contributions accumulate into the endpoint displacement $\delta^{(1)}_{\text{nearest}}$, the *coherent endpoint drift*, whose low-frequency, low-contrast profile defines the *gray-fog regime*. **Bottom:** TeleQuant feeds each discarded residual into the next pre-rounding activation, so consecutive residuals alternate in sign and cancel; integer codes enter the unchanged host GEMM.

Expanding the layerwise second moment reveals the missing temporal interactions:

$$\sum_{l=1}^{L} \mathbb{E}\left\|\delta_l^{(1)}\right\|_2^2 = \sum_{l=1}^{L}\sum_{t=0}^{T} \mathbb{E}\left\|c_t J_{t,l} e_{t,l}\right\|_2^2 + 2\sum_{l=1}^{L}\sum_{0\le t<u\le T} \mathbb{E}\langle c_t J_{t,l} e_{t,l},\, c_u J_{u,l} e_{u,l}\rangle. \tag{1}$$

The first term measures individual propagated errors; the second measures their reinforcement or cancellation across evaluations. The latter is governed by the cross-step interaction of those errors. These are uncentered second moments, and the full error additionally contains cross-layer terms. Figure 2 shows coherent drift and low-frequency error structure on the evaluated trajectories. The cross-step term in Equation 1 therefore suggests coordinating consecutive rounding errors at each site, before they accumulate.

4 TeleQuant: Sampler-Shaped Temporal Rounding

TeleQuant realizes this coordination by keeping a discarded residual and reweighting it at the next evaluation, with a gain matched to adjacent contributions. The remaining sensitivity and state errors motivate three choices: a damped gain at the selected sites (Section 4.1), INT8 storage with boundary resets (Section 4.2), and offline calibration for placement (Section 4.3).

4.1 Sampler-Shaped Residual Feedback

At each selected projection $l \in \mathcal{L}_{\text{shape}}$ and each classifier-free guidance stream, we initialize a residual buffer to $r_{-1,l} = 0$. Its stored approximation $\widehat{r}_{t-1,l}$ shifts the next pre-rounding activation:

$$u_{t,l} = z_{t,l} + \alpha_{t,l}\widehat{r}_{t-1,l}, \qquad q_{t,l} = \text{clip}(\lfloor u_{t,l} \rceil, q_{\min}, q_{\max})\,. \tag{2}$$

To determine $\alpha_{t,l}$, first consider exact storage and unsaturated rounding. The discarded residual $r_{t,l} = u_{t,l} - q_{t,l}$ gives $\bar{e}_{t,l} = \alpha_{t,l} r_{t-1,l} - r_{t,l}$. Define $a_{t,l} = c_t s_{t,l}$. The incoming contribution is then $a_{t,l}\alpha_{t,l} J_{t,l} r_{t-1,l}$, whereas the preceding outgoing contribution was $-a_{t-1,l} J_{t-1,l} r_{t-1,l}$. Their Jacobians need not coincide, but their scalar weights can be matched. With damping $\lambda_{t,l} \in (0, 1]$, this requires $a_{t,l}\alpha_{t,l} = \lambda_{t,l} a_{t-1,l}$. We therefore use

$$\alpha_{t,l} = \lambda_{t,l}\frac{c_{t-1}s_{t-1,l}}{c_t s_{t,l}}, \qquad r_{t,l} = \text{clip}(u_{t,l} - q_{t,l}, -0.5, 0.5)\,. \tag{3}$$

Appendix A.2 proves uniqueness for this matching target when $a_{t,l}$ and $J_{t,l}$ are nonzero. The ratio corrects changes in solver response and quantization scale; λ controls feedback strength. We initialize $\alpha_{0,l} = 0$. Figure 3 shows the host integration, and Section 4.2 specifies response-boundary resets.

Matching weights leaves the difference between consecutive Jacobians. The following theorem exposes this remainder and includes sites without feedback.

Theorem 1 (Trajectory drift under selective feedback)**.** *Assume exact residual storage, no saturation or resets, and $a_{t,l} \neq 0$ at feedback steps. Partition the sites into $\mathcal{L}_{\text{shape}}$ and $\mathcal{L}_{\text{static}}$. Shaped sites use the rounding rule in Equation 2 with the gain in Equation 3, starting from $r_{-1,l} = 0$ and $\|r_{t,l}\|_\infty \leq \rho \leq 0.5$. Let $\delta^{(1)}_{\text{static}} = \sum_{l \in \mathcal{L}_{\text{static}}} \sum_{t=0}^{T} a_{t,l} J_{t,l} \bar{e}^{\text{static}}_{t,l}$ and $\Delta_\lambda J_{t,l} = \lambda_{t+1,l} J_{t+1,l} - J_{t,l}$. Then the first-order surrogate satisfies*

$$\delta^{(1)} = \sum_{l \in \mathcal{L}_{\text{shape}}} \left[\sum_{t=0}^{T-1} a_{t,l} \Delta_\lambda J_{t,l} r_{t,l} - a_{T,l} J_{T,l} r_{T,l} \right] + \delta^{(1)}_{\text{static}}, \tag{4}$$

and consequently

$$\left\| \delta^{(1)} \right\|_2 \leq \rho \sum_{l \in \mathcal{L}_{\text{shape}}} \left(|a_{T,l}| \, \|J_{T,l}\|_{\infty \to 2} + \sum_{t=0}^{T-1} |a_{t,l}| \, \|\Delta_\lambda J_{t,l}\|_{\infty \to 2} \right) + \left\| \delta^{(1)}_{\text{static}} \right\|_2 . \tag{5}$$

In Equation 4, each residual enters adjacent evaluations with matched scalar weights. With constant Jacobians and $\lambda = 1$, the interior terms cancel and only the terminal residual remains. Damping reduces the pre-rounding shift at the cost of incomplete cancellation. State compression and nonlinear propagation enter through the bounds in Appendix A.4 and Appendix A.5. We use $\lambda = 0.75$, the best of the three fixed settings in Table 8; once sensitivities change, the bound in Equation 5 is minimized away from $\lambda = 1$, which Appendix E.2 analyzes. Appendix A.3 proves the identity.

4.2 State Compression and Boundary Handling

The recurrence adds a state read and write at each selected site. The half-bin bound follows from unsaturated nearest rounding and permits a fixed symmetric INT8 representation. Table 7 tests its fidelity against FP16 and INT4 state. We fuse state dequantization, feedback, and update with activation quantization in a Triton frontend (Tillet et al., 2019); integer codes pass to the unchanged host GEMM. Appendix A.4 bounds the storage error.

Boundary handling prevents an invalid residual from being carried forward. Saturated rows clear outgoing state because clipping error can exceed half a bin. Zero solver responses disable the undefined gain; sign changes reset state as a conservative deployment policy. Signed nonzero responses remain valid in the identity. The current error is retained in the correction of Appendix A.4, and Algorithm 1 specifies the update order.

4.3 Calibration-Guided Site Selection

Projections differ in how suitable they are for feedback, and reliable state is the first requirement. Theorem 1 favors slowly varying sensitivities, while the endpoint curvature bound in Appendix A.5 shows how nonlinear propagation can offset a first-order gain. We therefore screen sensitivity variation first, then test realized local error on paired calibration trajectories. The accepted sites remain fixed during generation.

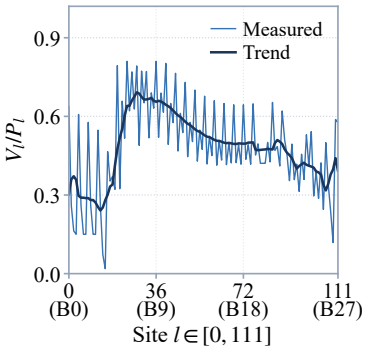


Figure 4: Jacobian variation atlas. Smoothness ratio V_l/P_l of Equation 6 over the 112 screened sites.

The first screen measures whether adjacent sensitivities vary slowly relative to their overall magnitude. We use an undamped norm proxy, estimated with Jacobian-vector products, for the theorem's damped induced-norm expression. Figure 4 shows why selection depends on measured site behavior. The scale $s_{t,l}$ enters both numerator and denominator on the probe, in the coordinates of the activation that the site rounds, which is the space its scale and Equation 2 are defined in; normalization makes the ratio comparable across sites. With the probe v_l of Appendix D.4, a candidate passes when

$$V_l/P_l \triangleq \frac{\sum_{t=0}^{T-1} |c_t| \, \|(J_{t+1,l} - J_{t,l})(s_{t,l} \odot v_l)\|_2}{\sum_{t=0}^{T} |c_t| \, \|J_{t,l}(s_{t,l} \odot v_l)\|_2} < \tau_{\text{tv}}. \tag{6}$$

The variation ratio screens sensitivity variation alone, so the second screen checks actual local errors in paired frozen and shaped rollouts, covering clipping, compressed state, and nonlinear propagation:

$$\Delta\mathcal{E}_l \triangleq \sum_{t=0}^{T} |c_t| \left\| e_{t,l}^{\text{frozen}} \right\|_2^2 - \sum_{t=0}^{T} |c_t| \left\| e_{t,l}^{\text{TeleQuant}} \right\|_2^2. \tag{7}$$

We retain candidates with $\Delta\mathcal{E}_l > 0$ and evaluate their final mask jointly, since individually favorable sites can interact. Appendix D.4 gives the probe, threshold, and calibration procedure. The single power of $|c_t|$ weights local error by response magnitude, so the score ranks sites by response-weighted local error. The sufficient endpoint condition in Appendix A.5 complements this screen. Appendix F.2 examines the selected sites.

5 Experiments

Section 5.1 specifies the setup: the image and video backbones, their evaluation datasets, the quantized hosts and matched FP16 references, the metrics, and the sampling configuration. Section 5.2 reports results on image generation, video generation, trajectory-level compensation, rounding and state ablations, and latency and memory.

5.1 Experimental Setup

Models. We evaluate DiT-XL/2 for class-conditional images (Peebles & Xie, 2023), PixArt-Σ and SDXL for text-to-image synthesis (Chen et al., 2024; Podell et al., 2024), and Open-Sora 1.2 for video generation (Zheng et al., 2024). Appendix B specifies their roles.

Datasets. The corresponding evaluation sets are 50,000 ImageNet samples, ViDiT-Q's 1,024-prompt `coco_1024` subset, SVDQuant's 5,000-prompt MJHQ subset, and the 251-prompt VBench composition. Host and feedback arms share prompts or class labels, seeds, and solver settings. Comparisons are within each backbone; Appendix C gives the manifests.

Baselines. We add TeleQuant to frozen ViDiT-Q W4A8 and SVDQuant W4A4 hosts (Zhao et al., 2025; Li et al., 2025). Matched FP16 outputs provide trajectory references, and ViDiT-Q W8A8 is a higher-precision video reference. Independent stochastic rounding tests memoryless randomization; unit-delay feedback tests residual reuse without the derived gain. TCEC tests solver-level compensation. Other cited methods motivate the comparison; Appendix D.2 specifies the evaluated arms.

Metrics. FID, sFID, and Inception Score assess distributional image quality. PSNR, SSIM (Wang et al., 2004), and AlexNet-LPIPS (Zhang et al., 2018) measure fidelity to matched FP16 outputs. Video evaluation adds six VBench dimensions, including overall video–text consistency (Huang et al., 2024b). The two views are complementary: the distributional metrics assess sample quality and paired fidelity measures agreement with the matched FP16 reference. The DiT distributional FP16 row is a published reference; paired fidelity uses matched generated outputs. Appendix D.5 specifies metric and uncertainty protocols.

Implementation details. We retain each backbone's schedule: 50-step DDIM for DiT-XL/2, 20-step DPM-Solver++ for PixArt-Σ, 50-step Euler-EDM for SDXL, and 100-step rectified flow for 16-frame Open-Sora 1.2 clips. Appendix D.1 gives resolutions and guidance scales. Feedback uses global damping $\lambda = 0.75$ and INT8 state. A mask calibrated on 32 instances is fixed during generation. Guidance streams keep separate residual state, initialized per sample. Arms share initial noise and conditioning, although subsequent trajectories can diverge as rounding changes later inputs. Appendix D specifies host integration, calibration, and profiling.

5.2 Results

Image generation. Table 1 reports gains in both distributional quality and paired fidelity across all three image backbones. The host calibration remains frozen, so the added effect is temporal feedback alone. All three paired metrics move in the same direction on every backbone. The recovered fraction of the remaining FP16 gap differs by backbone, with the two text-to-image hosts closing more of their distributional gap than the class-conditional DiT (Appendix F.3). Figure 5 shows the

Table 1: Image generation with ViDiT-Q and SVDQuant hosts, with and without TeleQuant. W4A8 and W4A4 denote 4-bit weights with 8-bit and 4-bit activations; 16/16 is unquantized FP16.

Model	Quantization Host	W/A	FID ↓	sFID ↓	IS ↑	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
DiT-XL/2	*FP16 Reference*	16/16	2.27	4.60	278.24	—	—	—
DiT-XL/2	ViDiT-Q Host	W4A8	5.32	6.48	248.50	19.59	0.680	0.286
	+ TeleQuant (Ours)	W4A8	**4.65**	**5.82**	**254.10**	**19.91**	**0.708**	**0.262**
PixArt-Σ	*FP16 Reference*	16/16	7.24	7.65	238.50	—	—	—
PixArt-Σ	ViDiT-Q Host	W4A8	8.42	9.15	224.60	15.25	0.590	0.448
	+ TeleQuant (Ours)	W4A8	**7.55**	**8.48**	**229.80**	**16.88**	**0.611**	**0.400**
SDXL	*FP16 Reference*	16/16	12.54	9.42	214.60	—	—	—
SDXL	SVDQuant Host	W4A4	14.86	11.70	195.42	21.02	0.745	0.288
	+ TeleQuant (Ours)	W4A4	**13.95**	**11.02**	**200.18**	**22.46**	**0.808**	**0.264**

Table 2: Text-to-video generation on VBench under the frozen ViDiT-Q host; higher is better.

Method	W/A	Imaging	Motion	Background	Subject	Human	Overall
FP16 Reference	16/16	63.68	96.28	96.13	90.28	88.50	26.21
ViDiT-Q Baseline	W8A8	63.48	96.14	95.84	90.24	88.10	26.06
ViDiT-Q Host	W4A8	61.07	95.69	95.23	88.72	85.40	25.94
ViDiT-Q + TeleQuant	W4A8	**62.42**	**96.15**	**96.10**	**90.15**	**87.12**	**26.05**

corresponding improvements in local contrast, facial detail, and background shading; Appendix E.5 provides larger examples. Entries are point estimates; Appendix E.4 reports paired bootstrap intervals for the additional arms.

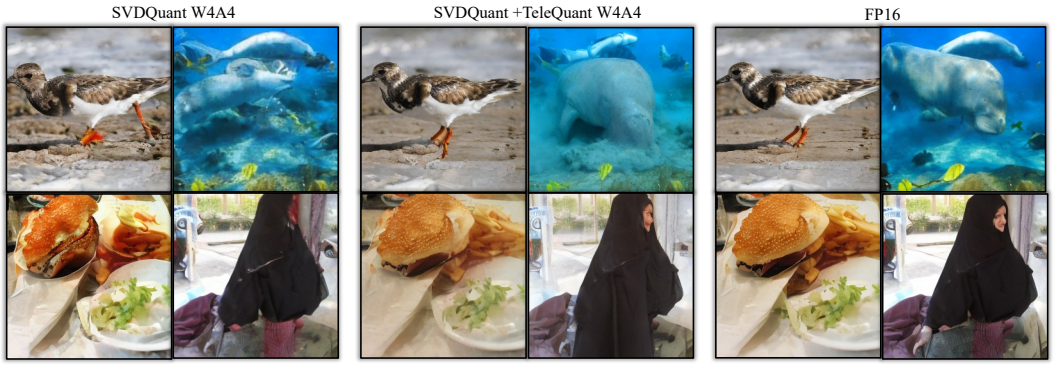
Video generation. Table 2 shows gains in all six dimensions over the frozen W4A8 host: imaging quality, motion smoothness, background consistency, subject consistency, human action, and overall consistency. Background and subject consistency approach the FP16 reference, while imaging quality and human action remain further from it. Against the higher-precision W8A8 baseline, feedback on the 4-bit host exceeds motion smoothness and background consistency, and stays within a tenth of it on the remaining two consistency dimensions; on imaging quality and human action it narrows the gap to that baseline. The recovered fractions separate along the same lines: motion smoothness, background consistency, and subject consistency recover most of their remaining gap, whereas imaging quality, human action, and overall video–text consistency recover less (Table 14). Since the mechanism acts on the rounding-error sequence rather than on one frame, this ordering is consistent with the cross-step diagnostics in Figure 2. The skyline sequence in Figure 6 illustrates better preserved structure at matched temporal indices. Because VBench assesses perceptual and semantic properties, we separately test fidelity to FP16 frames and their temporal differences in Appendix E.1, where both measurements improve while aggregate spatial fidelity is preserved.

Table 3: Trajectory-fidelity comparison with TCEC under the same frozen host.

	DiT-XL/2			PixArt-Σ		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ViDiT-Q	19.59	0.680	0.286	15.25	0.590	0.448
+ TCEC	19.74	0.689	0.279	15.68	0.601	0.431
+ TeleQuant	**19.91**	**0.708**	**0.262**	**16.88**	**0.611**	**0.400**

Trajectory-level compensation. Table 3 compares TCEC (Liu et al., 2026a) with TeleQuant under matched ViDiT-Q hosts and sampling conditions. Feedback achieves higher PSNR and SSIM and lower LPIPS on both backbones. This comparison distinguishes correcting a denoiser's prediction from changing the internal rounding decisions that produce it. The latter also alters propagation through the remaining denoiser operations.

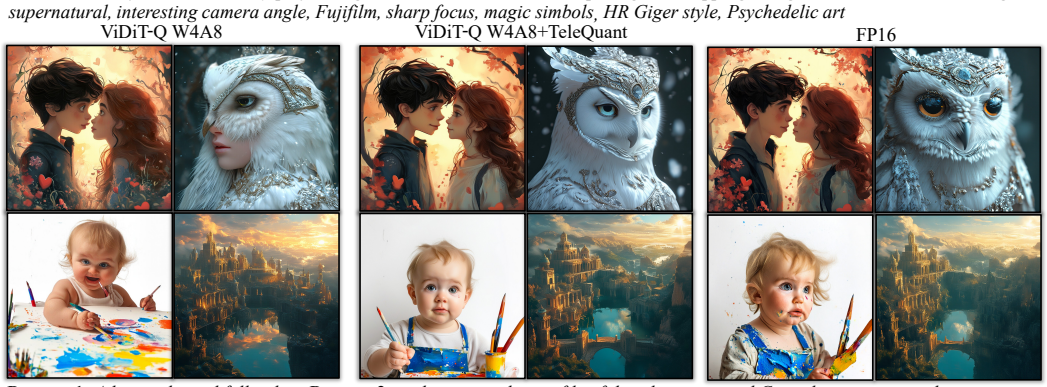
Rounding and state ablations. Figure 7 tests whether storing a residual is sufficient. Both feedback arms use the same INT8 state and frozen ViDiT-Q host. Unit-delay reuse gives a smaller gain, while TeleQuant gives the best joint PSNR–LPIPS result. The matched gain targets the mismatch that unweighted reuse leaves, derived in Appendix A.2. The three rows therefore test the deployed system, in which damping, solver response, and scale matching act together. Stochastic rounding also degrades paired reconstruction, showing that pointwise unbiasedness alone is insufficient after



Caption 1: *A bird standing on some rocks near the water.* **Caption 2**: *A number of large whales swimming under the ocean surface.* **Caption 3**: *A hamburger and fries sit on top of paper.* **Caption 4**: *A person standing outside with their back facing the camera.*



Prompt 1: *morgan freeman headshot, hyperrealistic, 4k, colour graded, wearing old shashank redemption hat, looking at camera* **Prompt 2**: *beautiful photography advertising, instagramad image for engagement ring, engagement ring photography, neutral background* **Prompt 3**: *portrait of a cute girl in a fairy forest from the movie avatar in cartoon style* **Prompt 4**: *Photo of a horror movie scene of a secret society performing ritual, cinematic, dramatic, pentagram, dripping art, glam art, aliens, ambient lights, supernatural, interesting camera angle, Fujifilm, sharp focus, magic simbols, HR Giger style, Psychedelic art*



Prompt 1: *A boy and a girl fall in love* **Prompt 2**: *anthropomorphic profile of the white snow owl Crystal priestess , art deco painting, pretty and expressive eyes, ornate costume, mythical, ethereal, intricate, elaborate, hyperrealism, hyper detailed, 3D, 8K, Ultra Realistic, high octane, ultra resolution, amazing detail, perfection, In frame, photorealistic, cinematic lighting, visual clarity, shading , Lumen Reflections, Super-Resolution, gigapixel, color grading, retouch, enhanced, PBR, Blender, V-ray, Procreate, zBrush, Unreal Engine 5, cinematic, volumetric, dramatic, neon lighting, wide angle lens ,no digital painting blur* **Prompt 3**: *Bright scene, aerial view, ancient city, fantasy, gorgeous light, mirror reflection, high detail, wide angle lens* **Prompt 4**: *A baby painter trying to draw very simple picture, white background*

Figure 5: Image generation with matched latents and prompts. SDXL uses SVDQuant W4A4 with 50-step Euler-EDM at 1024×1024; PixArt-Σ uses ViDiT-Q W4A8 with 20-step DPM-Solver++ at 512×512. Feedback preserves the host configuration in each comparison.

propagation: a memoryless rule removes the deterministic residual structure that nearest rounding leaves for feedback to coordinate. Nearest rounding minimizes local squared error on a fixed grid; feedback instead coordinates the sequence encountered during sampling. Appendix F.1 traces the resulting drift before decoding. Across the ablation, unweighted reuse recovers about a third of the PSNR gap to the complete design, while stochastic rounding falls 9.2 dB below nearest rounding on the same grid.



Prompt: *A cinematic city skyline at dusk viewed across a river, with illuminated skyscrapers and reflections on the water; the camera pans smoothly along the waterfront while the buildings remain stable anddetailed*

Figure 6: Video generation with matched latents, prompts, and frame indices. Open-Sora 1.2 uses the frozen ViDiT-Q W4A8 host with 100-step rectified flow over 16 frames and guidance scale 4.0.

Rounding	W/A	Residual	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	Role
FP16 reference	16/16	None	—	—	—	Unquantized
Nearest (RTN)	W4A8	None	19.59	0.680	0.286	Frozen PTQ
Stochastic	W4A8	None	10.39	0.212	0.686	Memoryless
Unit-delay, $\alpha = 1$	W4A8	INT8	19.69	0.696	0.282	Unweighted
TeleQuant (ours)	W4A8	**INT8**	**19.91**	**0.708**	**0.262**	**Telescoping**



Figure 7: Rounding-mechanism ablation on DiT-XL/2 under the frozen ViDiT-Q host. Top: paired fidelity per mechanism; bottom: matched samples. *Unit feedback* is unit-delay reuse at $\alpha = 1$; *sampler-shaped* is TeleQuant.

The operating-point studies test accuracy against added cost. Table 7 shows that INT8 closely tracks FP16-state fidelity at half the buffer storage, whereas INT4 underperforms the host. Table 8 supports the damped default over the two tested endpoints. Finer damping in Table 10 yields additional gains: offline per-layer adaptation is inexpensive at inference, while per-step adaptation incurs repeated estimation cost. The global scalar keeps the main comparisons free of that extra adaptation procedure. These controlled sweeps use DiT-XL/2; the other backbones evaluate the transferred configuration.

Latency and memory. TeleQuant retains most of the frozen host's acceleration over FP16. End-to-end latency is $1.03\times$–$1.05\times$ that of the host on the profiled image pipelines, and peak allocation grows only marginally. Pipeline peaks include temporary allocations and therefore differ from buffer-only storage. Fusing residual dequantization, addition, and requantization avoids a separate pass over activations, although state transfer and scheduling still add cost. Appendix E.6 reports operator latency separately from complete pipeline time, which includes text encoding, repeated denoising, and decoding.

6 Conclusion

TeleQuant coordinates activation rounding across diffusion evaluations through solver- and scale-aware feedback. Its first-order identity exposes a terminal residual and damped sensitivity variation, motivating calibration-guided placement and fused INT8 state handling. Image and video evaluations show improved fidelity within frozen quantizers at modest overhead on the profiled pipelines. The scalar-response approximation and empirical site selection remain limitations; extending the analysis to additional solvers and precision schedules is future work.

Reproducibility Statement

Appendix A provides proofs and analytical solver responses. Appendix B specifies the evaluation backbones and Appendix C the manifests. Appendix D details sampling, host integration, calibration, state management, and profiling. Paired comparisons share checkpoints, guidance schedules, and seeds. Code and evaluation utilities are available at https://anonymous.4open.science/r/TeleQuant-8217/. Additional measurements are reported in Appendix E, and mechanism diagnostics in Appendix F.

AI Use Statement

We used generative AI tools for three purposes in this work. First, AI tools assisted with editing the manuscript text for readability, including grammar, phrasing, and improving transitions between sections. Second, AI tools assisted with retrieval and discovery of related work, helping us locate and summarize candidate papers in the post-training quantization and diffusion sampling literature, which we independently verified against their original sources before citation. Third, AI tools assisted with research execution in two specific ways: (i) downloading and reproducing the released code of prior baseline methods (ViDiT-Q (Zhao et al., 2025), SVDQuant (Li et al., 2025), etc.) to establish working reference implementations for comparison, and (ii) helping us analyze and propose more reasonable experimental designs, such as candidate ablation structures and comparison protocols, which the authors then evaluated and adopted or rejected. These fall under the required-disclosure categories of "implement methods" and "design or provide feedback on research methodology or experiments."

We have not used generative AI tools for any other aspect of this work. In particular, we did not use AI to formulate the theoretical claims or derive the proofs (including Theorem 1 and its proof in Appendix A), to decide experimental configurations (datasets, sampling schedules, guidance scales, and solver settings; Appendix C and Table 4), to perform hyperparameter tuning (including the damping coefficient sweep in Table 8 and the bit-width and residual-precision sweeps in Table 6 and Table 7), to execute the experiments or collect measurements, to generate synthetic data, or to interpret or draw conclusions from the experimental results reported in Section 5. All experimental configurations, hyperparameter choices, and result interpretations are the authors' own decisions and analysis.

All AI-assisted work was actively verified by the authors before inclusion in this paper. Reproduced baseline code was checked against the original repositories and papers, and its outputs were validated against the numbers those works report for the backbones they evaluate, before being used as a reference in our comparisons (Tables 1 to 3). Experimental designs suggested by AI tools were independently reviewed by the authors, who determined their validity, revised them where necessary, and ran all resulting experiments and measurements themselves. AI-suggested text edits were checked against the original technical meaning by all authors before acceptance. We take full responsibility for the accuracy and integrity of the derivations, experimental results, and final manuscript.

References

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yann Bouquet, Alireza Khodamoradi, Kristof Denolf, and Mathieu Salzmann. KroQuant: Kronecker-structured block transforms for efficient post-training quantization of diffusion transformers. arXiv preprint arXiv:2607.21446, 2026.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *European Conference on Computer Vision (ECCV)*, 2024.

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.

Yongsen Cheng, Kai Liu, Kaiwen Tao, Junxian Li, Zhixin Wang, Zhikai Chen, Renjing Pei, and Yulun Zhang. PermuQuant: Lowering per-group quantization error by reordering channels for diffusion models. arXiv preprint arXiv:2605.09503, 2026.

Jia Deng, Wei Dong, Richard Socher, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

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

Ning Ding, Jing Han, Yuchuan Tian, Chao Xu, Kai Han, and Yehui Tang. Post-training quantization for diffusion transformer via hierarchical timestep grouping. arXiv preprint arXiv:2503.06930, 2025.

Xin Ding, Xin Li, Haotong Qin, and Zhibo Chen. Q&C: When quantization meets cache in efficient generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhenyuan Dong and Sai Qian Zhang. DiTAS: Quantizing diffusion transformers via enhanced activation smoothing. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4606–4615, 2025.

Weilun Feng, Chuanguang Yang, Haotong Qin, Xiangqi Li, Yu Wang, Zhulin An, Libo Huang, Boyu Diao, Zixiang Zhao, Yongjun Xu, and Michele Magno. Q-VDiT: Towards accurate quantization and distillation of video-generation diffusion transformers. In *International Conference on Machine Learning (ICML)*, pp. 16956–16976, 2025.

Weilun Feng, Chuanguang Yang, Haotong Qin, Mingqiang Wu, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Yulun Zhang, Michele Magno, and Yongjun Xu. QuantSparse: Comprehensively compressing video diffusion transformer with model quantization and attention sparsification. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=4TAG3aQljJ.

Robert W. Floyd and Louis Steinberg. An adaptive algorithm for spatial greyscale. *Proceedings of the Society for Information Display*, 17(2):75–77, 1976.

Weizhi Gao, Zhichao Hou, Junqi Yin, Feiyi Wang, Linyu Peng, and Xiaorui Liu. Modulated diffusion: Accelerating generative modeling with modulated quantization. In *International Conference on Machine Learning (ICML)*, pp. 18337–18362, 2025.

Seokho Han, Dongwei Wang, Jinhee Kim, Yiran Chen, Kang Eun Jeon, Huanrui Yang, and Jong Hwan Ko. TASQ: Temporal-adaptive bit sparsification quantization for diffusion models. arXiv preprint arXiv:2608.03057, 2026.

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 (NeurIPS)*, 2023.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. EfficientDM: Efficient quantization-aware fine-tuning of low-bit diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. TFMQ-DM: Temporal feature maintenance quantization for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7362–7371, 2024a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Donghyun Lee, Jitesh Chavan, Duy Nguyen, Sam Huang, Liming Jiang, Priyadarshini Panda, Timo Mertens, and Saurabh Shukla. OrbitQuant: Data-agnostic quantization for image and video diffusion transformers. arXiv preprint arXiv:2607.02461, 2026.

Basile Lewandowski, Simon Kurz, Aditya Shankar, Robert Birke, Jian-Jia Chen, and Lydia Y. Chen. TMPDiff: Temporal mixed-precision for diffusion models. arXiv preprint arXiv:2603.14062, 2026.

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 component for 4-bit diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-Diffusion: Quantizing diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Zhiteng Li, Hanxuan Li, Junyi Wu, Kai Liu, Haotong Qin, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang. DVD-Quant: Data-free video diffusion transformers quantization. In *International Conference on Learning Representations (ICLR)*, 2026.

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 *European Conference on Computer Vision (ECCV)*, 2014.

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

Songwei Liu, Chao Zeng, Chenqian Yan, Xurui Peng, Xing Wang, Fangmin Chen, and Xing Mei. Error propagation mechanisms and compensation strategies for quantized diffusion models. In *International Conference on Machine Learning (ICML)*, 2026a.

Xinyan Liu, Huihong Shi, Yang Xu, and Zhongfeng Wang. TaQ-DiT: Time-aware quantization for diffusion transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 36: 6225–6229, 2026b. doi: 10.1109/TCSVT.2026.3652275.

Xuewen Liu, Zhikai Li, Junrui Xiao, Mengjuan Chen, Jianquan Li, and Qingyi Gu. EDA-DM: Enhanced distribution alignment for post-training quantization of diffusion models. arXiv preprint arXiv:2401.04585, 2024.

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 around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025. doi: 10.1007/s11633-025-1562-4.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *International Conference on Machine Learning (ICML)*, 2020.

Steven R. Norsworthy, Richard Schreier, and Gabor C. Temes (eds.). *Delta-Sigma Data Converters: Theory, Design, and Simulation*. IEEE Press, 1997.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 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. In *International Conference on Learning Representations (ICLR)*, 2024.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Richard Schreier and Gabor C. Temes. *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2005.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Rundong Su, Jintao Zhang, Zhihang Yuan, Haojie Duanmu, Jianfei Chen, and Jun Zhu. 6Bit-Diffusion: Inference-time mixed-precision quantization for video diffusion models. arXiv preprint arXiv:2603.18742, 2026.

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.

Philippe Tillet, H. T. Kung, and David Cox. Triton: An intermediate language and compiler for tiled neural network computations. In *ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)*, pp. 10–19, 2019. doi: 10.1145/3315508.3329973.

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

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *International Conference on Learning Representations (ICLR)*, 2022.

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.

Qinkai Xu, Yijin Liu, Yang Chen, Lin F. Yang, Li Erran Li, and Yuxiang Fu. VETA-DiT: Variance-equalized and temporally adaptive quantization for efficient 4-bit diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.52202/085713-5113. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e067d57f9f113dc49dadcfefce1738d6-Abstract-Conference.html.

Lianwei Yang, Haokun Lin, Tianchen Zhao, Yichen Wu, Hongyu Zhu, Ruiqi Xie, Zhenan Sun, Yu Wang, and Qingyi Gu. LRQ-DiT: Log-rotation post-training quantization of diffusion transformers for image and video generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026. doi: 10.1109/TCSVT.2026.3723090.

Jiaji Zhang, Ruichao Sun, Hailiang Zhao, Jiaju Wu, Peng Chen, Hao Li, Yuying Liu, Kingsum Chow, Gang Xiong, and Shuiguang Deng. SegQuant: A semantics-aware and generalizable quantization framework for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43717–43726, 2026a. URL https://optisys-zju.github.io/segquant/.

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

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shaoqiu Zhang, Zizhong Ding, Kaicheng Yang, Junyi Wu, Xianglong Yan, Xi Li, Bingnan Duan, Jianping Fang, and Yulun Zhang. AdaTSQ: Pushing the Pareto frontier of diffusion transformers via temporal-sensitivity quantization. arXiv preprint arXiv:2602.09883, 2026b.

Xun Zhang, Kaicheng Yang, Hongliang Lu, Haotong Qin, Yong Guo, and Yulun Zhang. Q-DiT4SR: Exploration of detail-preserving diffusion transformer quantization for real-world image super-resolution. In *International Conference on Machine Learning (ICML)*, 2026c.

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.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Enshu Liu, Rui Wan, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. ViDiT-Q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404, 2024.

A PROOFS

Theorem 1 describes cancellation in a scalar-response, first-order surrogate. We first derive this surrogate from the sampler, then obtain the matching gain and prove the cancellation identity. State compression, resets, and nonlinear endpoint error are derived in the following sections. Throughout, $t = 0, \dots, T$ indexes denoiser evaluations in execution order, as defined in Section 3; we suppress the layer index for single-site arguments.

A.1 TERMINAL RESPONSE OF DIFFUSION SOLVERS

Exact linearization and the scalar surrogate. For a differentiable sampler $x_{t+1} = G_t(x_t, f_\theta(x_t, t))$, let A_t be its total state derivative and $B_t = \partial G_t / \partial f_t$. The first-order state perturbation satisfies

$$\begin{aligned} d_{t+1} &= A_t d_t + B_t \eta_t, \qquad \eta_t = \sum_l J_{t,l} e_{t,l}, \qquad d_0 = 0, \\ d_{T+1} &= \sum_{t=0}^{T} H_t \eta_t, \qquad H_t = A_T \cdots A_{t+1} B_t. \end{aligned} \tag{8}$$

The product H_t records how a perturbation reaches the final state, including its effect on subsequent predictions. A multistep method uses an augmented state and a terminal projection. Replacing its matrix-valued transfer H_t by $c_t I$ gives the surrogate in Section 3 and the explicit approximation bound

$$\begin{aligned} \delta^{(1)} &= \sum_{t,l} c_t J_{t,l} e_{t,l}, \\ d_{T+1} - \delta^{(1)} &= \sum_{t,l} (H_t - c_t I) J_{t,l} e_{t,l}, \\ \|d_{T+1} - \delta^{(1)}\|_2 &\le \sum_{t,l} \|H_t - c_t I\|_{2\to 2} \, \|J_{t,l} e_{t,l}\|_2. \end{aligned} \tag{9}$$

The scalar approximation makes an analytical feedback gain possible without carrying a full transfer matrix during generation. The theorem applies exactly to this surrogate; the displayed remainder quantifies the approximation to the full transfer. Equation 1 expands its layerwise second moments, while the full second moment also contains cross-layer terms. Weight quantization remains in the frozen host.

Schedule-only responses. Use descending solver indices $k = K, \dots, 1$, where $K = T + 1$ and $k = K - t$. With future solver-consumed predictions held fixed,

$$x_{k-1} = A_k x_k + B_k f_k \quad \Longrightarrow \quad c_k = B_k \prod_{j=1}^{k-1} A_j, \qquad \prod_{j=1}^{0} A_j = 1. \tag{10}$$

These A_k, B_k are scalar schedule coefficients, distinct from the total derivatives above. For deterministic DDIM with $f_k = \epsilon_k$ and positive schedule denominators,

$$A_k = \sqrt{\frac{\bar\alpha_{k-1}}{\bar\alpha_k}}, \qquad B_k = \sqrt{1 - \bar\alpha_{k-1}} - A_k \sqrt{1 - \bar\alpha_k}. \tag{11}$$

Euler has $c_k^v = \tau_{k-1} - \tau_k$ when its consumed prediction is velocity. Euler-EDM written in denoised-data form instead gives

$$v_k = \frac{x_k - D_k}{\sigma_k}, \qquad A_k^D = \frac{\sigma_{k-1}}{\sigma_k}, \qquad B_k^D = 1 - A_k^D, \qquad c_k^D = B_k^D \prod_{j=1}^{k-1} A_j^D. \tag{12}$$

The Jacobian must use the same prediction parameterization as c_k. A final data-form step to $\sigma_0 = 0$ gives zero response for earlier frozen predictions, for which feedback is disabled. The deployed SDXL sampler instead consumes the noise prediction under the σ-scaled Euler update of Table 4, so its response is the corresponding step size $c_k^\epsilon = \sigma_{k-1} - \sigma_k$, the same $\Delta\sigma$ form as the velocity case, and its prediction Jacobian is likewise taken with respect to ϵ_k. Consecutive sigmas differ on the deployed schedule, so no evaluation carries a zero response and disabling does not arise there; the terminal step to $\sigma_0 = 0$ does not zero that response either, unlike the data form above. Independent internal stages obey the same product rule, with $B_k f_k$ replaced by $\sum_j B_{k,j} f_{k,j}$ and $c_{k,j} = B_{k,j} \prod_{i<k} A_i$.

For DPM-Solver++ 2M midpoint updates (Lu et al., 2025), define $\ell_k = \log(\alpha_k/\sigma_k)$, $h_k = \ell_{k-1} - \ell_k$, and $\nu_k = h_{k+1}/h_k$. Here α_k is the diffusion amplitude. Then

$$\begin{aligned} x_{k-1} &= A_k x_k + B_k D_k + C_k D_{k+1}, \qquad A_k = \frac{\sigma_{k-1}}{\sigma_k}, \\ B_k &= -\alpha_{k-1}(e^{-h_k} - 1)\left(1 + \frac{1}{2\nu_k}\right), \qquad C_k = \frac{\alpha_{k-1}(e^{-h_k} - 1)}{2\nu_k}, \\ c_k &= \underbrace{B_k \prod_{j=1}^{k-1} A_j}_{\text{current prediction}} + \underbrace{\mathbf{1}_{\{k\geq 2\}} C_{k-1} \prod_{j=1}^{k-2} A_j}_{\text{history reuse}}. \end{aligned} \tag{13}$$

Startup uses $C_K = 0$ and $B_K = -\alpha_{K-1}(e^{-h_K} - 1)$; a final first-order step uses its actual B_1, C_1. For noise-prediction networks, the conversion $D_k = (x_k - \sigma_k \epsilon_k)/\alpha_k$ also determines the prediction Jacobian.

A.2 Uniqueness of the Feedback Gain

The scalar surrogate separates the quantities the gain can match from the Jacobian variation it cannot remove. At one site, write $a_t = c_t s_t$. With exact storage and unsaturated rounding,

$$\begin{aligned} u_t &= z_t + \alpha_t r_{t-1}, \qquad q_t = \mathrm{round}(u_t), \qquad r_t = u_t - q_t, \\ \bar{e}_t &= q_t - z_t = (u_t - z_t) - (u_t - q_t) = \alpha_t r_{t-1} - r_t. \end{aligned} \tag{14}$$

For $t \geq 1$, $a_t \neq 0$, and $J_t \neq 0$, the locally aligned target is $\lambda_t a_{t-1} J_t r$ for every antecedent residual r. Its scalar matching condition is

$$\begin{aligned} a_t \alpha_t J_t r = \lambda_t a_{t-1} J_t r \quad \text{for all } r &\iff (a_t \alpha_t - \lambda_t a_{t-1}) J_t = 0 \\ &\iff \alpha_t = \lambda_t \frac{a_{t-1}}{a_t} = \lambda_t \frac{c_{t-1} s_{t-1}}{c_t s_t}. \end{aligned} \tag{15}$$

No full-rank assumption is needed: a nonzero matrix cannot be annihilated by a nonzero scalar. This proves uniqueness for the stated matching target, rather than global optimality over rounding rules.

To examine departures from this target, let $p_t = a_t r_t$, $p_{-1} = 0$, and $\Delta_\lambda J_t = \lambda_{t+1} J_{t+1} - J_t$. For a different scalar gain b_t, set $\kappa_t = a_t b_t - \lambda_t a_{t-1}$ for $t \geq 1$ and $\kappa_0 = 0$, with $b_0 = 0$. With exact state and no resets,

$$\begin{aligned} \sum_{t=0}^{T} a_t J_t \bar{e}_t &= \sum_{t=0}^{T} J_t\big(\lambda_t p_{t-1} - p_t + \kappa_t r_{t-1}\big) \\ &= -J_T p_T + \sum_{t=0}^{T-1} \Delta_\lambda J_t \, p_t + \underbrace{\sum_{t=0}^{T} \kappa_t J_t r_{t-1}}_{\text{gain mismatch}}. \end{aligned} \tag{16}$$

TeleQuant removes this mismatch. Unit-delay feedback matches this target for all residuals only when $a_t = \lambda_t a_{t-1}$ at every nondegenerate step. Exact cancellation of the actual preceding sensitivity additionally requires $\lambda_t J_t = J_{t-1}$. At $a_t = 0$, incoming feedback is disabled; response-sign resets are handled below.

A.3 Proof of Theorem 1

Proof. Using the residual identity in Equation 14, assume exact storage, no saturation or resets, and the nonzero-response condition of the theorem. Write $a_t = c_t s_t$ and introduce the weighted residual $p_t = a_t r_t$, with $p_{-1} = 0$. The gain in Equation 3 gives

$$a_t \bar{e}_t = \lambda_t p_{t-1} - p_t, \qquad \|p_t\|_\infty \leq \rho |a_t|. \tag{17}$$

For a shaped site l, summation by parts now yields

$$\begin{aligned} \delta_l^{(1)} &= \sum_{t=0}^{T} J_t\big(\lambda_t p_{t-1} - p_t\big) \\ &= \sum_{t=0}^{T-1} \lambda_{t+1} J_{t+1} p_t - \sum_{t=0}^{T} J_t p_t \\ &= \underbrace{-J_T p_T}_{\text{terminal residual}} + \underbrace{\sum_{t=0}^{T-1} \big(\lambda_{t+1} J_{t+1} - J_t\big) p_t}_{\text{sensitivity variation}}. \end{aligned} \tag{18}$$

Let $\Delta_\lambda J_t = \lambda_{t+1} J_{t+1} - J_t$. Applying the induced norm inequality to each term gives

$$\begin{aligned} \|\delta_l^{(1)}\|_2 &\le \|J_T\|_{\infty\to 2}\|p_T\|_\infty + \sum_{t=0}^{T-1} \|\Delta_\lambda J_t\|_{\infty\to 2}\|p_t\|_\infty \\ &\le \rho \left(|a_T|\,\|J_T\|_{\infty\to 2} + \sum_{t=0}^{T-1} |a_t|\,\|\Delta_\lambda J_t\|_{\infty\to 2} \right). \end{aligned} \tag{19}$$

Restoring the layer index and adding the unshaped contribution,

$$\delta^{(1)} = \sum_{l\in\mathcal{L}_{\text{shape}}} \delta_l^{(1)} + \delta_{\text{static}}^{(1)}, \qquad \|\delta^{(1)}\|_2 \le \sum_{l\in\mathcal{L}_{\text{shape}}} \|\delta_l^{(1)}\|_2 + \|\delta_{\text{static}}^{(1)}\|_2. \tag{20}$$

Substituting the single-site identity into Equation 20 gives Equation 4. Applying the single-site bound gives Equation 5. Neither step assumes cancellation between layers. □

The gain absorbs changes in solver response and activation scale into p_t. What remains in the interior is

$$\Delta_\lambda J_t = \lambda_{t+1} \underbrace{(J_{t+1} - J_t)}_{\text{temporal variation}} - \underbrace{(1-\lambda_{t+1}) J_t}_{\text{damping contribution}}. \tag{21}$$

Thus, if $\lambda_t = 1$ and $J_t = J$ throughout the segment, $\delta_l^{(1)} = -Jp_T$: all interior residuals cancel, even when c_t and s_t vary. For $\lambda_t < 1$, the second term in Equation 21 remains even with constant sensitivity. The identity therefore separates temporal variation from the cost of damping. Since residuals and sensitivities depend on the realized trajectory, the deployed damping is selected by the empirical comparison in Table 8.

A.4 State Compression and Tokenwise Scales

The theorem isolates the ideal mechanism. The deployed operator additionally compresses and resets state, while dynamic hosts may use a different scale for each token. We treat these two extensions separately.

Compressed and reset state. Let r_t be the outgoing state after clipping and resets, $\widehat{r}_t = r_t + \xi_t$ its stored approximation, and b_t the applied gain. Define the discarded residual $d_t = u_t - q_t - r_t$ and use κ_t from Equation 16. Direct substitution gives

$$\begin{aligned} \bar{e}_t &= b_t(r_{t-1} + \xi_{t-1}) - r_t - d_t, \\ a_t\bar{e}_t &= \lambda_t p_{t-1} - p_t + \underbrace{\kappa_t r_{t-1} + a_t b_t \xi_{t-1} - a_t d_t}_{v_t}, \\ \delta_l^{(1)} &= -J_T p_T + \sum_{t=0}^{T-1} \Delta_\lambda J_t\, p_t + \underbrace{\sum_{t=0}^{T} J_t v_t}_{R_{\text{state},l}}. \end{aligned} \tag{22}$$

The residual v_t collects three departures from ideal cancellation: gain mismatch, storage error, and discarded state. They enter one additive correction, so their propagation can be assessed without repeating the telescoping proof. Taking $\rho = 1/2$ for the deployed outgoing state, its norm is bounded by

$$\begin{aligned} \|R_{\text{state},l}\|_2 &\le \sum_{t=0}^{T} \|J_t\|_{\infty\to 2}\|v_t\|_\infty \\ &\le \sum_{t=0}^{T} \|J_t\|_{\infty\to 2} \left(\rho|\kappa_t| + |a_t b_t|\,\|\xi_{t-1}\|_\infty + |a_t|\,\|d_t\|_\infty\right). \end{aligned} \tag{23}$$

For the symmetric INT8 format in Algorithm 1, $\widehat{r} = \text{round}(254r)/254$ implies $\|\xi\|_\infty \le 1/508$. Gain clipping or incoming resets enter through κ_t; saturation and outgoing resets enter through d_t. A reset prevents an unreliable residual from being reused, but does not erase the current quantization error. The propagated bound still depends on the gains and sensitivities, so accuracy follows from the complete recurrence. Summing $R_{\text{state},l}$ over shaped sites gives the full correction.

Tokenwise scales. Let S_t be a positive diagonal scale matrix with aligned token coordinates, and set $W_t = c_t S_t$ for nonzero c_t. Under the ideal-state assumptions, define $p_t = W_t r_t$ and

$$\begin{aligned} \Lambda_t &= \lambda_t W_t^{-1} W_{t-1}, \\ W_t\bar{e}_t &= W_t(\Lambda_t r_{t-1} - r_t) = \lambda_t p_{t-1} - p_t, \\ \delta_l^{(1)} &= \sum_{t=0}^{T} J_t W_t \bar{e}_t = -J_T p_T + \sum_{t=0}^{T-1} \Delta_\lambda J_t\, p_t. \end{aligned} \tag{24}$$

Consequently, the bound becomes $\rho\|J_T W_T\|_{\infty\to 2} + \rho \sum_{t<T} \|\Delta_\lambda J_t W_t\|_{\infty\to 2}$. Affine zero points also preserve the physical-error relation:

$$\zeta_t = S_t^{-1} x_t + z_t^0, \qquad \widehat{x}_t = S_t(q_t - z_t^0) \quad \Longrightarrow \quad \widehat{x}_t - x_t = S_t(q_t - \zeta_t). \tag{25}$$

A.5 Second-Order Error Bound

Accounting for state errors still leaves a first-order model. To determine when its advantage survives nonlinear propagation, let e_l stack the physical perturbations at one site, and let $F_l(e_l)$ be the endpoint with other perturbations fixed at zero. Assume F_l is twice continuously differentiable on the segment $[0, e_l]$ and $\|D^2 F_l\| \le L_l$ there. For $g_l = DF_l(0)e_l$, Taylor's integral formula yields

$$\begin{aligned} F_l(e_l) - F_l(0) &= g_l + R_l, \\ R_l &= \int_0^1 (1-u) D^2 F_l(u e_l)[e_l, e_l]\, du, \\ \|R_l\|_2 &\le \int_0^1 (1-u) L_l \|e_l\|_2^2\, du = \tfrac{1}{2} L_l \|e_l\|_2^2 =: b_l. \end{aligned} \tag{26}$$

Writing $[x]_+ = \max\{x, 0\}$, the triangle and reverse triangle inequalities give

$$[\|g_l\|_2 - b_l]_+^2 \;\le\; \|F_l(e_l) - F_l(0)\|_2^2 \;\le\; (\|g_l\|_2 + b_l)^2. \tag{27}$$

Apply this interval to the RTN arm R and shaped arm S under the same endpoint map. Whenever the expectations are finite,

$$\begin{aligned} &\mathbb{E}\big[\|F_l(e_{l,R}) - F_l(0)\|_2^2 - \|F_l(e_{l,S}) - F_l(0)\|_2^2\big] \\ &\quad \ge \mathbb{E}\big[[\|g_{l,R}\|_2 - b_{l,R}]_+^2 - (\|g_{l,S}\|_2 + b_{l,S})^2\big] =: \Delta\mathcal{E}_l^{\mathrm{2nd}}. \end{aligned} \tag{28}$$

Thus $\Delta\mathcal{E}_l^{\mathrm{2nd}} > 0$ is sufficient for endpoint improvement: the first-order advantage survives the nonlinear remainder. If g_l is replaced by the scalar surrogate, its approximation error must be added to b_l. Joint shaping instead uses a stacked perturbation map, including cross-layer curvature.

The calibration score in Equation 7 measures local error energy, while the endpoint condition in Equation 28 is stated on the shaped response. For example, let $e_R = (1, -1)$ and $e_S = (\varepsilon, \varepsilon)$ with $0 < \varepsilon < 1$, under a constant unit response:

$$\|e_S\|_2^2 = 2\varepsilon^2 < 2 = \|e_R\|_2^2, \qquad |\mathbf{1}^\top e_S|^2 = 4\varepsilon^2 > 0 = |\mathbf{1}^\top e_R|^2. \tag{29}$$

Lower local energy can therefore coexist with larger endpoint error. Moreover, RTN already minimizes squared error for a fixed activation and grid; the paired-rollout screen measures changes along distinct trajectories. Placement is therefore decided by measurement, and the accepted mask is evaluated as a whole after selection.

B Benchmark Models

The four backbones in Section 5.1 cover class-conditioned images, text-conditioned images, and video generation. Their roles below distinguish controlled mechanism studies from evaluation across architectures and quantization hosts. All host weights and quantization rules remain fixed when feedback is enabled. Evaluation manifests are specified in Appendix C and execution protocols in Appendix D.

DiT-XL/2. DiT-XL/2 (Peebles & Xie, 2023) is the class-conditional Diffusion Transformer used for the main ImageNet comparison and the controlled rounding, state-precision, and placement studies. Its repeated Transformer projections provide a common setting for isolating temporal feedback while holding the spatial quantizer fixed. The main comparison uses a ViDiT-Q host; additional Uniform PTQ and SVDQuant configurations are identified separately in Appendix E.2.

PixArt-Σ. PixArt-Σ (Chen et al., 2024) is a text-conditioned Diffusion Transformer. We evaluate its frozen ViDiT-Q W4A8 configuration to test feedback under dynamic activation scaling. This backbone also supports the matched comparison with TCEC in Table 3.

SDXL. SDXL (Podell et al., 2024) uses a convolutional UNet denoiser with attention, extending the evaluation beyond Transformer denoisers. Its SVDQuant W4A4 host combines a quantized main branch with a high-precision low-rank branch. TeleQuant acts on the main branch; the low-rank branch retains its original computation.

Open-Sora 1.2. Open-Sora 1.2 (Zheng et al., 2024) is the text-to-video backbone used to evaluate spatial fidelity and consistency across video frames. We compare the frozen ViDiT-Q W4A8 host with and without feedback and include the W8A8 host as a higher-precision reference. Residuals are propagated across denoising evaluations, with separate state for each guidance stream.

C BENCHMARK DATASETS

The models in Appendix B use the following evaluation manifests. Comparisons between a host and its TeleQuant extension share prompts or class labels, initial latents, and sampling settings. Dataset-specific scores are compared within each backbone, since sample counts and reference distributions differ across benchmarks.

ImageNet. DiT-XL/2 is evaluated on class-conditional ImageNet-256 $\times$ 256 (Deng et al., 2009). Distributional evaluation uses 50,000 generated samples pooled from five independently seeded runs, with the ADM evaluation harness (Dhariwal & Nichol, 2021) and ImageNet-1K training reference statistics. Paired reconstruction uses 50,000 class–seed instances, each matched to an FP16 trajectory. Per-instance PSNR, SSIM, and LPIPS are cached for the uncertainty analysis in Appendix E.4.

MS-COCO. PixArt-Σ uses ViDiT-Q's `coco_1024` manifest, a curated 1,024-prompt subset of MS-COCO 2017 (Lin et al., 2014). Each of the FP16, host, and host + TeleQuant arms generates the complete manifest once. The same prompt and seed pairs are used for trajectory reconstruction and the TCEC comparison.

MJHQ. SDXL uses SVDQuant's 5,000-prompt MJHQ subset (Li et al., 2025). Each evaluated arm generates the complete manifest once, using identical text encodings and initial latents across arms. This evaluation follows the selected host configuration, with its scores kept separate from the MS-COCO subset.

VBench. Open-Sora 1.2 uses ViDiT-Q's 251-prompt curated VBench composition (Huang et al., 2024b). We evaluate imaging quality, motion smoothness, background consistency, subject consistency, human action, and overall consistency. Overall consistency denotes the video–text consistency dimension of VBench. Human action is evaluated on the eligible human-action prompts. Frame-level reconstruction and temporal-difference results use the same generated clips.

D IMPLEMENTATION DETAILS

The following protocols connect the recurrence in Section 4 to the comparisons in Section 5. We first fix the sampling conditions and host interfaces, then specify state execution and offline mask calibration. The final subsections define how quality and runtime are measured. Main experiments use a fixed generation-time configuration; adaptive damping is an additional ablation in Appendix E.3.

D.1 SAMPLING AND MATCHED EVALUATION

Table 4 specifies the sampling configurations used for the paired comparisons. For each backbone, the host and TeleQuant arms share the checkpoint, scheduler, guidance scale, initial noise, text or class conditioning, and VAE decoder. The DiT FP16 distributional row in Table 1 is a published reference; its paired fidelity metrics instead use separately generated FP16 trajectories under the stated evaluation schedule. Dataset manifests are defined in Appendix C.

Table 4: Sampling settings for the paired generation experiments. Video resolution is specified per frame.

Backbone	Resolution	Sampler	Steps	CFG
DiT-XL/2	256 $\times$ 256	DDIM	50	1.5
PixArt-Σ	512 $\times$ 512	DPM-Solver++	20	4.5
SDXL	1024 $\times$ 1024	Euler-EDM	50	7.5
Open-Sora 1.2	240 $\times$ 424, 16 frames	Rectified flow	100	4.0

D.2 BASELINE AND HOST INTEGRATION

The comparisons use FP16 outputs as paired references, frozen quantizers as hosts, and alternative temporal rules to isolate the feedback mechanism. We specify these roles separately because changing the host and changing its rounding rule test different hypotheses.

Unquantized FP16 reference controls. For each architecture, we compute an unquantized FP16 reference under identical weights, guidance scales, ODE solver parameters, random seeds, and VAE decoders. This provides the reference for paired trajectory fidelity metrics.

Naive uniform RTN baseline. The uniform Round-to-Nearest baseline applies symmetric min-max dynamic quantization at each step:

$$s_{t,l} = \frac{\max(\max(|X_{t,l}|), \varepsilon_s)}{2^{b-1} - 1}, \qquad q_{t,l} = \text{clip}\left(\left\lfloor \frac{X_{t,l}}{s_{t,l}} \right\rceil, -2^{b-1}, 2^{b-1} - 1\right), \tag{30}$$

where $\varepsilon_s > 0$ prevents division by zero for an all-zero tensor. This baseline excludes spatial and temporal calibration.

ViDiT-Q integration protocol. The main DiT-XL/2, PixArt-Σ, and Open-Sora 1.2 comparisons use frozen ViDiT-Q hosts (Zhao et al., 2025). For PixArt-Σ and Open-Sora 1.2, integration uses the official release, retaining its channel transformations, activation scaling, precision choices, and GEMM dispatch. Feedback is inserted at the pre-rounding point in Equation 2. Under per-token scaling, the scale $s_{x,t}$ and gain α_t are length-M vectors, with one value per token broadcast across channels. The kernel applies $\boldsymbol{\alpha}_t \odot \widehat{\mathbf{R}}_{t-1}$ row-wise; the scale-ratio derivation in Appendix A.4 ensures that this update retains the weighted-residual identity under its stated assumptions. Algorithm 1 specifies the stored-state update.

SVDQuant integration protocol. For SDXL, the host is the SVDQuant W4A4 configuration of Nunchaku, the inference engine released with the method (Li et al., 2025), reproduced in PyTorch on SDXL-base-1.0 at the released INT4/rank-32 UNet checkpoint configuration: activation outliers are absorbed into a parallel low-rank FP16 branch while the high-rank residual is quantized to 4 bits, and the host supplies its own activation quantization and matrix multiplication path. TeleQuant is inserted immediately before that activation quantizer, in the same smoothed domain the quantizer operates in, so the intervention changes only the INT4 activation codes and their dynamic scales: the low-rank activation is read from the unmodified input, which preserves the released rank-32 low-rank branch, and the weight codebooks and host GEMM dispatch stay untouched. Evaluation uses the 5,000-prompt MJHQ subset of Appendix C.

Controlled temporal ablations. The rounding arms in Section 5.2 are standard round-to-nearest, independent stochastic rounding, unit-delay feedback with $\alpha = 1$, and TeleQuant. They share the frozen ViDiT-Q host, initial conditions, and solver schedule, and each arm changes only the rounding rule applied at the selected projection sites: the host weights, activation scales, zero-points, and GEMM dispatch are unchanged. Stochastic rounding draws each activation's decision independently between the two neighboring grid points with probability matching, so its error is locally unbiased and carries no deterministic structure for a fixed gain to act on. The resulting trajectories can differ because each rounding rule changes the inputs to later evaluations. Both feedback arms use the same INT8 state format.

TCEC provides a separate trajectory-level comparison in Table 3, using the same frozen ViDiT-Q W4A8 hosts, initial conditions, and sampling schedules on DiT-XL/2 and PixArt-Σ. It corrects accumulated quantization error at the sampling step rather than inside the denoiser. Write μ_t for the prediction the solver consumes at evaluation t and $\widetilde{\mu}_t$ for the host's quantized prediction. The baseline rescales the quantized prediction by a coefficient $K_{t,i}$ per evaluation and per output channel i, so the corrected prediction is $(1 - K_t) \odot \widetilde{\mu}_t$, and fits those coefficients on the calibration set by minimizing the regularized reconstruction objective (Liu et al., 2026a)

$$\mathcal{L}(K) = \sum_{t,i,j,k} \left[(1 - K_{t,i})\widetilde{\mu}_{t,i,j,k} - \mu_{t,i,j,k}\right]^2 + \lambda_1 \sum_{t,i} K_{t,i}^2. \tag{31}$$

Here μ is the matched full-precision prediction at the same input along the host's calibration trajectory, obtained by replaying the unquantized denoiser on the recorded input through the same path that produces the FP16 references of Appendix D.1; it is not a ground-truth image. Setting the derivative to zero gives the closed form

$$K_{t,i} = \frac{\sum_{b,j,k} \widetilde{\mu}_{t,i,j,k}^2 - \sum_{b,j,k} \widetilde{\mu}_{t,i,j,k}\, \mu_{t,i,j,k}}{\sum_{b,j,k} \widetilde{\mu}_{t,i,j,k}^2 + \lambda_1}, \tag{32}$$

with the sums taken over the calibration batch and the non-channel axes of the prediction. The two moment sums are accumulated in a single pass over the same 32 calibration instances used in Appendix D.4, and every step of the schedule must be visited, so the cached table carries one row per denoising step and one column per output channel. We set $\lambda_1 = 10^{-3}$ on both backbones and record it, together with the host, the schedule, and the solver-response table, in the coefficient artifact. Generation then performs one lookup and one scaling of the prediction per evaluation, and inserts nothing into the denoiser: the host's rounding rule, activation scales, and GEMM dispatch are those of the frozen host arm. This comparison therefore tests compensation at the solver level and feedback at internal rounding sites as separate interventions.

D.3 Residual State and Fused Execution

Algorithm 1 specifies the per-step lifecycle. Cached INT8 residuals are dequantized in SRAM, added before host rounding, and updated in the same activation pass. All main configurations use global damping $\lambda = 0.75$ and INT8 state; the operator implements these fixed settings without fitting parameters during generation. The design rationale is in Section 4.2; the accuracy–cost sweeps are in Appendix E.2.

Algorithm 1 TeleQuant Fused Noise-Shaping Operator

Input: normalized activation $\mathbf{Z}_t \in \mathbb{R}^{M\times K}$, quantized weight $\mathbf{W}_{\text{int}} \in \mathbb{Z}^{N\times K}$, dynamic scale $\mathbf{s}_{x,t} \in \mathbb{R}^{M}_{>0}$ defined per token or as a scalar under per-tensor scaling, weight scale s_w, cached INT8 residual $\mathbf{R}_{t-1} \in \mathbb{Z}^{M\times K}$, analytical feedback gain $\boldsymbol{\alpha}_t \in \mathbb{R}^{M}$ defined per token and broadcast across K, or as a scalar under per-tensor scaling, and the previous solver-response sign $\sigma_{t-1} \triangleq \text{sign}(c_{t-1})$.
Output: projection output $\mathbf{Y}_t \in \mathbb{R}^{M\times N}$, updated residual buffer $\mathbf{R}_t \in \mathbb{Z}^{M\times K}$, and current sign $\sigma_t \triangleq \text{sign}(c_t)$.
$\widehat{\mathbf{R}}_{t-1} \leftarrow \mathbf{R}_{t-1}/254$; set incoming gain and state to zero on initialization, $c_t = 0$, or $\sigma_t \neq \sigma_{t-1}$ *// in-SRAM residual dequantization*
$\mathbf{U}_t \leftarrow \mathbf{Z}_t + \boldsymbol{\alpha}_t \odot \widehat{\mathbf{R}}_{t-1}$ *// pre-rounding modulation; one gain per row, broadcast over K under per-token scaling*
$\mathbf{Q}_t \leftarrow \text{clip}(\lfloor \mathbf{U}_t \rceil, q_{\min}, q_{\max})$ *// fused integer discretization*
$\mathbf{r}_t \leftarrow \text{clip}(\mathbf{U}_t - \mathbf{Q}_t, -0.5, 0.5)$ *// coordinate residual*
Set $\mathbf{r}_t[m,:] \leftarrow \mathbf{0}$ for every row m where any coordinate of $\mathbf{U}_t[m,:]$ reaches $q_{\min}$ or $q_{\max}$, or $c_t = 0$, or $\sigma_t \neq \sigma_{t-1}$ *// anti-windup clamp: reset on either-direction saturation or solver-response sign change*
$\mathbf{R}_t \leftarrow \text{round}(254\mathbf{r}_t)$; compute $\mathbf{Y}_t$ by host GEMM dispatch and host output dequantization
return $\mathbf{Y}_t, \mathbf{R}_t, \sigma_t$

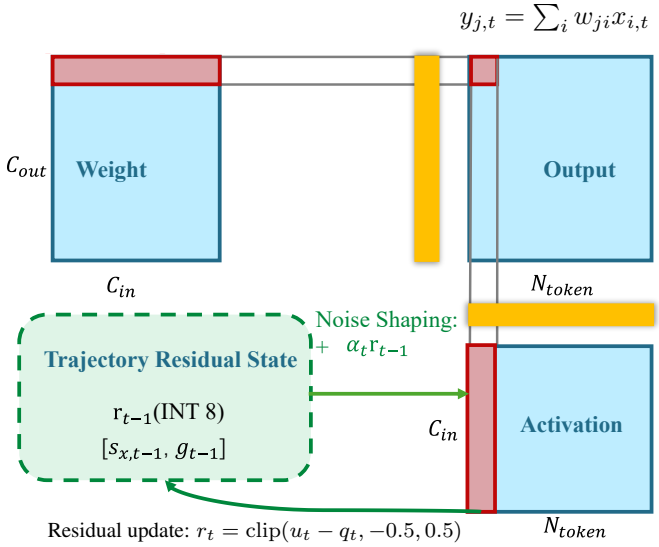


Figure 8: Fused operator microarchitecture. Activation feedback precedes host integer GEMM dispatch, with compressed residual state maintained separately. The schematic writes solver response c_{t-1} as g_{t-1} and uses a transposed matrix orientation relative to $Y = XW^\top$. Algorithm 1 specifies rounding, state updates, and boundary resets.

Algorithm 1 defines a single update; efficient sampling also requires executing that update without repeatedly moving intermediate tensors. Figure 8 illustrates the tiled implementation. Each activation tile and its previous residual are loaded once for feedback addition, host rounding, boundary checking, and state requantization. The integer codes pass to the host GEMM with its original scale and zero-point corrections. Per-token gains are broadcast across channels as derived in Appendix A.4.

Conditional and unconditional guidance streams are keyed separately and never read each other's residual rows. Where the host evaluates the two streams as separate denoiser calls, each site holds one INT8 buffer per stream; where it batches them into a single call, the two streams occupy disjoint slices of one per-site buffer. All buffers are zero-initialized at the start of a generation and released at completion, and each stored residual is replaced in place by the evaluation that consumes it, so a step's write never races the read that preceded it. Boundary-hit rows clear outgoing state; zero or sign-changing solver responses disable incoming feedback and reset outgoing state. These policies are part of the deployed recurrence rather than the idealized no-reset identity.

D.4 Offline Site Calibration

Section 4.3 defines two screens with different roles: sensitivity variation identifies candidate sites, and paired rollouts test their realized local error. We apply both on 32 calibration instances, then fix the accepted mask for generation. This count sets the offline calibration budget. The variation screen records how rapidly each site's sensitivity changes, and the paired rollout decides placement; Figure 4 reports the recorded ratios for the reported DiT-XL/2 configuration, where they peak near 0.8 and stay below the threshold of Equation 6. Algorithm 2 gives the complete procedure.

For each site, a Gaussian probe $v_l \sim \mathcal{N}(0, I)$ is reused across evaluations to compute the scaled probe responses $J_{t,l}(s_{t,l} \odot v_l)$ by Jacobian-vector products, with $s_{t,l}$ the scale of Equation 2 broadcast over the site's tokens. The identities $\mathbb{E}_v\|J_{t,l}S_{t,l}v\|_2^2 = \|J_{t,l}S_{t,l}\|_F^2$ and $\mathbb{E}_v\|(J_{t+1,l} - J_{t,l})S_{t,l}v\|_2^2 = \|(J_{t+1,l} - J_{t,l})S_{t,l}\|_F^2$, where $S_{t,l}$ is the diagonal scaling that $s_{t,l}$ induces on the coordinates of the activation the site rounds, motivate substituting these probes into Equation 6. Reusing the probe retains the temporal comparison. The probe avoids constructing full Jacobians, which keeps calibration cost low. Sites with a zero denominator are excluded. The threshold $\tau_{\text{tv}} = 0.85$ sets the variation screen, and the second screen of Equation 7 adds the measured local-error gain on top of the ratio; the sufficient endpoint condition is derived separately in Appendix A.5.

Algorithm 2 Calibration-Guided Site Selection

Require: candidate sites $\{1, \dots, L\}$, calibration prompts $\mathcal{D}$ of size $|\mathcal{D}| = 32$, threshold $\tau_{\text{tv}} = 0.85$, damping $\lambda = 0.75$

run one frozen rollout on $\mathcal{D}$ shared across all l; cache the scaled probe responses $J_{t',l}(s_{t,l} \odot v_l)$ at $t' = t$ and $t' = t + 1$, together with $a_{t,l} = c_t s_{t,l}$, for every site l and step t
$\mathcal{L}_{\text{cand}} \leftarrow \emptyset$
for $l = 1, \dots, L$ **do**
    skip the site if the denominator is zero; otherwise compute V_l/P_l via Equation 6 from the cached scaled probe responses
    if $V_l/P_l < \tau_{\text{tv}}$ **then**
        $\mathcal{L}_{\text{cand}} \leftarrow \mathcal{L}_{\text{cand}} \cup \{l\}$
    end if
end for
$\mathcal{L}_{\text{shape}} \leftarrow \emptyset$
for $l \in \mathcal{L}_{\text{cand}}$ **do**
    run one additional shaped rollout on $\mathcal{D}$ with $\lambda = 0.75$ active at site l
    compute $\Delta\mathcal{E}_l$ via Equation 7 from the paired frozen/shaped errors
    if $\Delta\mathcal{E}_l > 0$ **then**
        $\mathcal{L}_{\text{shape}} \leftarrow \mathcal{L}_{\text{shape}} \cup \{l\}$
    end if
end for
return $\mathcal{L}_{\text{shape}}$ {recorded once per backbone; no runtime re-selection}

The frozen trajectory supplies baseline errors shared across sites. Each surviving candidate is then tested in a separate shaped rollout with feedback enabled only at that site. Runs share prompts or class labels, initial latents, guidance, and schedule nodes; their actual states can differ. We retain a candidate when the score in Equation 7 is positive. This is a measured local-error criterion, whose distinction from a sufficient endpoint condition is derived in Appendix A.5. The feedback shift is bounded by $|\alpha_{t,l}|s_{t,l}/2$, so its physical size also depends on the gain.

With L_c surviving candidates, the second screen requires $32L_c$ shaped trajectories in addition to the shared frozen baselines. Jacobian-vector products contribute a separate cost per candidate and evaluation. A candidate rejected by the ratio costs no shaped rollout, so the cheaper screen runs first. The final selected mask is evaluated jointly; if it is empty, generation reduces to the frozen host. The masks are backbone-specific because sensitivities and calibration trajectories differ.

D.5 Evaluation Metrics and Uncertainty Protocol

Quality is evaluated along two complementary axes: distributional image quality and fidelity to matched FP16 outputs. We define their metrics below, followed by the uncertainty protocol. The video benchmark dimensions and eligible prompt subsets are specified in Appendix C.

Fréchet Inception Distance (Heusel et al., 2017) evaluates the squared Wasserstein-2 distance between multivariate Gaussian representations of Inception-v3 pool3 activations for generated and reference image feature distributions. Spatial FID (Dhariwal & Nichol, 2021) incorporates intermediate spatial feature representations to

penalize high-frequency structural distortions and repetitive artifacts. Inception Score (Salimans et al., 2016) quantifies class-conditional certainty and marginal diversity.

We compare each quantized output with its FP16 counterpart generated from the same initial conditions. PSNR measures pixel error through $10 \log_{10}(1/\mathrm{MSE})$ for pixels in $[0, 1]$. SSIM measures local luminance, contrast, and structure (Wang et al., 2004). AlexNet-LPIPS measures perceptual distance with calibrated deep features (Zhang et al., 2018). These complementary metrics assess reference fidelity, and the distributional metrics are reported alongside them.

The paired-bootstrap protocol resamples complete generation instances with replacement for 10,000 trials, keeping the arms paired. Table 11 reports the 2.5th and 97.5th percentile endpoints. Its DiT rows evaluate Uniform PTQ, whereas the main-table DiT rows use ViDiT-Q. The main tables report point estimates; the appendix reports these explicit interval endpoints for the stated evaluation arms. Comparisons in this paper report point-estimate changes with paired intervals. Population FID and IS are evaluated on sample sets, so interval reporting applies to the paired per-instance metrics.

D.6 Hardware and Timing Protocol

Runtime measurements use isolated workloads to avoid interference from unrelated generation jobs. The two platforms are an NVIDIA GeForce RTX 4060 with 8 GB memory and an NVIDIA A100-SXM4 with 40 GB memory. The former profiles the projection path on a consumer GPU; the latter supports the operator sweep and full image pipelines. Quality-evaluation manifests remain those of Appendix C.

Timing procedure. To reduce thermal, frequency-scaling, and cache-state variability:

1. Every benchmark execution begins with 100 untimed burn-in iterations to stabilize GPU core clock frequencies and thermal envelopes.
2. Standalone kernel benchmarks alternate baseline and TeleQuant execution across 1,500 profiling iterations per round.
3. Between consecutive iterations, we flush GPU L2 cache buffers by allocating and zeroing a disjoint dummy buffer exceeding the physical L2 cache size of 24MB for RTX 4060 and 40MB for A100.
4. GPU timing is captured via asynchronous CUDA events recorded with `cudaEventRecord`, with synchronization outside GPU-timed blocks. End-to-end wall-clock timing additionally requires a synchronized host timer covering CPU scheduling and dispatch.

The operator benchmark measures standalone execution of the fused Triton operator with tiled residual unscaling, feedback addition, dynamic clamping, INT8 state casting, and row-sum accumulation prior to integer GEMM dispatch. The pipeline benchmark covers wall-clock execution of the full inference pipeline, including all reverse sampling steps, text encoding, scheduling updates, and VAE decoding. This separation isolates the cost of our operator from pipeline-level overheads. Results under both tiers are reported in Appendix E.6.

E Additional Experimental Results

Section 5.2 establishes the main fidelity and efficiency comparisons. Here we examine their supporting evidence: frame-level video errors, choices of bit width and residual state, damping granularity, statistical uncertainty, and operator versus pipeline cost. Video examples accompany frame-level measurements, while image galleries complement the fidelity comparisons. Protocols are in Appendix D; interpretation of trajectory and placement diagnostics is reserved for Appendix F.

E.1 Frame Reconstruction and Temporal Differences

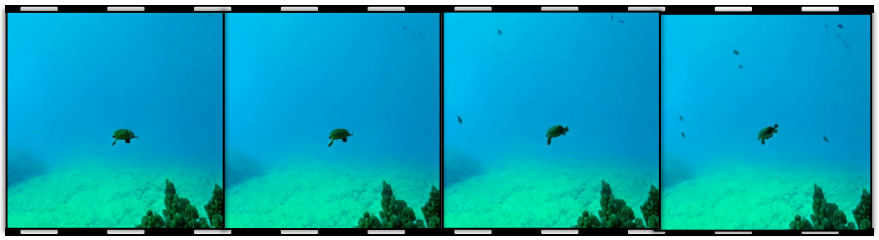
The VBench scores in Table 2 assess perceptual quality and consistency. To examine fidelity to FP16 more directly, Table 5 measures frame reconstruction and inter-frame differences on the same 251-prompt manifest. These temporal-difference measures complement VBench motion smoothness, which uses frame-interpolation consistency. Both metric groups improve relative to the frozen host, with a majority of paired instances improving on each metric. The aggregate results thus show temporal improvement alongside preserved spatial fidelity.

The video metrics assess several independent properties. We therefore examine subject and background structure, frame appearance, and motion in matched examples. Each arm shares the prompt, seed, and sampling settings in Table 4. Only the rounding path changes. In Figure 9, the turtle and coral retain shapes closer to the FP16 sequence with feedback. Figure 10 shows the complementary appearance effect: snow and sky contrast remain closer to the reference. These examples connect the consistency and imaging-quality dimensions of Table 2 to visible differences within a clip.

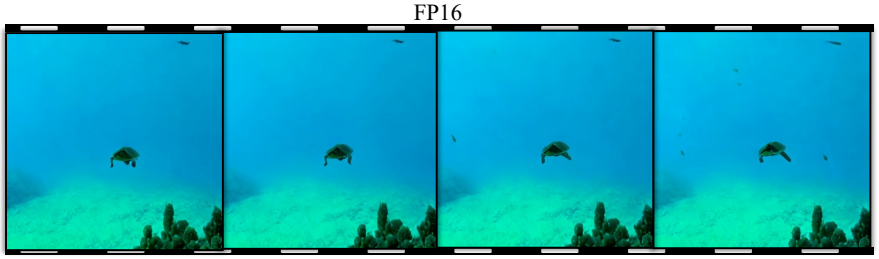
Figure 11 examines a moving subject at close range. Changes in head pose are judged by whether identity and facial structure survive the motion. In the displayed sequence, feedback retains the rounded ears and facial outline while allowing the pose to change. The example therefore complements the static appearance comparison: the desired outcome is stable structure during motion.

Table 5: Frame-level and inter-frame differential metrics, Open-Sora 1.2, VBench manifest, matched seeds. Win rates count the paired instances retained for each metric, and the Δ column is the difference of the two preceding columns at the printed precision.

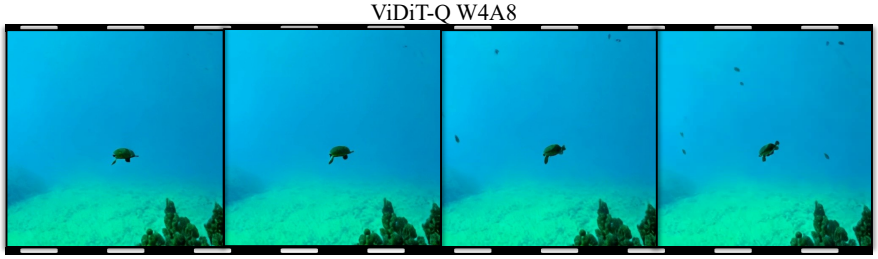
Metric	ViDiT-Q Host	+ TeleQuant, ours	Δ	Win-Rate
Frame PSNR / dB ↑	21.35	**21.38**	+0.03	70.0%
Frame SSIM ↑	0.740	**0.741**	+0.001	70.0%
Temporal PSNR / dB ↑	40.44	**40.49**	+0.05	80.0%
Temporal L_1 Drift $\times 10^{-2}$ ↓	1.21	**1.18**	−0.03	80.0%



FP16



ViDiT-Q W4A8



ViDiT-Q W4A8+TeleQuant(W4A8)

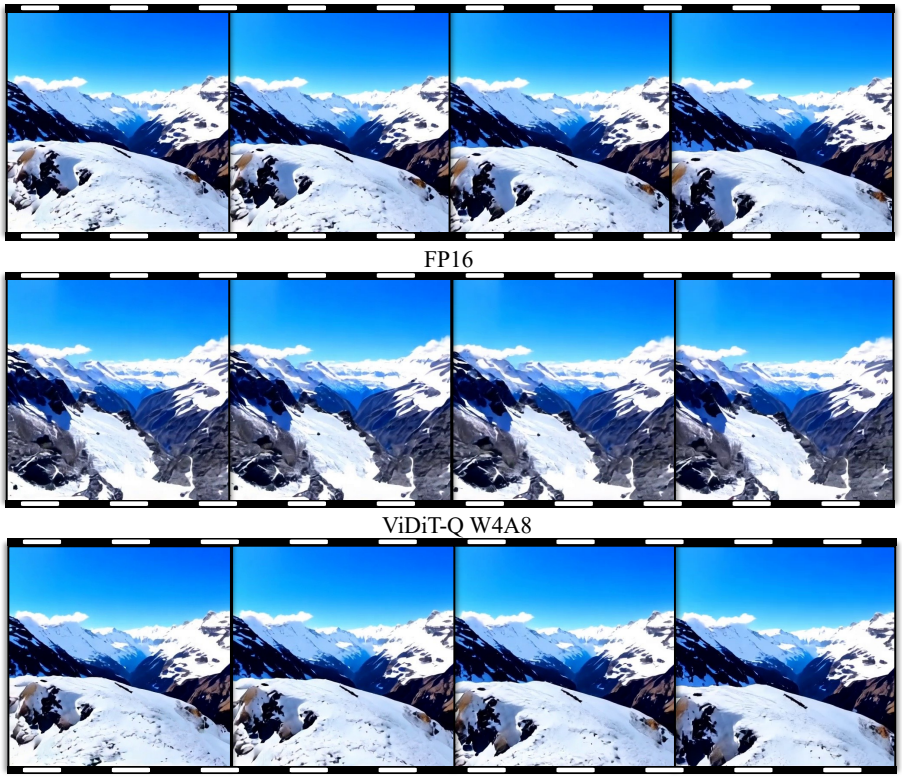
Prompt: *A serene underwater scene featuring a sea turtle swimming through a coral reef. The turtle, with its greenish-brown shell, is the main focus of the video, swimming gracefully towards the right side of the frame. The coral reef, teeming with life, is visible in the background, providing a vibrant and colorful backdrop to the turtle's journey. Several small fish, darting around the turtle, add a sense of movement and dynamism to the scene. The video is shot from a slightly elevated angle, providing a comprehensive view of the turtle's surroundings. The overall style of the video is calm and peaceful, capturing the beauty and tranquility of the underwater world.*

Figure 9: Subject and background structure on Open-Sora 1.2. The matched turtle sequence compares FP16, frozen ViDiT-Q W4A8, and feedback at identical frame indices.

Table 2 assesses consistency across the full manifest of evaluated videos, and Table 5 measures paired frame and temporal-difference errors on the same clips. Together with the frames above, these measurements connect the appearance of a sequence to the reported dimensions.

E.2 Bit Width, Residual Precision, and Damping

These studies vary different parts of the deployed system. Bit width changes the host approximation; residual precision changes the stored feedback; damping changes its strength. All use the 50,000 matched ImageNet trajectories and FP16 counterparts described in Appendix C, with the host specified separately for each experiment.



Prompt: *A majestic snow-covered mountain valley with rugged rocky cliffs and blue sky, filmed from a slowly moving camera; preserve sharp mountain textures and stable colors across frames.*

Figure 10: Frame appearance under the same video protocol. Feedback better retains the FP16 reference's sky and snow contrast in the displayed mountain sequence.

Host bit width. Table 6 evaluates feedback on Uniform PTQ and SVDQuant configurations. These are additional host settings, distinct from the main ViDiT-Q comparison. Their paired rows isolate the benefit of enabling feedback at a fixed host and bit width. Improvements across PSNR, SSIM, and LPIPS show that feedback improves fidelity on hosts without spatial calibration as well. Uniform W4A8 leaves more room for reconstruction recovery than Uniform W8A8 in this study. The gain also persists with SVDQuant, so it holds across calibrated and uncalibrated hosts. The W4A4 row changes the host together with the precision, so its fidelity is specific to that configuration.

Table 6: DiT-XL/2 fidelity across quantization regimes, 50,000 matched trajectories. This table isolates TeleQuant's contribution on a naive Uniform PTQ baseline without spatial calibration and on SVDQuant, as a bit-width sweep distinct from the ViDiT-Q-based damping sweep of Table 8.

Model	Host / Baseline	W/A	PSNR / dB ↑	SSIM ↑	AlexNet LPIPS ↓
DiT-XL/2	*FP16 Reference*	16/16	—	—	—
DiT-XL/2	Uniform PTQ	W8A8	29.842	0.9082	0.0681
DiT-XL/2	Uniform + TeleQuant	W8A8	**30.003**	**0.9119**	**0.0655**
DiT-XL/2	Uniform PTQ	W4A8	17.637	0.5993	0.3862
DiT-XL/2	Uniform + TeleQuant	W4A8	**18.511**	**0.6676**	**0.2863**
DiT-XL/2	SVDQuant Host	W4A4	22.477	0.7512	0.1994
DiT-XL/2	SVDQuant + TeleQuant	W4A4	**22.894**	**0.7635**	**0.1882**

Residual-state precision sweep. After fixing the host, we vary only the precision used to carry residuals between evaluations. Table 7 compares symmetric INT4 and INT8 state with uncompressed FP16 state on frozen ViDiT-Q at $\lambda = 0.75$. This isolates the accuracy–memory trade-off introduced by state compression. INT8 state



Prompt: *A close-up brown bear in a green forest, with both rounded ears and detailed fur clearly visible; the bear moves its head gently while the forest background remains consistent.*

Figure 11: Selected frames from a 16-frame bear clip. The prompt asks for gentle head motion with a consistent forest background. All arms use matched generation settings.

closely matches the reconstruction fidelity of FP16 state while halving persistent buffer memory, making it the deployed precision. INT4 state trades fidelity for a further reduction in buffer size. The near-lossless behavior is specific to INT8 under this experiment.

Table 7: Residual-state precision sweep, DiT-XL/2, frozen ViDiT-Q host, $\lambda = 0.75$.

Residual-State Precision	PSNR / dB ↑	State Memory
INT4	18.950	14 MB
INT8 (deployed)	**19.911**	**28 MB**
FP16 (uncompressed)	19.918	56 MB

Damping factor λ sweep. Table 8 tests three fixed strengths under the same frozen ViDiT-Q host. At $\lambda = 0$, Equation 3 disables feedback and recovers the main-table host. At $\lambda = 1$, the gain fully matches adjacent scalar weights, giving the ideal cancellation case when sensitivities are constant. The intermediate setting $\lambda = 0.75$ yields the best measured fidelity among these alternatives. We therefore adopt this intermediate setting as the deployed default.

The measured ordering follows from the interior term of Equation 5. Its interior term is minimized by the least-squares choice $\lambda^{\star}_{t+1,l} = \langle J_{t,l}, J_{t+1,l} \rangle / \|J_{t+1,l}\|_2^2$, which equals one only when the sensitivity is unchanged and falls below one where it grows during sampling. A single scalar approximates that ratio across steps and sites, and the per-layer and per-step variants in Table 10 improve on it, consistent with a nonuniform ratio. Equation 21 retains a damping contribution even at constant sensitivity. The bound is a worst-case norm over trajectories under exact storage without resets, so it sits above the measured gaps. Damping shrinks the pre-rounding shift, and with it the higher-order and storage contributions.

The comparison evaluates the complete deployed recurrence, including compressed state, boundary handling, and nonlinear propagation. The sweep therefore measures the combined effect of these departures. The analytical separation of sensitivity variation and damping is given by Equation 21.

Table 8: Damping factor λ sweep, DiT-XL/2, frozen ViDiT-Q host.

Damping Factor λ	**PSNR / dB ↑**
$\lambda = 0$ (no feedback, recovers the ViDiT-Q host)	19.59
$\lambda = 0.75$ (best tested, ours)	**19.911**
$\lambda = 1$ (un-damped)	19.805

Rounding configurations and host context. Table 9 adds the uncalibrated Uniform PTQ starting point to the main rounding comparison. Its gap to TeleQuant combines spatial calibration with temporal feedback. The three frozen-host arms share ViDiT-Q, so their relative ordering isolates the rounding configuration. The matched-host interpretation is given in Section 5.2; the extra reference here quantifies the combined improvement of calibration and feedback.

Table 9: Rounding comparison, DiT-XL/2, W4A8. The three frozen-ViDiT-Q arms are as in Figure 7; the remaining row is the uncalibrated Uniform PTQ reference from Table 6, which uses nearest rounding without spatial calibration, so its gap to TeleQuant covers calibration and the rounding rule jointly. Δ PSNR is TeleQuant minus each row.

Rounding Mechanism	**PSNR / dB ↑**	**AlexNet LPIPS ↓**	**Δ PSNR**
Uncalibrated min–max PTQ (nearest)	17.637	0.3862	$+2.274$
Stochastic Rounding	10.392	0.686	$+9.519$
Unit-delay ($\alpha = 1$, uniform gain)	19.69	0.282	$+0.221$
TeleQuant (ours)	**19.911**	**0.262**	—

E.3 Damping Granularity

The indexed gain in Equation 3 permits finer damping than the global value used above. Table 10 tests whether that flexibility improves the trade-off on frozen ViDiT-Q W4A8. Global damping fixes $\lambda = 0.75$. Per-layer damping calibrates one λ_l per site on the same 32-instance set used for placement. Per-step damping estimates λ_t during sampling using a finite-difference approximation to the Jacobian norm. The host and residual format are fixed, so the comparison isolates adaptation granularity.

Table 10: Damping granularity ablation, DiT-XL/2, frozen ViDiT-Q W4A8: latency relative to the global-scalar default.

Damping Scheme	**PSNR / dB ↑**	**SSIM ↑**	**LPIPS ↓**	**Relative Latency**
Global scalar, $\lambda = 0.75$, ours	19.91	0.708	0.262	—
Per-layer adaptive λ_l	19.97	0.712	0.257	$1.003\times$
Per-step adaptive λ_t	20.02	0.715	0.253	$1.118\times$

Both adaptive variants track the global scalar closely. Offline per-layer adaptation offers a small additional gain at low inference cost, whereas per-step adaptation pays for repeated Jacobian-norm estimation. The results therefore support different operating choices: per-layer damping can be attractive when additional calibration is acceptable, while the global scalar provides a simple default across hosts. We retain that default in the main comparisons, where the rounding mechanism is evaluated on its own.

E.4 Paired Uncertainty and Bootstrap Confidence Intervals

Table 11 supplements the evaluated arms with 95% percentile-bootstrap intervals using the paired protocol in Appendix D.5, quantifying sampling uncertainty alongside the point estimates. The DiT rows cover the Uniform PTQ bit-width study; PixArt-Σ and SDXL use the main-table hosts.

The intervals characterize each arm's uncertainty under its evaluation manifest.

E.5 Additional Image Examples

The image comparisons extend the same matched-input protocol to artistic and photographic prompts. Figure 12 shows PixArt-Σ under W4A8, where feedback better preserves contours and local contrast across the displayed scenes. Figure 13 examines SVDQuant W4A4 on SDXL, where the visible differences include facial detail, fabric texture, and background grain. These are distinct host configurations, so the relevant comparison is within each gallery. Their aggregate fidelity and distributional scores are reported in Table 1.

Table 11: 95% paired bootstrap confidence intervals, [2.5%, 97.5%], 10,000 resamples.

Model / Arm	PSNR / dB ↑	SSIM ↑	AlexNet LPIPS ↓
DiT-XL/2 Uniform PTQ W4A8	[17.5645, 17.7085]	[0.5960, 0.6028]	[0.3828, 0.3897]
DiT-XL/2 Uniform PTQ W4A8 + TeleQuant	**[18.4210, 18.6015]**	**[0.6639, 0.6714]**	**[0.2826, 0.2901]**
PixArt-Σ ViDiT-Q W4A8 (`coco_1024`)	[15.1120, 15.3944]	[0.5824, 0.5966]	[0.4410, 0.4546]
PixArt-Σ ViDiT-Q W4A8 + TeleQuant	**[16.7431, 17.0235]**	**[0.6035, 0.6177]**	**[0.3928, 0.4064]**
SDXL SVDQuant W4A4 (MJHQ-5K)	[20.8842, 21.1648]	[0.7385, 0.7519]	[0.2825, 0.2937]
SDXL SVDQuant W4A4 + TeleQuant	**[22.3215, 22.6061]**	**[0.8016, 0.8150]**	**[0.2586, 0.2698]**



Figure 12: PixArt-Σ image examples at 512×512 with 20-step DPM-Solver++. FP16, ViDiT-Q W4A8, and feedback share latents, prompts, and guidance scale 4.5.



Figure 13: SDXL image examples at 1024×1024 with 50-step Euler-EDM. FP16, SVDQuant W4A4, and feedback share latents, prompts, and guidance scale 7.5.

E.6 Operator and Pipeline Costs

The added recurrence has two distinct costs: executing the fused operator and carrying its state through the complete generation pipeline. We measure them separately using Appendix D.6, and report each cost at its own tier.

Operator latency. Table 12 measures the fused frontend on an A100 with fixed projection dimensions $K = N = 4096$ and varying sequence length. This operator sweep complements the RTX 4060 projection-path measurement in Table 13; the execution and timing procedures are specified in Appendix D.6.

Table 12: Frontend operator latency, A100, varying sequence length M with $K = N = 4096$.

Sequence Length M	$M = 256$	$M = 1024$	$M = 2048$	$M = 4096$
Frontend Latency / ms	0.0302	0.0665	0.1014	0.1768

Frontend latency grows more slowly than sequence length over the tested shapes in Table 12. This measures the added operator over a fixed projection width. The complete-pipeline measurements below determine whether its cost remains modest when composed with host execution.

Pipeline latency and memory.

Table 13 compares complete PixArt-Σ and SDXL generation pipelines, including text encoding, sampling, and VAE decoding. The PixArt configuration uses eight projection sites with one residual per site: the released pipeline evaluates the conditional and unconditional guidance streams in a single denoiser call, so the two streams occupy disjoint slices of one per-site buffer rather than one buffer each (Appendix D.3). Each host and its TeleQuant extension share the sampling settings in Table 4, with a backbone-specific FP16 timing reference. Figure 14 reports the same paired overhead across repeated timing rounds.

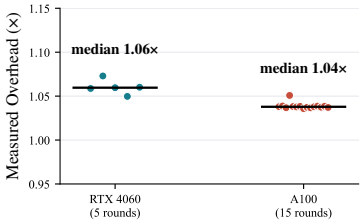


Figure 14: Variation across hardware profiling rounds.

The overhead in Table 13 is small enough that both profiled image pipelines retain most of their frozen-host speedup over FP16. Persistent INT8 buffers occupy 28 MB in the DiT state-precision experiment, whereas peak allocation also depends on activation reuse and workload. Persistent buffer size and pipeline peak memory therefore describe different quantities.

The two timing tiers answer different questions. The operator sweep holds projection width fixed and varies sequence length, isolating the cost of the feedback frontend. The pipeline rows include all invocations and the surrounding generation stages. Within each tier, paired runs hold the host and workload fixed; this is why overhead is normalized to that host. The consumer and datacenter experiments time different workloads, so each tier reports timings within its own GPU class.

Table 13: Deployment latency on consumer and datacenter GPUs. Projection-path rows time one invocation; pipeline rows report latency per generated image. Relative overhead uses the corresponding frozen host as its denominator. Timing and cache-control procedures are specified in Appendix D.6.

Device / Benchmark	**Path / Framework**	**Latency**	**Delta**	**Speedup rel. FP16**	**Relative Overhead**
RTX 4060, 8GB	DiT-XL/2 ViDiT-Q host	8.300 ms	—	—	Baseline
Projection path	DiT-XL/2 + TeleQuant	8.798 ms	+0.498 ms	—	1.06×
A100, 40GB	FP16, Unquantized	3,310.00 ms	—	1.00×	—
E2E, PixArt-Σ	ViDiT-Q host, W4A8	2,720.00 ms	—	1.22×	Baseline
	+ TeleQuant (ours)	2,850.00 ms	+130.00 ms	**1.16×**	1.05×
A100, 40GB	FP16, Unquantized	5,790.00 ms	—	1.00×	—
E2E, SDXL	SVDQuant host, W4A4	4,510.00 ms	—	1.28×	Baseline
	+ TeleQuant (ours)	4,650.00 ms	+140.00 ms	**1.25×**	1.03×

F Additional Analysis

The results in Appendix E describe performance under additional settings. Here we examine how they relate to the proposed mechanism: where error grows along a trajectory, which sites benefit from feedback, how much of a host's fidelity gap is recovered, and how the recurrence composes with existing quantization and temporal methods.

F.1 Error Along the Sampling Trajectory

To locate where trajectories diverge, Figure 15 traces the relative latent drift, the deviation of each intermediate latent from its matched FP16 counterpart, during sampling on DiT-XL/2, PixArt-Σ, and SDXL. The three are drawn solid with circle markers, dashed with square markers, and dotted with diamond markers, respectively, so line style and marker shape carry the same model identity. This locates the observed benefit before image decoding and complements the cross-step diagnostics in Figure 2.

In the displayed nearest-rounding trajectories, accumulated drift grows as sampling proceeds. With TeleQuant, the latent deviation is reduced across the evaluated backbones. The negative mean lag-1 autocorrelation reported in panel b of Figure 2 provides complementary evidence of altered cross-step error structure. Together, these diagnostics are consistent with residual cancellation along the trajectory and show that the observed improvement is already present before VAE decoding.

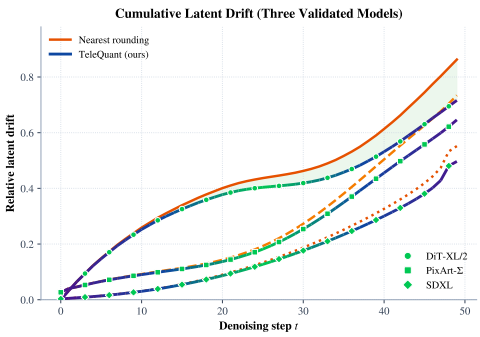


Figure 15: Cumulative latent drift dynamics. Each curve is the relative latent drift of the intermediate latent from its matched FP16 trajectory over sampling steps. TeleQuant lowers the drift trajectory across the evaluated model families.

F.2 Site Selection Diagnostics

We next examine whether all projections benefit equally. The smoothness screen in Section 4.3 uses an undamped Frobenius-norm proxy, while Theorem 1 involves damped variation in the induced $\ell_\infty \to \ell_2$ norm. The inequality $\|M\|_{\infty\to 2} \le \sqrt{d}\|M\|_F$ depends on the input dimension d, so projection rankings need not coincide. The atlas evaluates placement empirically under the calibrated mask.

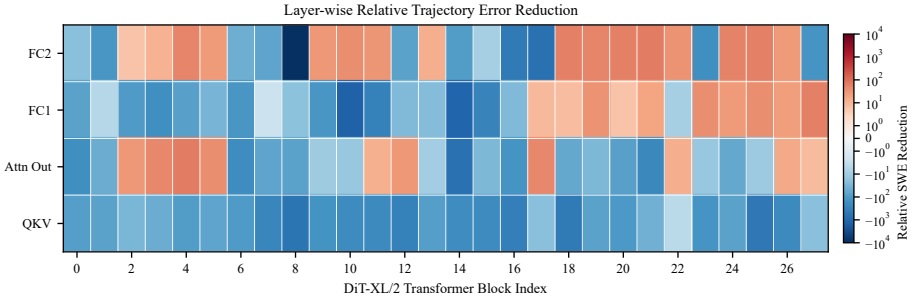


Figure 16: Layer-wise relative trajectory error reduction, DiT-XL/2, 28 Transformer blocks × 4 projection sites. Warm hues: positive measured reduction. Cool hues: increased measured error. The colour scale is the relative reduction in solver-weighted error (SWE) at each site, a signed ratio on a symlog scale rather than a percentage, so zero marks no change. This statistic is the site-wise relative analogue of $\Delta\mathcal{E}_l$ in Equation 7.

Figure 16 reports a DiT-specific relative-error statistic over 28 blocks and four projection sites. Positive responses are concentrated at the Attn Out and FC2 sites, and recur in shallow, middle, and late blocks. An additional band appears at FC1 in blocks 17–27, whereas QKV is negative across all 28 blocks. QKV and FC1 both precede nonlinear operations, yet their measured responses differ. Placement therefore follows measured site behavior rather than projection name. Negative values indicate increased measured error.

F.3 Normalized Gap Recovery Ratio

The preceding diagnostics concern error evolution and site selection. To relate them to endpoint quality, we summarize how much of each frozen host's gap to its FP16 reference is recovered. For host B and metric m,

define

$$\text{Recovery}(B) = \frac{m(B + TeleQuant) - m(B)}{m(\text{FP16}) - m(B)} \times 100\%, \tag{33}$$

where m is oriented so that larger values indicate better quality, for example by negating FID. The ratio requires a finite, nonzero FP16-to-host gap, which holds for the metrics reported below. Table 14 reports recovery for selected image metrics and the six VBench dimensions. DiT distributional recovery uses the published FP16 reference, which was produced under its own sampling configuration rather than the matched 50-step schedule of the host rows, while paired reconstruction uses matched FP16 generations. Datasets, samplers, and precisions differ across backbones, so each ratio describes gap closure within its own setting.

Table 14: Fraction of the FP16-to-host gap recovered using Equation 33. Image metrics use Table 1, with the matched FP16 arm as the reference (SSIM 1.0); video dimensions use Table 2.

Backbone	**Metric**	**Recovery(B) (%)**
DiT-XL/2, ViDiT-Q, W4A8, ImageNet-50K	FID	21.97
	sFID	35.11
	IS	18.83
	SSIM	8.75
PixArt-Σ, ViDiT-Q, W4A8, `coco_1024`	FID	73.73
	SSIM	5.12
SDXL, SVDQuant, W4A4, MJHQ-5K	FID	39.22
	SSIM	24.71
Open-Sora 1.2, video, 251-prompt VBench	Imaging Quality	51.72
	Motion Smoothness	77.97
	Background Consistency	96.67
	Subject Consistency	91.67
	Human Action	55.48
	Overall Consistency	40.74

PixArt-Σ shows the largest FID recovery, and SDXL the largest SSIM recovery. Cross-metric percentages use distinct scales and denominators; the spectral evidence in Figure 2 is reported separately.

F.4 Activation Feedback and Host Composition

The feedback rule targets activation rounding because these tensors change at every denoising evaluation, while the host weights remain fixed. Feedback tracks the activation-rounding residual sequence, and zero gain recovers the host rule in Equation 2. Unselected sites retain that rule but can receive altered upstream inputs. Weight quantization still contributes trajectory-dependent prediction error. A fixed weight error acts on changing activations, so its effect varies over the trajectory as well.

ViDiT-Q transforms activation channels and updates scales; SVDQuant separates high-precision and quantized branches. In SDXL, only the quantized branch receives feedback. The residual is stored in the coordinates of the quantized branch, so the scale ratio compares like quantities at successive evaluations. Compatibility denotes composition with these frozen operations, without assuming statistical independence. Table 1 evaluates the joint system, and Appendix D.2 specifies its integration.

F.5 Relationship to Existing Temporal Methods

The categories in Section 2 distinguish intervention points. Here we examine what can be composed with the recurrence and what requires a different analysis. The evaluated host interfaces are specified in Appendix D.2.

Calibration and fine-tuning modify the quantized model before generation (Liu et al., 2024; Huang et al., 2024a; He et al., 2024). Time-aware transformations and precision policies additionally change the quantizer encountered at each evaluation (Ding et al., 2025; Liu et al., 2026b; Zhang et al., 2026c). These methods can influence temporal correlations. The operative distinction is whether a method explicitly matches successive rounding residuals after calibration. Our gain assumes consistent activation coordinates and accounts for changing scales. A changing basis or feature layout would also require transporting the stored residual between coordinate systems.

Uniform and nonuniform quantizers have different residual bounds. OrbitQuant uses a normalized rotated basis with a shared Lloyd–Max codebook (Lee et al., 2026); LRQ-DiT uses logarithmic weight quantization and adaptive activation rotations (Yang et al., 2026). The half-bin bound in Theorem 1 applies to uniform activation

rounding in normalized coordinates. Nonuniform activation cells require cell-dependent error bounds, while changes to weight representation remain part of the frozen host. The two evaluated integrations cover a W4A8 host and a W4A4 host, spanning both activation precisions.

MoDiff is a close temporal comparison because it also carries quantization information between evaluations (Gao et al., 2025). It quantizes the difference from a reconstructed activation and accumulates the corresponding layer-output update. TeleQuant instead rounds the current host activation after a bounded residual shift. Its gain is chosen to match adjacent contributions to a solver-weighted trajectory surrogate. These recurrences use different state and execution paths. Consequently, unit-delay feedback in Figure 7 serves as a control for our gain.

Caching and sparse attention introduce additional approximation errors. Q&C corrects the distributional effects of combining quantization and caching (Ding et al., 2026); 6Bit-Diffusion caches temporal block differences (Su et al., 2026). QuantSparse restores information omitted by sparse attention using temporal residual correction (Feng et al., 2026). The rounding-state bound covers the activation-rounding residual, while caching and sparse attention add their own errors. Composing them with feedback accounts for skipped or approximated operations and recalibrates the selected sites. The current experiments retain the host's evaluation schedule.

Fast solvers change schedule nodes and prediction weights (Liu et al., 2022; Bao et al., 2022; Zhang & Chen, 2023). TCEC adds solver-level compensation (Liu et al., 2026a). Either change affects the transfer used to define c_t, so the coefficient must be derived for the actual deployed update and prediction parameterization. Appendix A.1 makes this dependence explicit. The comparison in Table 3 tests separate interventions.

Adaptive rounding chooses decisions during calibration (Nagel et al., 2020), while QDrop randomizes quantization during reconstruction (Wei et al., 2022). TeleQuant updates its state during generation. Its connection to classical noise shaping is the reuse of discarded quantization error (Norsworthy et al., 1997; Floyd & Steinberg, 1976). For constant $\alpha > 0$ and white residuals, the scalar power factor is $|\alpha e^{-i\omega} - 1|^2 = 1 + \alpha^2 - 2\alpha \cos \omega$, smaller at low denoising-step frequencies. This stationary example motivates the spectral diagnostic; the theorem handles changing gains and sensitivities. Denoising-step frequency and video-frame frequency are distinct.

The evidence supports different parts of the proposed mechanism. Figure 2 examines error structure, Figure 7 tests the complete gain, and Table 7 tests compression. The placement atlas reports site behavior under the deployed mask, and the calibration count and variation threshold are held fixed across the reported comparisons.