FROM PREDICTABLE DISCREPANCY TO RESIDUAL VARIANCE: CORRECTING QUANTIZED DDIM SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) reduces diffusion-model inference cost, but denoiser-output discrepancies perturb subsequent sampling states. We introduce State-Conditioned Predictive Error Correction (SPEC), a two-stage output corrector for a fixed quantized denoiser under DDIM-family sampling. At sparse refresh steps, its temporal branch learns an output-matching offset for the quantized denoiser's time condition from full-precision supervision on the same latent state; the DDIM grid and transition coefficients remain nominal. A full-resolution branch is then trained on trajectories recollected with this adjustment active to predict the remaining same-state output discrepancy. On held-out trajectories, the Predictable Error Ratio (PER) shows that SPEC reduces the summed squared same-state output discrepancy between the quantized and full-precision denoisers by 82.7%–88.1% across the evaluated bitwidths. For stochastic DDIM, Student-t-estimated variance reallocation (t-VR) models the remaining post-SPEC residual variance and reallocates the nominal stochastic variance between its estimated contribution and the explicit Gaussian innovation. With 100-step stochastic DDIM on CIFAR-10 at W4A8, FID falls from 6.84 for Q-Diffusion to 5.83 for SPEC and 5.66 with t-VR.

1 INTRODUCTION

Diffusion models require repeated denoiser evaluations during sampling Ho et al. (2020); Song et al. (2021a). Post-training quantization (PTQ) reduces this cost with low-precision weights and activations Li et al. (2023), but each denoiser-output discrepancy perturbs the next latent state and thereby changes subsequent denoiser inputs. Correcting that discrepancy at one step therefore also changes the states on which later corrections must operate.

Timestep-aware calibration and multi-step quantizer optimization Li et al. (2023); Lee et al. (2025), along with cross-timestep distillation Li & Du (2025), improve the quantized model. StepbaQ corrects sampling steps from calibrated error statistics Chen et al. (2024b), while PTQD and D^2-DPM correct mean and variance effects using calibrated or Gaussian error statistics He et al. (2023); Zeng et al. (2025). We study a fixed quantized denoiser with the DDIM grid and coefficients fixed: how can its time condition reduce the output discrepancy at the current state, how should the remaining correction be learned on the resulting trajectory, and how should its residual contribution be incorporated into stochastic sampling?

DDIM-family transitions are affine in a single denoiser prediction, so a same-state output discrepancy enters the next update directly. State-Conditioned Predictive Error Correction (SPEC) predicts a sparse temporal-condition offset from the current latent and quantized output, using supervision from a small discrete search that matches the quantized output to the nominal-time full-precision output at the same latent. Only the quantized denoiser's time input changes; the DDIM grid and transition coefficients retain the nominal timestep. A full-resolution branch, trained on trajectories recollected with this adjustment active, predicts the remaining same-state output discrepancy from the current latent and adjusted quantized output. For stochastic DDIM, t-VR then estimates the second-order contribution of the post-SPEC residual and reallocates the nominal variance budget between that contribution and the Gaussian innovation.

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.

On held-out trajectories, Predictable Error Ratio (PER) indicates that SPEC reduces the summed squared output difference between the quantized and full-precision denoisers evaluated on the same latent states; the reduction ranges from 82.7% to 88.1% across the evaluated bitwidths. For stochastic DDIM, Student-t-estimated variance reallocation (t-VR) adjusts the noise budget using modeled post-SPEC residual variance. On CIFAR-10 at W4A8 (4-bit weights and 8-bit activations), Fréchet Inception Distance (FID) falls from 6.84 to 6.04 with temporal adjustment, 5.83 with SPEC, and 5.66 with t-VR; SPEC increases theoretical bit operations (BOPs) by 13.6% in this setting.

Our main contributions are:

- We develop SPEC for a fixed quantized denoiser. At sparse steps, its state-conditioned predictor learns a temporal-condition offset from same-state output matching with a full-precision teacher while keeping the DDIM grid and coefficients nominal. A second branch predicts the remaining full-resolution discrepancy on recollected, temporally adjusted trajectories. Held-out PER measures the combined error reduction.
- For stochastic DDIM, we apply t-VR after SPEC. Rather than modeling the original quantization error, it uses a Student-t working marginal to estimate the post-correction residual's average second-order contribution and reallocates the nominal variance budget between this contribution and the explicit Gaussian innovation. We evaluate SPEC across bitwidths, architectures, and deterministic and stochastic DDIM settings.

2 Related Work

Temporal correction. Quantization can disturb timestep-dependent features; TFMQ-DM addresses this during quantizer reconstruction Huang et al. (2023). Separately, sampling errors can move generated states away from the distribution associated with their nominal timestep Ning et al. (2023); Yu et al. (2026). Time-Shift Sampler selects an effective sampling time from state-variance statistics Li et al. (2024), while StepbaQ interprets quantization-induced latent drift as a stepback and uses calibrated error statistics to correct the sampling step and trajectory Chen et al. (2024b). These methods preserve temporal features or adjust state–time alignment; SPEC instead targets the quantized denoiser's output discrepancy on the same current state. Its state-conditioned time offset is trained against the full-precision prediction and applied at sparse refresh steps. The physical DDIM grid and transition coefficients remain nominal, so the offset controls denoiser output without estimating a new physical diffusion time.

Sampling-time error correction. Diffusion PTQ methods improve the quantized denoiser through multi-timestep reconstruction, temporal-feature preservation, or adaptive quantization Li et al. (2023); Huang et al. (2023); Wang et al. (2024); Huang et al. (2024); Chen et al. (2024a); Hoang et al. (2026). AccuQuant simulates multiple denoising steps during quantizer optimization Lee et al. (2025), whereas CTEC distills compensation for previous-step errors into the quantized model Li & Du (2025). Sampling-time approaches instead act on a deployed quantized denoiser. TAC applies timestep-channel reconstruction and input-bias compensation, and QNCD estimates inter-step noise at runtime Yao et al. (2024); Chu et al. (2024). D^2-DPM estimates an affine conditional error mean from the current quantized output under timestep-wise Gaussian statistics Zeng et al. (2025). Q-Drift uses calibrated drift statistics, while IEC iteratively refines predictions at test time Ryu et al. (2026); Zhong et al. (2026). Unlike an output-only affine estimate, SPEC conditions its full-resolution post-alignment corrector on both the current latent and adjusted quantized output, allowing state- and coordinate-dependent corrections. Training it on trajectories recollected after temporal adjustment also matches the states induced by that branch.

Residual-variance calibration for stochastic sampling. Quantization error can alter both the mean and the effective variance of a reverse diffusion transition. PTQD decomposes quantization noise into correlated and uncorrelated components, corrects the associated bias, and uses Variance Schedule Calibration to absorb the estimated variance of the uncorrelated component into the diffusion variance He et al. (2023). D^2-DPM likewise separates mean and variance deviations induced by quantization and corrects them using timestep-wise Gaussian statistics Zeng et al. (2025). These calibrations characterize quantization-induced error, whereas the variance relevant after SPEC is that of the remaining post-correction residual. Our t-VR fits this residual on corrected trajectories with

a zero-location Student-t working marginal, motivated by its observed heavy tails. It estimates the residual's average second-order contribution and reduces the explicit Gaussian innovation accordingly, keeping their modeled contributions within the original stochastic sampling budget.

3 METHOD

3.1 PROBLEM FORMULATION AND PREDICTABLE ERROR DECOMPOSITION

We develop SPEC for deterministic and stochastic generalized DDIM. For a transition from timestep t to $s < t$, the reverse update has the affine form

$$\boldsymbol{x}_s = A_{t,s}\boldsymbol{x}_t + B_{t,s}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \sigma_{t,s}\boldsymbol{z}_t, \qquad \boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

where $A_{t,s}$, $B_{t,s}$, and $\sigma_{t,s}$ are the standard generalized-DDIM coefficients given in Appendix A.1, Eq. (A.6). In particular, $B_{t,s}$ determines how the denoiser prediction affects the subsequent latent state.

To distinguish the instantaneous discrepancy between the quantized and full-precision denoisers from differences caused by evaluating them on different sampling trajectories, we evaluate both models on the same latent state $\boldsymbol{x}_t^q$. We obtain this state by running the quantized sampler recursively: each step uses the state produced by the preceding step. The two denoiser predictions are

$$\boldsymbol{\epsilon}_t^q = Q_\theta(\boldsymbol{x}_t^q, t), \qquad \boldsymbol{\epsilon}_t^{\mathrm{fp}|q} = F_\theta(\boldsymbol{x}_t^q, t), \qquad \boldsymbol{e}_t = \boldsymbol{\epsilon}_t^q - \boldsymbol{\epsilon}_t^{\mathrm{fp}|q}. \tag{2}$$

We refer to $\boldsymbol{e}_t$ as the *same-input quantization discrepancy*. Substituting $\boldsymbol{\epsilon}_t^q = \boldsymbol{\epsilon}_t^{\mathrm{fp}|q} + \boldsymbol{e}_t$ into Eq. (1) introduces the additional term $B_{t,s}\boldsymbol{e}_t$ into the next-state update. Throughout, we use $\boldsymbol{c}_t = (\boldsymbol{x}_t, Q_\theta(\boldsymbol{x}_t, t), t)$ to denote the observable condition at whichever current sampling state $\boldsymbol{x}_t$ is under consideration.

With the quantized backbone fixed, we learn a lightweight correction predictor under a prescribed architecture, inference precision, and temporal-refresh schedule to control its inference overhead. We define the post-correction residual as the discrepancy remaining after subtracting the learned correction, yielding

$$\boldsymbol{e}_t = \widehat{\boldsymbol{m}}_\omega(\boldsymbol{c}_t) + \boldsymbol{r}_t, \tag{3}$$

where $\widehat{\boldsymbol{m}}_\omega(\boldsymbol{c}_t)$ is the learned state-conditioned correction and $\boldsymbol{r}_t$ is its *post-correction residual*.

We measure the empirical reduction using the *Predictable Error Ratio* (PER). For M state–timestep pairs with discrepancies $\boldsymbol{e}_i$ and residuals $\boldsymbol{r}_i$, we compute

$$\widehat{\mathrm{PER}} = 1 - \frac{\sum_{i=1}^{M} \|\boldsymbol{r}_i\|_2^2}{\sum_{i=1}^{M} \|\boldsymbol{e}_i\|_2^2}. \tag{4}$$

PER measures aggregate discrepancy-energy reduction, pooling energies across independently generated complete-SPEC test trajectories with the predictor fixed (Appendix B.1). We call discrepancy predictable to the extent that the trained corrector reduces its energy on held-out states; this does not assume exact recovery of $\boldsymbol{e}_t$. Figure 1 shows PER of 82.7%–88.1% across the evaluated bitwidths, while the discrepancy energy relative to the full-precision prediction increases from 13.1% under W8A16 to 54.0% under W4A6. Thus, predictive correction removes most observed discrepancy energy even under aggressive quantization; the remaining residual motivates variance modeling.

3.2 STATE-CONDITIONED PREDICTIVE ERROR CORRECTION (SPEC)

We parameterize the structured correction $\widehat{\boldsymbol{m}}_\omega(\boldsymbol{c}_t)$ in Eq. (3) using two sequential operations: state-dependent temporal-condition adjustment followed by full-resolution error prediction.

Given an effective temporal condition $\widetilde{t}$, the quantized denoiser produces the temporally adjusted prediction

$$\widetilde{\boldsymbol{\epsilon}}_t^q = Q_\theta(\boldsymbol{x}_t, \widetilde{t}). \tag{5}$$

The error branch then estimates the remaining structured discrepancy $\widehat{\boldsymbol{e}}_t$, yielding

$$\boldsymbol{\epsilon}_t^{\mathrm{corr}} = \widetilde{\boldsymbol{\epsilon}}_t^q - \widehat{\boldsymbol{e}}_t. \tag{6}$$

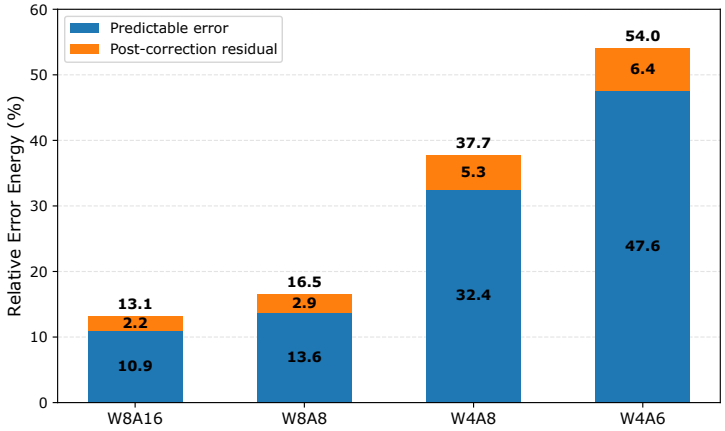


Figure 1: Empirical reduction of the same-input discrepancy by the frozen SPEC predictor on independently generated test trajectories across bitwidth settings. The total bar height reports the summed squared norm of the same-input discrepancy, normalized by the corresponding summed squared norm of the full-precision prediction. PER partitions this total into the portion removed by the trained corrector and the post-correction residual.

Accordingly, the total state-conditioned correction relative to the nominal quantized prediction is

$$\widehat{\boldsymbol{m}}_\omega(\boldsymbol{c}_t) = Q_\theta(\boldsymbol{x}_t, t) - Q_\theta(\boldsymbol{x}_t, \widetilde{t}) + \widehat{\boldsymbol{e}}_t. \tag{7}$$

The first term moves the prediction along the quantized denoiser's response to its time condition. Because this control is a single scalar, the second term predicts the remaining output-tensor discrepancy elementwise.

Both corrections are implemented by a unified lightweight predictor P_ω with a shared lightweight encoder and two output branches. The temporal branch P^t_ω predicts a state-dependent temporal offset, while the error branch P^e_ω estimates the remaining full-resolution discrepancy.

Temporal-Condition Prediction. For a state $\boldsymbol{x}^q_t$ collected from the original quantized trajectory, the oracle temporal offset obtained by a small discrete search around the nominal timestep is

$$\delta^\star_t = \arg\min_{\delta \in \Delta_t} \ \mathrm{SmoothL1}\left(Q_\theta(\boldsymbol{x}^q_t, t+\delta), F_\theta(\boldsymbol{x}^q_t, t)\right), \tag{8}$$

where $\Delta_t \subset \mathbb{Z}$ denotes a local integer-offset candidate set. Only the temporal condition of the quantized denoiser is varied during the search; the full-precision teacher remains evaluated at the nominal timestep t. Thus, $\delta^\star_t$ is an output-matching target for the quantized denoiser, rather than an estimate of the state's physical diffusion time.

The temporal branch predicts $\widehat{\delta}_t = P^t_\omega(\boldsymbol{c}_t)$ and is trained using

$$\mathcal{L}_{\mathrm{time}} = \mathbb{E}_{\boldsymbol{c}_t}\left[\left(\widehat{\delta}_t - \delta^\star_t\right)^2\right]. \tag{9}$$

Reevaluating the quantized denoiser with an adjusted temporal condition at every reverse step would introduce substantial additional cost. We therefore perform temporal refresh only at a sparse subset $\mathcal{R}$ of timesteps. At inference, the predicted offset is rounded to the nearest integer and clipped to the valid timestep range before forming $\widetilde{t}$. For $t \in \mathcal{R}$, we use the resulting adjusted condition; otherwise, we set $\widetilde{t} = t$ and reuse the nominal quantized prediction. The candidate set Δ_t, clipping range, and refresh schedule $\mathcal{R}$ are detailed in Appendix B.1.

Full-Resolution Post-Alignment Discrepancy Prediction. Temporal-condition adjustment uses one scalar offset to change the denoiser's response to its time input. It cannot represent an arbitrary residual tensor, so the second branch predicts the remaining full-resolution post-alignment discrepancy as

$$\widehat{\boldsymbol{e}}_t = P^e_\omega\left(\boldsymbol{x}_t, \widetilde{\boldsymbol{\epsilon}}^q_t, \widetilde{t}\right), \tag{10}$$

where $\boldsymbol{x}_t$ denotes the current sampling state. The prediction $\widehat{\boldsymbol{e}}_t \in \mathbb{R}^{C\times H\times W}$ has the same shape as the denoiser output and supplies a separate correction for each channel and latent position.

Trajectory-Aware Staged Training. Temporal adjustment changes subsequent sampling states. We therefore train the branches sequentially, recollecting temporally adjusted trajectories before training the error branch on the states it will encounter.

In Stage I, we run the frozen quantized sampler and collect states $\boldsymbol{x}_t^q$ from the original quantized sampling trajectory. For each state, the observable condition $\boldsymbol{c}_t$ is instantiated using $\boldsymbol{x}_t = \boldsymbol{x}_t^q$. The temporal branch is optimized using Eq. (9) with the oracle offset $\delta_t^\star$ from Eq. (8), while the error branch remains frozen.

After Stage I, we freeze both the temporal branch and the shared encoder, enable temporal-condition adjustment, and rerun the sampler to collect a new trajectory. Let $\boldsymbol{x}_t^a$ denote a state from the resulting temporally adjusted trajectory. On the same state, we evaluate

$$\widetilde{\boldsymbol{\epsilon}}_t^q = Q_\theta(\boldsymbol{x}_t^a, \widetilde{t}), \quad \boldsymbol{\epsilon}_t^{\mathrm{fp}|a} = F_\theta(\boldsymbol{x}_t^a, t). \quad . \tag{11}$$

The full-precision teacher remains evaluated at the nominal timestep t, because $\widetilde{t}$ adjusts only the temporal condition of the quantized denoiser. The same-state discrepancy remaining after temporal-condition adjustment is therefore

$$\boldsymbol{e}_t^a = \widetilde{\boldsymbol{\epsilon}}_t^q - \boldsymbol{\epsilon}_t^{\mathrm{fp}|a}. \tag{12}$$

In Stage II, only the error branch is optimized to predict $\boldsymbol{e}_t^a$. Since $\boldsymbol{\epsilon}_t^{\mathrm{corr}} = \widetilde{\boldsymbol{\epsilon}}_t^q - \widehat{\boldsymbol{e}}_t$, direct error supervision is equivalent to reconstructing the full-precision prediction, yielding

$$\mathcal{L}_{\mathrm{err}} = \|\widehat{\boldsymbol{e}}_t - \boldsymbol{e}_t^a\|_2^2 = \left\|\boldsymbol{\epsilon}_t^{\mathrm{corr}} - \boldsymbol{\epsilon}_t^{\mathrm{fp}|a}\right\|_2^2. \tag{13}$$

Under the ϵ-parameterization, the noise prediction is proportional to the score at a fixed timestep Song et al. (2021b). We therefore complement elementwise error reconstruction with directional and magnitude consistency, defined by

$$\mathcal{L}_{\mathrm{dir}} = 1 - \frac{\left\langle \boldsymbol{\epsilon}_t^{\mathrm{corr}}, \boldsymbol{\epsilon}_t^{\mathrm{fp}|a} \right\rangle}{\|\boldsymbol{\epsilon}_t^{\mathrm{corr}}\|_2 \left\|\boldsymbol{\epsilon}_t^{\mathrm{fp}|a}\right\|_2 + \varepsilon}, \qquad \mathcal{L}_{\mathrm{mag}} = \left(\frac{\|\boldsymbol{\epsilon}_t^{\mathrm{corr}}\|_2}{\left\|\boldsymbol{\epsilon}_t^{\mathrm{fp}|a}\right\|_2 + \varepsilon} - 1 \right)^2. \tag{14}$$

Here, $\varepsilon > 0$ is a small numerical-stability constant. The complete Stage-II objective is $\mathcal{L}_{\mathrm{corr}} = \mathcal{L}_{\mathrm{err}} + \lambda_{\mathrm{geo}}(\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{mag}})$, with $\lambda_{\mathrm{geo}} = 0.5$ by default.

Under common-noise coupling and a local Lipschitz condition, the endpoint RMS state deviation from the full-precision sampler has an upper bound determined by the per-step RMS prediction discrepancies along each sampler's own trajectory. This bound is smaller for complete SPEC if its per-step discrepancies are no greater than those of the uncorrected quantized sampler at every transition, with a strict reduction at any one transition. Appendix A.2 derives this sufficient condition and reports supporting estimates from held-out W4A8 trajectories.

Lightweight implementation. The two branches share a compact convolutional encoder composed of three 3×3 convolution–GroupNorm–SiLU blocks with 32 channels. The temporal branch applies global average pooling followed by a two-layer MLP to predict a scalar temporal offset, while the error branch uses two 1×1 convolutions to produce a full-resolution correction tensor. The shared encoder contains only 27.9K parameters. Alternative encoder designs and the accuracy–efficiency comparison are reported in Table C-2 of Appendix C.1. Detailed training settings are given in Appendix B.1.

3.3 Student-t-Estimated Variance Reallocation (t-VR) for Post-Correction Residuals

After Stage II, we freeze both SPEC branches and recollect complete-SPEC trajectories under the original Gaussian noise schedule. For a state $\boldsymbol{x}_t^c$ visited before variance reallocation, we define the same-state post-correction residual as

$$\boldsymbol{r}_t^c = \boldsymbol{\epsilon}_t^{\mathrm{corr}}(\boldsymbol{x}_t^c) - F_\theta(\boldsymbol{x}_t^c, t). \tag{15}$$

Here, $\boldsymbol{\epsilon}_t^{\text{corr}}(\boldsymbol{x}_t^c)$ denotes the complete SPEC prediction at $\boldsymbol{x}_t^c$ and nominal timestep t. These residuals are used to fit timestep-wise variance estimates, which are held fixed during SPEC+t-VR sampling.

Under the oracle full-precision reference used in our second-order analysis, which assumes zero full-precision prediction error relative to the forward noise, generalized DDIM couples its prediction-direction and Gaussian coefficients so that their variances recover the prescribed conditional marginal variance. A centered post-SPEC residual with average per-coordinate variance v_t adds $B_{t,s}^2 v_t$ beyond the original Gaussian budget $\sigma_{t,s}^2$. We use this budget as the reference for one-step second-order calibration; Appendix A.3 relates it to the complete generalized-DDIM marginal.

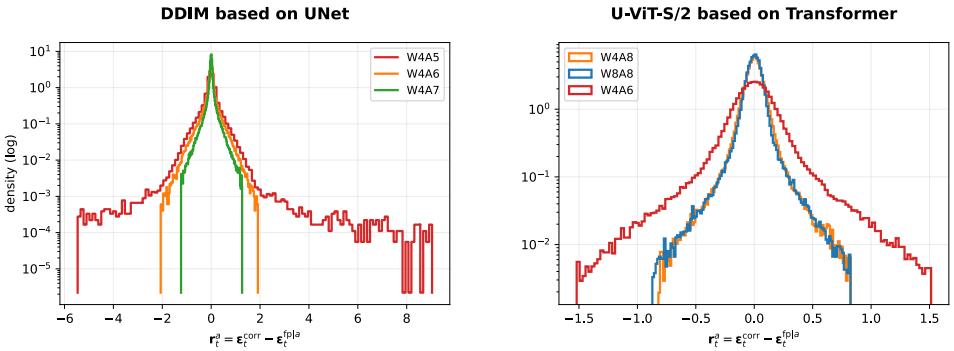


Figure 2: Empirical densities of the post-correction residual on a logarithmic scale for DDIM with a U-Net backbone (left) and U-ViT-S/2 with a Transformer backbone (right).

Figure 2 shows that post-correction residuals concentrate near zero and retain substantial tails across both architectures. Pooling residuals with state- and coordinate-dependent scales can produce such heavy tails. Appendix A.3 discusses this motivation, the aggregate residual-mean diagnostic, and fitted-parameter stability. We use a zero-location Student-t working marginal to estimate v_t as

$$r_{t,k}^c \sim t_{\nu_t}(0, s_t), \qquad \nu_t > 2,\ s_t > 0, \tag{16}$$

where k indexes coordinates, ν_t controls tail heaviness, and s_t is the scale. The fitted marginal yields the scalar variance estimate

$$\widehat{v}_t = \frac{\widehat{\nu}_t}{\widehat{\nu}_t - 2}\widehat{s}_t^2. \tag{17}$$

Table 5 compares this estimate with the empirical second moment under the same variance-reallocation rule.

For a stochastic generalized-DDIM transition, we first hold the nominal prediction coefficient $B_{t,s}$ fixed and subtract the modeled residual contribution from the Gaussian variance budget:

$$\widetilde{\sigma}_{t,s}^2 = \sigma_{t,s}^2 - B_{t,s}^2 \widehat{v}_t, \tag{18}$$

provided that $B_{t,s}^2 \widehat{v}_t \le \sigma_{t,s}^2$. We then update the prediction coefficient using the generalized-DDIM relation

$$\widetilde{B}_{t,s} = \sqrt{1 - \bar{\alpha}_s - \widetilde{\sigma}_{t,s}^2} - A_{t,s}\sqrt{1 - \bar{\alpha}_t}, \qquad A_{t,s} = \sqrt{\frac{\bar{\alpha}_s}{\bar{\alpha}_t}}. \tag{19}$$

These sequential coefficient updates require no joint solve; Appendix A.3 analyzes their conditional covariance under idealized assumptions.

For the final transition, we set $\eta = 0$ and skip t-VR, yielding $\widetilde{\sigma}_{t,s} = \sigma_{t,s} = 0$ and $\widetilde{B}_{t,s} = B_{t,s}$. The resulting SPEC+$t$-VR transition is then given by

$$\boldsymbol{x}_s = A_{t,s}\boldsymbol{x}_t + \widetilde{B}_{t,s}\boldsymbol{\epsilon}_t^{\text{corr}} + \widetilde{\sigma}_{t,s}\boldsymbol{z}_t, \qquad \boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{20}$$

The resulting coefficients remain within the generalized-DDIM parameterization, while the modeled residual and the remaining Gaussian innovation share the original one-step stochastic variance budget.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate the proposed method on both pixel- and latent-space diffusion models. For pixel-space generation, we use pretrained models on CIFAR-10 (32×32) Krizhevsky & Hinton (2009) and CelebA (64×64) Liu et al. (2015) with 100-step DDIM sampling Song et al. (2021a), generating 50,000 images per setting. Main results use deterministic sampling ($\eta = 0$), while CIFAR-10 is additionally evaluated with $\eta = 1$ to study residual-aware variance reallocation. We further evaluate U-ViT-S/2 Bao et al. (2023) on CIFAR-10 to examine transferability to a Transformer-based denoiser. For latent-space generation, we use LDM-4 on LSUN-Bedroom and LDM-8 on LSUN-Church at 256×256 resolution Rombach et al. (2022); Yu et al. (2015), generating 3,000 images per setting. We additionally evaluate class-conditional LDM-4 on ImageNet at 256×256 resolution using 20-step DDIM with $\eta = 1$ and classifier-free guidance (CFG) scale 3.0. Each ImageNet setting generates 10,000 images, with 10 images per class across all 1,000 classes. We report FID Heusel et al. (2017) for all datasets and additionally report Inception Score (IS) Salimans et al. (2016) for CIFAR-10 and CelebA.

We denote X-bit weight and Y-bit activation quantization as WXAY. Main experiments consider W8A8 and W4A8, with W4A5–W4A7 used to characterize the low-bit operating boundary. Detailed quantization, predictor training, inference, residual calibration, and reproducibility settings are provided in Appendices B.1, B.1, and B.2.

Result provenance and comparison protocol. Results marked with † are reproduced under our unified evaluation protocol; ‡ denotes values reported in the original papers and provides literature context rather than a controlled ranking. All FP32, Q-Diffusion, and D^2-DPM entries are reproduced by us. Claims of matched improvement refer to Q-Diffusion evaluated with the same backbone, sampler, quantization setting, and metric pipeline. Appendix B.3 reports the metric definitions and sample counts.

4.2 COMPARATIVE RESULTS

Table 1: Quantization results for 100-step DDIM ($\eta = 0$) on CIFAR-10 (32×32) and CelebA (64×64).

Dataset	Method	Bitwidth (W/A)	FID	IS
CIFAR-10	FP32†	32/32	4.22	9.12
	Ours	8/8	**3.58**	9.27
	TAC Yao et al. (2024)‡	8/8	3.68	**9.49**
	Q-Diffusion Li et al. (2023)†	8/8	3.80	9.36
	TDQ (PTQ) So et al. (2023)‡	8/8	5.99	8.85
	QNCD Chu et al. (2024)‡	8/8	3.61	8.42
	EDA-DM Liu et al. (2024)‡	8/8	3.73	9.40
	Ours	4/8	4.81 ± 0.03	9.10
	TAC‡	4/8	4.89	9.15
	Q-Diffusion†	4/8	4.89	9.12
	QNCD‡	4/8	4.92	**9.61**
	EDA-DM‡	4/8	**4.03**	9.43
CelebA	FP32†	32/32	17.26	2.61
	Ours	8/8	**19.37**	**2.60**
	Q-Diffusion†	8/8	21.84	2.55
	Ours	4/8	**23.61**	**2.74**
	Q-Diffusion†	4/8	25.23	2.70

Table 4: Ablation and theoretical deployment cost across stochastic sampling settings.

Setting	Method	W/A	FID	Size (MB)	BOPs/step
DDIM on CIFAR-10 100 steps,$\eta = 1$	FP32†	32/32	5.78	142.82	6597
	Q-Diffusion†	8/8	4.31	35.76	798
	Temporal-Condition Prediction	8/8	4.23	35.87	863.6
	SPEC	8/8	4.19	35.87	884.1
	SPEC + t-VR	8/8	**4.05**	35.87	884.1
	Q-Diffusion†	4/8	6.84	17.83	399
	Temporal-Condition Prediction	4/8	6.04	17.94	432.7
	SPEC	4/8	5.83	17.94	453.2
	SPEC + t-VR	4/8	**5.66**	17.94	453.2
DDIM on CIFAR-10 100 steps,$\eta = 0.5$	FP32†	32/32	4.51	142.82	6597
	Q-Diffusion†	8/8	3.97	35.76	798
	SPEC + t-VR	8/8	$\mathbf{3.90 \pm 0.02}$	35.87	884.1
	Q-Diffusion†	4/8	5.55	17.83	399
	SPEC + t-VR	4/8	$\mathbf{5.49 \pm 0.05}$	17.94	453.2
LDM-4 on LSUN-Bedroom 200 steps,$\eta = 1$	FP32†	32/32	8.35	1096.2	107.17
	Q-Diffusion†	8/8	10.78	274.1	12.97
	D^2-DPM Zeng et al. (2025)†	8/8	10.71	274.1	12.97
	SPEC	8/8	**10.63**	274.2	13.58
	SPEC + t-VR	8/8	10.65	274.2	13.58
	Q-Diffusion†	4/8	11.03	137.0	6.48
	D^2-DPM†	4/8	11.60	137.0	6.48
	SPEC	4/8	10.87	137.1	6.83
	SPEC + t-VR	4/8	**10.76**	137.1	6.83
LDM-4 on ImageNet, 20 steps, $\eta = 1$	FP32†	32/32	17.56	1603.7	102.2
	Q-Diffusion†	4/8	23.69	400.9	3.19
	SPEC + t-VR	4/8	**19.67**	401.9	3.67

Per-nominal-step BOPs include the amortized predictor and eight refresh denoiser calls (GBOPs for DDIM; TBOPs for LDM-4). CIFAR-10 $\eta = 0.5$ FIDs are mean $\pm$ sample standard deviation over three seeds (Appendix C.1).

Table 2: 100-step U-ViT-S/2 ($\eta = 1$) results on CIFAR-10.

Method	W8A8		W4A8		W4A6
	FID	IS	FID	IS	FID
FP32†	4.09	9.30	4.09	9.30	4.09
Ours	**5.70**	**9.17**	**11.88**	**9.08**	**113.3**
Q-Diffusion†	7.00	9.01	14.80	8.89	130.7

Table 3: 400-step LDM-8 ($\eta = 0$) results on LSUN-Church.

Method	W8A8	W4A8
	FID	FID
FP32†	7.33	7.33
Q-Diffusion†	**7.50**	8.17
Ours	7.56	**7.90**

Under 100-step deterministic DDIM, Table 1 shows that SPEC lowers matched Q-Diffusion FID from 3.80 to 3.58 (W8A8) and 4.89 to 4.81 (W4A8) on CIFAR-10, and from 21.84 to 19.37 and 25.23 to 23.61 on CelebA. IS remains comparable. CelebA W4A8 has higher IS than FP32 but higher FID, so a quantization gap remains.

On the Transformer-based U-ViT-S/2 backbone (Table 2), SPEC lowers FID from 7.00 to 5.70 (W8A8) and 14.80 to 11.88 (W4A8) relative to Q-Diffusion, with IS gains in both settings. These results extend the observed benefit beyond the U-Net backbone; W4A8 retains a substantial gap to FP32.

On LSUN-Church with 400-step deterministic LDM-8 sampling (Table 3), SPEC improves W4A8 FID from 8.17 to 7.90 but trails Q-Diffusion under W8A8 (7.56 versus 7.50).

4.3 Ablation and Deployment Efficiency

Table 4 separates the effects of SPEC and t-VR under stochastic sampling and reports model size and theoretical BOPs. The objective ablation is in Appendix C.1.

On CIFAR-10 W4A8, the full-resolution branch lowers FID from 6.04 to 5.83 after temporal correction, and t-VR lowers it to 5.66. At W8A8, the corresponding CIFAR-10 FIDs fall from 4.23 to 4.19 and then 4.05. On LSUN-Bedroom, t-VR improves W4A8 FID from 10.87 to 10.76 but changes W8A8 FID from 10.63 to 10.65. ImageNet W4A8 FID falls from 23.69 to 19.67 relative to Q-Diffusion.

At intermediate stochasticity ($\eta = 0.5$), SPEC+t-VR also lowers CIFAR-10 FID from 3.97 to 3.90 ± 0.02 at W8A8 and from 5.55 to 5.49 ± 0.05 at W4A8. The gain therefore persists beyond the $\eta = 1$ setting, although it is smaller than in the corresponding $\eta = 1$ comparisons.

The sequential CIFAR-10 gains distinguish the error branch from variance reallocation. The LSUN-Bedroom W8A8 reversal shows that the latter is not uniformly beneficial, even where SPEC alone improves FID. The 20-step ImageNet result extends the observed improvement to latent-space sampling with fewer reverse steps. SPEC BOP overhead, including temporal refresh, is 13.6% on CIFAR-10 W4A8 and 5.4% on LSUN-Bedroom W4A8; t-VR adds negligible arithmetic cost. Offline costs are in Appendix B.2.

4.4 Residual-Variance Estimation and Operating Range

Table 5: Post-SPEC residual-variance estimation on CIFAR-10 with 100-step DDIM ($\eta = 1$).

Method	W4A5	W4A6	W4A7
	FID ↓	FID ↓	FID ↓
Q-Diffusion†	76.32	29.52	10.52
SPEC + Empirical-VR†	78.25	25.13	9.31
SPEC + t-VR	**60.35**	**19.42**	**6.45**

With the same residuals and reallocation rule, t-VR improves on the Gaussian-fit Empirical-VR He et al. (2023) by 2.86/5.71/17.90 FID at W4A7/W4A6/W4A5. The Gaussian fit improves W4A6 and W4A7 over Q-Diffusion but worsens W4A5 (78.25 versus 76.32). Student-t estimation gives the lowest FID at each bitwidth, with the largest absolute advantage over the Gaussian fit at W4A5. Even so, W4A5 remains unreliable (60.35 FID); fitted-parameter diagnostics are in Appendix A.3.

5 Conclusion

This work treats quantized-denoiser error as a state-conditioned output discrepancy rather than a fixed calibration offset. SPEC combines sparse temporal adjustment with full-resolution residual prediction; on held-out trajectories, it reduces the summed squared same-state discrepancy by 82.7–88.1% across the evaluated bitwidths. For stochastic DDIM, Student-t variance reallocation further improves FID in most matched settings, although its benefit is not uniform. The two stages thus target distinct quantities in DDIM sampling: reducible same-state discrepancy and post-correction residual variance. Very low activation precision and third-order DPM-Solver++ transfer remain challenging (Appendix C.2).

AI Use Statement

Generative AI assisted with research framing, related-work analysis, feedback on method exposition and theoretical reasoning, proof writing and checking, interpretation and presentation of existing experimental results, manuscript editing, and LaTeX formatting. The authors reviewed the AI-assisted material and take responsibility for all claims, results, and the final manuscript.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

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. *arXiv preprint arXiv:2406.17343*, 2024a.

Yi-Chung Chen, Zhi-Kai Huang, and Jing-Ren Chen. StepbaQ: Stepping backward as correction for quantized diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-1712.

Huanpeng Chu, Wei Wu, Chengjie Zang, and Kun Yuan. QNCD: Quantization noise correction for diffusion models. *arXiv preprint arXiv:2403.19140*, 2024.

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*, volume 36, 2023.

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*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Dung Anh Hoang, Cuong Pham, Trung Le, Jianfei Cai, and Thanh-Toan Do. Gradient-aligned calibration for post-training quantization of diffusion models. *arXiv preprint arXiv:2602.01289*, 2026.

Haocheng Huang, Jiaxin Chen, Jinyang Guo, Ruiyi Zhan, and Yunhong Wang. TCAQ-DM: Timestep-channel adaptive quantization for diffusion models. *arXiv preprint arXiv:2412.16700*, 2024.

Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. TFMQ-DM: Temporal feature maintenance quantization for diffusion models. *arXiv preprint arXiv:2311.16503*, 2023.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Seunghoon Lee, Jeongwoo Choi, Byunggwan Son, Jaehyeon Moon, Jeimin Jeon, and Bumsub Ham. AccuQuant: Simulating multiple denoising steps for quantizing diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3679.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *International Conference on Learning Representations*, 2024.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-Diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023.

Yanxi Li and Chengbin Du. Optimizing quantized diffusion models via distillation with cross-timestep error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18530–18538, 2025. doi: 10.1609/aaai.v39i17.34039.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sooyoung Ryu, Mathieu Salzmann, and Saqib Javed. Q-Drift: Quantization-aware drift correction for diffusion model sampling. *arXiv preprint arXiv:2603.18095*, 2026.

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*, 2016.

Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

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

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate post-training quantization for diffusion models. *arXiv preprint arXiv:2305.18723*, 2024.

Yuzhe Yao, Feng Tian, Jun Chen, Haonan Lin, Guang Dai, Yong Liu, and Jingdong Wang. Timestep-aware correction for quantized diffusion models. In *European Conference on Computer Vision*, pp. 215–232, 2024.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Meng Yu, Lei Sun, Jianhao Zeng, Xiangxiang Chu, and Kun Zhan. Elucidating the SNR–t bias of diffusion probabilistic models. *arXiv preprint arXiv:2604.16044*, 2026.

Qian Zeng, Jie Song, Han Zheng, Hao Jiang, and Mingli Song. D^2-DPM: Dual denoising for quantized diffusion probabilistic models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9806–9814, 2025. doi: 10.1609/aaai.v39i9.33063.

Yunshan Zhong, Weiqi Yan, and Yuxin Zhang. Test-time iterative error correction for efficient diffusion models. In *International Conference on Learning Representations*, 2026. arXiv:2511.06250.

A Mathematical Details

We give the generalized-DDIM coefficients, the trajectory-error analysis, and the one-step variance-reallocation derivation used in the main text.

A.1 Generalized DDIM Affine Form

Let

$$\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i). \tag{A.1}$$

For a transition from timestep t to $s < t$, generalized DDIM first estimates the clean sample as

$$\widehat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}}, \tag{A.2}$$

and performs the reverse update

$$\boldsymbol{x}_s = \sqrt{\bar{\alpha}_s}\,\widehat{\boldsymbol{x}}_0 + \sqrt{1-\bar{\alpha}_s-\sigma_{t,s}^2}\,\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \sigma_{t,s}\boldsymbol{z}_t, \qquad \boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{A.3}$$

where

$$\sigma_{t,s} = \eta\sqrt{\frac{1-\bar{\alpha}_s}{1-\bar{\alpha}_t}}\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_s}}. \tag{A.4}$$

Here, $\eta = 0$ gives deterministic DDIM, whereas $\eta > 0$ introduces stochasticity.

Substituting Eq. (A.2) into Eq. (A.3) yields

$$\begin{aligned}\boldsymbol{x}_s = &\sqrt{\frac{\bar{\alpha}_s}{\bar{\alpha}_t}}\,\boldsymbol{x}_t \\ &+ \left(\sqrt{1-\bar{\alpha}_s-\sigma_{t,s}^2} - \sqrt{\frac{\bar{\alpha}_s}{\bar{\alpha}_t}}\sqrt{1-\bar{\alpha}_t}\right)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \sigma_{t,s}\boldsymbol{z}_t.\end{aligned} \tag{A.5}$$

Therefore, the coefficients in Eq. (1) are

$$A_{t,s} = \sqrt{\frac{\bar{\alpha}_s}{\bar{\alpha}_t}}, \qquad B_{t,s} = \sqrt{1-\bar{\alpha}_s-\sigma_{t,s}^2} - A_{t,s}\sqrt{1-\bar{\alpha}_t}. \tag{A.6}$$

For later use, write the standard prediction coefficient as a function of the Gaussian-noise scale u gives

$$B_{t,s}^{\mathrm{DDIM}}(u) = \sqrt{1-\bar{\alpha}_s-u^2} - A_{t,s}\sqrt{1-\bar{\alpha}_t}, \qquad B_{t,s} = B_{t,s}^{\mathrm{DDIM}}(\sigma_{t,s}). \tag{A.7}$$

During t-VR, $A_{t,s}$ retains its original value, while $\widetilde{\sigma}_{t,s}$ is obtained explicitly from Eq. (18). The prediction coefficient is then recomputed as $\widetilde{B}_{t,s} = B_{t,s}^{\mathrm{DDIM}}(\widetilde{\sigma}_{t,s})$, as specified in Eq. (19). Thus, the Gaussian coefficient is adjusted first and the prediction coefficient is updated afterward within the generalized-DDIM parameterization.

A.2 SPEC Trajectory-Error Propagation

This section connects the per-step same-input discrepancy to the accumulated state deviation along generalized-DDIM sampling by comparing analytical endpoint bounds under a common-noise coupling. We analyze SPEC without t-VR so that all coupled trajectories use the same Gaussian coefficient. The residual-variance analysis is given separately in Appendix A.3.

Consider a reverse sequence $t_K > t_{K-1} > \cdots > t_1 > t_0 = 0$, where K is the number of reverse transitions in the chosen sampling schedule. We abbreviate the generalized-DDIM coefficients for $t_k \to t_{k-1}$ by A_k, B_k, and σ_k. Their explicit forms are given in Appendix A.1. We couple the full-precision, uncorrected quantized, and SPEC-corrected samplers using the same initial noise and the same innovation $\boldsymbol{z}_k$ at every transition.

Here, $m = q$ denotes the uncorrected quantized trajectory and $m = c$ denotes the SPEC-corrected trajectory. Define the realized per-step discrepancies along these trajectories by

$$\boldsymbol{u}_k^q = Q_\theta(\boldsymbol{x}_{t_k}^q, t_k) - F_\theta(\boldsymbol{x}_{t_k}^q, t_k), \qquad \boldsymbol{u}_k^c = \boldsymbol{\epsilon}_{t_k}^{\mathrm{corr}}(\boldsymbol{x}_{t_k}^c) - F_\theta(\boldsymbol{x}_{t_k}^c, t_k). \tag{A.8}$$

Thus, the prediction used by trajectory m at $\boldsymbol{x}_{t_k}^m$ is $F_\theta(\boldsymbol{x}_{t_k}^m, t_k) + \boldsymbol{u}_k^m$. The coupled transitions can therefore be written uniformly as

$$\boldsymbol{x}_{t_{k-1}}^{\mathrm{fp}} = A_k \boldsymbol{x}_{t_k}^{\mathrm{fp}} + B_k F_\theta(\boldsymbol{x}_{t_k}^{\mathrm{fp}}, t_k) + \sigma_k \boldsymbol{z}_k, \tag{A.9}$$

$$\boldsymbol{x}_{t_{k-1}}^{m} = A_k \boldsymbol{x}_{t_k}^{m} + B_k \left[F_\theta(\boldsymbol{x}_{t_k}^{m}, t_k) + \boldsymbol{u}_k^m\right] + \sigma_k \boldsymbol{z}_k. \tag{A.10}$$

Let $\boldsymbol{\Delta}_k^m = \boldsymbol{x}_{t_k}^m - \boldsymbol{x}_{t_k}^{\mathrm{fp}}$. Because the common innovation cancels, subtracting the two transitions yields

$$\boldsymbol{\Delta}_{k-1}^m = A_k \boldsymbol{\Delta}_k^m + B_k \left[F_\theta(\boldsymbol{x}_{t_k}^m, t_k) - F_\theta(\boldsymbol{x}_{t_k}^{\mathrm{fp}}, t_k)\right] + B_k \boldsymbol{u}_k^m. \tag{A.11}$$

Assume that the full-precision denoiser is locally Lipschitz over the visited states, where "locally" restricts the assumption to the trajectory region:

$$\|F_\theta(\boldsymbol{x}, t_k) - F_\theta(\boldsymbol{y}, t_k)\|_2 \le L_k \|\boldsymbol{x} - \boldsymbol{y}\|_2 . \tag{A.12}$$

Define the RMS state deviation and the two RMS per-step discrepancies as

$$D_k^m = \left(\mathbb{E}\|\boldsymbol{\Delta}_k^m\|_2^2\right)^{1/2}, \tag{A.13}$$

$$R_k^q = \left(\mathbb{E}\|\boldsymbol{u}_k^q\|_2^2\right)^{1/2}, \tag{A.14}$$

$$R_k^c = \left(\mathbb{E}\|\boldsymbol{u}_k^c\|_2^2\right)^{1/2}. \tag{A.15}$$

Taking Euclidean norms in Eq. (A.11), applying the triangle inequality, and then using the local Lipschitz condition gives the following pointwise bound for every coupled trajectory:

$$\|\boldsymbol{\Delta}_{k-1}^m\|_2 \le (|A_k| + |B_k| L_k) \|\boldsymbol{\Delta}_k^m\|_2 + |B_k| \|\boldsymbol{u}_k^m\|_2.$$

Taking the L_2 norm over the trajectory randomness and applying Minkowski's inequality then gives the RMS one-step bound

$$D_{k-1}^m \le (|A_k| + |B_k| L_k)\, D_k^m + |B_k| R_k^m. \tag{A.16}$$

The coefficient multiplying D_k^m bounds the propagation of an existing state deviation through transition k. This motivates the following propagation factors:

$$\rho_k := |A_k| + |B_k| L_k, \qquad G_k := |B_k| \prod_{j=1}^{k-1} \rho_j, \tag{A.17}$$

where the empty product equals one. Here, ρ_k is the one-step deviation propagation factor. A discrepancy introduced at transition k first enters the state with coefficient $|B_k|$ and then passes through transitions $k-1, \ldots, 1$; hence, G_k is its propagation weight in the endpoint upper bound. Both quantities follow from expanding the one-step inequality and introduce no additional assumption. For each reverse transition $t_k > t_{k-1}$, the generalized-DDIM coefficients satisfy $A_k > 0$ and $B_k < 0$, so $\rho_k > 0$ and $G_k > 0$.

Proposition (coupled endpoint-bound tightening). Under the common-noise coupling and the local Lipschitz condition above, both $m = q$ and $m = c$ satisfy

$$D_0^m \le \mathcal{U}_m := \sum_{k=1}^{K} G_k R_k^m. \tag{A.18}$$

Here, D_0^m is the actual RMS state deviation at the final sampling state, whereas $\mathcal{U}_m$ denotes its accumulated analytical upper bound. The bound sums the RMS discrepancy introduced at each transition after weighting it by its possible propagation to the endpoint; $\mathcal{U}_m$ is not an additional model output or error term.

A directly measurable sufficient condition is

$$R_k^c \le R_k^q, \qquad k = 1, \ldots, K, \tag{A.19}$$

with strict inequality at some transition. Because all G_k are positive, this condition directly gives

$$\mathcal{U}_c = \sum_{k=1}^{K} G_k R_k^c < \sum_{k=1}^{K} G_k R_k^q = \mathcal{U}_q. \tag{A.20}$$

Proof. The coupled trajectories start from the same initial noise, so $D_K^m = 0$. Recursively expanding Eq. (A.16) gives Eq. (A.18). Equation (A.19) and the positive weights G_k then give Eq. (A.20). □

Equation (A.20) also shows that reductions at transitions with larger G_k have greater influence on the endpoint bound. The result concerns the RMS deviation from the coupled full-precision sampler and does not order individual realized trajectories or directly guarantee distribution-level generation quality.

Empirical support for the per-step condition.

We evaluate the per-step quantities in Eq. (A.19) using independently held-out W4A8 DDIM sampling trajectories. The uncorrected quantized sampler and SPEC start from matched initial noise. At every reverse transition, each prediction is compared with the full-precision denoiser evaluated on exactly the same current state.

Let M denote the number of held-out trajectories, let d denote the number of elements in a denoiser output, and let $\boldsymbol{u}_{k,i}^m$ denote the realization of the discrepancy defined in Eq. (A.8) on trajectory i. We estimate the per-element version of R_k^m by

$$\widehat{R}_k^m = \left[\frac{1}{Md} \sum_{i=1}^{M} \left\| \boldsymbol{u}_{k,i}^m \right\|_2^2 \right]^{1/2}, \qquad m \in \{q, c\}. \tag{A.21}$$

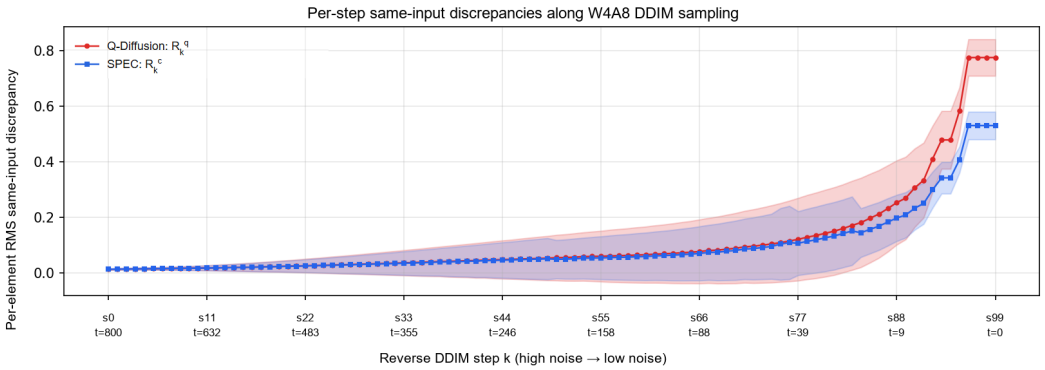


Figure A-1: Empirical assessment of the per-step condition in Eq. (A.19) on held-out W4A8 DDIM trajectories. The curves report the per-element RMS same-input quantization discrepancy of Q-Diffusion, $\widehat{R}_k^q$, and the post-correction residual of SPEC, $\widehat{R}_k^c$, evaluated on their respective sampling states. Shaded regions denote one standard deviation across held-out trajectories.

As shown in Figure A-1, the empirical estimates exhibit the ordering

$$\widehat{R}_k^c < \widehat{R}_k^q, \qquad k = 1, \ldots, K, \tag{A.22}$$

at every evaluated reverse transition. The discrepancy reduction becomes particularly pronounced in the late low-noise region, where the uncorrected per-step quantization discrepancy increases rapidly. These measurements provide timestep-wise empirical support for the sufficient condition in Eq. (A.19) and, through Eq. (A.20), for a tighter corrected endpoint bound.

A.3 Post-SPEC Residual Modeling and Variance Reallocation

The DDIM noise-prediction objective depends on the prescribed $q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$, rather than a unique joint distribution across timesteps Song et al. (2021a). Generalized DDIM therefore retains these

one-time marginals while varying their cross-time coupling, allowing the trained denoiser to be reused with different sampling stochasticity. To expose this freedom, assume the nominal representation $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and an oracle full-precision prediction $F_\theta(\boldsymbol{x}_t, t) = \boldsymbol{\epsilon}_t$. The $t \to s$ affine update then gives

$$\boldsymbol{x}_s^{\mathrm{fp}} = \sqrt{\bar{\alpha}_s}\boldsymbol{x}_0 + c_{t,s}\boldsymbol{\epsilon}_t + \sigma_{t,s}\boldsymbol{z}_t, \qquad c_{t,s}^2 + \sigma_{t,s}^2 = 1 - \bar{\alpha}_s, \tag{A.23}$$

where $c_{t,s} = \sqrt{1 - \bar{\alpha}_s - \sigma_{t,s}^2}$ and $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is independent of $\boldsymbol{\epsilon}_t$. Thus, the target conditional mean and covariance are $\boldsymbol{\mu}_s^0 = \sqrt{\bar{\alpha}_s}\boldsymbol{x}_0$ and $\boldsymbol{Q}_s^0 = (1-\bar{\alpha}_s)\boldsymbol{I}$. At a fixed target timestep s, Eq. (A.23) divides this marginal variance between inherited noise and a fresh Gaussian innovation; it does not fix their cross-time coupling. With $\eta = 0$, $\sigma_{t,s} = 0$ and the step is deterministic given $\boldsymbol{x}_t$, inheriting all target noise through $\boldsymbol{\epsilon}_t$. Feasible $\eta > 0$ introduces an independent innovation and reduces the inherited-noise coefficient while preserving the same target marginal. The target variance depends on s; $\sigma_{t,s}^2$ is instead the transition variance conditional on $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$.

The post-SPEC residual enters the affine update through $B_{t,s}$, so the nominal coefficients can inflate the target covariance. This motivates t-VR, which reduces the explicit innovation and recomputes the coupled prediction coefficient. Under the assumptions below, this update moves the post-SPEC conditional covariance toward the prescribed DDIM marginal covariance.

Let $\boldsymbol{r}_t^c$ denote the post-correction residual in Eq. (3), evaluated on the complete-SPEC state $\boldsymbol{x}_t^c$ before variance reallocation. Under the oracle reference, the complete SPEC prediction is $\boldsymbol{\epsilon}_t + \boldsymbol{r}_t^c$. For the idealized second-order comparison, assume

$$\mathbb{E}[\boldsymbol{r}_t^c \mid \boldsymbol{x}_0] = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{r}_t^c \mid \boldsymbol{x}_0) = v_t\boldsymbol{I}, \qquad \mathrm{Cov}(\boldsymbol{\epsilon}_t, \boldsymbol{r}_t^c \mid \boldsymbol{x}_0) = \mathbf{0}.$$

Independence of $\boldsymbol{z}_t$ from the residual follows from the sampler construction, whereas the last orthogonality relation is an analytical assumption because the residual depends on the current state.

We empirically assess the centered-residual assumption at the aggregate level. Let $r_{i,t,j,p}^c$ denote the residual element of calibration sample i, channel j, and spatial position p. We compute

$$\widehat{\mu}_{r,t,j} = \frac{1}{N_t M}\sum_{i=1}^{N_t}\sum_{p=1}^{M} r_{i,t,j,p}^c, \qquad \widehat{v}_{r,t,j} = \frac{1}{N_t M}\sum_{i=1}^{N_t}\sum_{p=1}^{M}\left(r_{i,t,j,p}^c - \widehat{\mu}_{r,t,j}\right)^2, \tag{A.24}$$

and quantify relative mean-shift energy as $\kappa_{t,j} = \widehat{\mu}_{r,t,j}^2 / \widehat{v}_{r,t,j}$.

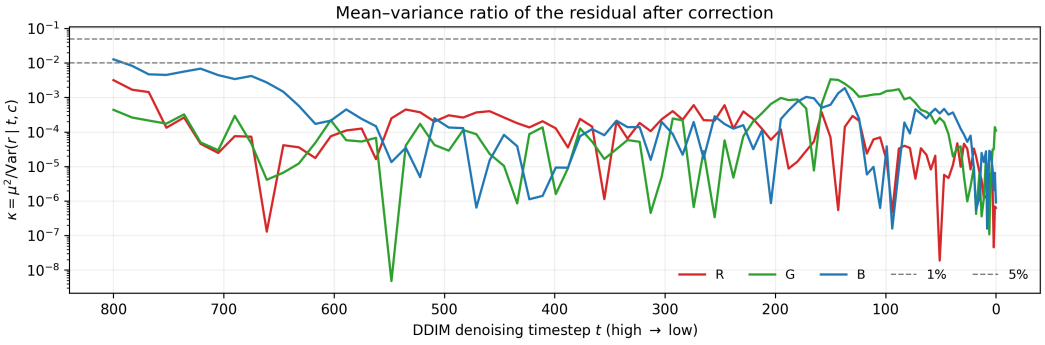


Figure A-2: Squared-mean-to-variance ratio $\kappa_{t,j}$ of the post-correction residual across DDIM timesteps and output channels. The dashed lines indicate ratios of 1% and 5%.

Figure A-2 shows that $\kappa_{t,j}$ remains below 5% in all evaluated cases and below 1% in nearly all cases. Thus, channelwise mean-shift energy is small relative to residual variation across calibration samples and spatial positions. This aggregate diagnostic provides empirical support for treating the residual as approximately centered over the calibration distribution.

With the original coefficients, the residual produces the conditional covariance inflation

$$\mathrm{Cov}(\boldsymbol{x}_s^{\mathrm{SPEC}} \mid \boldsymbol{x}_0) = \boldsymbol{Q}_s^0 + B_{t,s}^2 v_t \boldsymbol{I}. \tag{A.25}$$

Equation (A.25) characterizes the covariance inflation before variance reallocation. We next compare it with a feasible t-VR update.

Proposition (conditional marginal-covariance contraction). Under the oracle-reference and nominal-state conditions above, together with the centered-isotropic-residual and orthogonality assumptions, let $\widehat{v}_t \geq 0$ satisfy $B_{t,s}^2 \widehat{v}_t \leq \sigma_{t,s}^2$. Applying the variance and coefficient updates in Eqs. (18) and (19) preserves the target conditional mean and satisfies

$$\mathbf{0} \preceq \underbrace{\mathrm{Cov}(\boldsymbol{x}_s^{t\text{-VR}} \mid \boldsymbol{x}_0) - \boldsymbol{Q}_s^0}_{\widetilde{B}_{t,s}^2 v_t \boldsymbol{I}} \preceq \underbrace{\mathrm{Cov}(\boldsymbol{x}_s^{\mathrm{SPEC}} \mid \boldsymbol{x}_0) - \boldsymbol{Q}_s^0}_{B_{t,s}^2 v_t \boldsymbol{I}}. \tag{A.26}$$

Proof. Feasibility gives $0 \leq \widetilde{\sigma}_{t,s} \leq \sigma_{t,s}$. For a reverse step $t > s$, the magnitude $|B_{t,s}^{\mathrm{DDIM}}(u)|$ in Eq. (A.6) is nondecreasing over feasible u. Hence $|\widetilde{B}_{t,s}| \leq |B_{t,s}|$. The generalized-DDIM coefficient relation keeps the oracle-reference contribution equal to $\boldsymbol{Q}_s^0$, while the centered residual contributes $\widetilde{B}_{t,s}^2 v_t \boldsymbol{I}$. Combining these facts with Eq. (A.25) proves Eq. (A.26).

Because both deviations in Eq. (A.26) are isotropic and positive semidefinite, the result also implies

$$\left\|\mathrm{Cov}(\boldsymbol{x}_s^{t\text{-VR}} \mid \boldsymbol{x}_0) - \boldsymbol{Q}_s^0\right\|_{\mathrm{F}} \leq \left\|\mathrm{Cov}(\boldsymbol{x}_s^{\mathrm{SPEC}} \mid \boldsymbol{x}_0) - \boldsymbol{Q}_s^0\right\|_{\mathrm{F}}. \tag{A.27}$$

These bounds describe the target-time marginal $\boldsymbol{x}_s \mid \boldsymbol{x}_0$: t-VR preserves its prescribed mean and reduces the residual-induced excess variance in every direction. The resulting marginal therefore retains the DDIM target location and has a spread closer to the target: its covariance mismatch in Frobenius norm cannot increase and strictly decreases when $v_t > 0$ and $|\widetilde{B}_{t,s}| < |B_{t,s}|$. This second-order comparison does not characterize the marginal's potentially non-Gaussian shape.

The residual densities in Fig. 2 show broad tails. State- and coordinate-dependent scales provide a plausible mechanism: pooling approximately Gaussian residuals across these scales produces a heavier-tailed marginal than a single Gaussian. A Gaussian model with a matched variance can represent the aggregate second moment, but it assigns too little probability to the observed large residuals. With finite calibration data, underrepresented tail samples can therefore yield an underestimated second moment; variance reallocation would then subtract too little from the Gaussian innovation budget, leaving part of the residual contribution uncompensated. The zero-location Student-t working marginal allows its degrees of freedom to capture heavier tails and uses the fitted variance as the second-moment estimate, provided the fit describes the calibration residuals.

Student-t parameter stability. The fitted variance in Eq. (17) is finite for $\widehat{\nu}_t > 2$, but a fit close to this boundary can amplify variance estimation error. We examine the fitted degrees of freedom and implied variance across timesteps in Fig. A-3. The density and fitting objective are given in Appendix B.2.

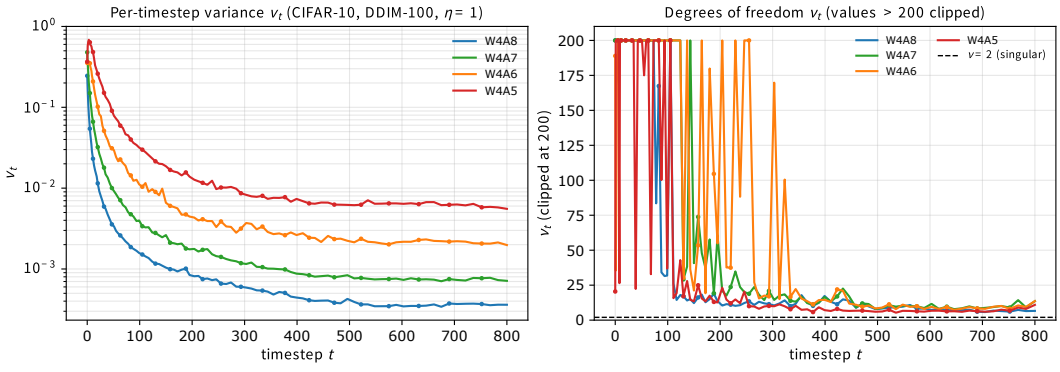


Figure A-3: Timestep-wise Student-t-implied residual variance $\widehat{v}_t$ (left) and fitted degrees of freedom $\widehat{\nu}_t$ (right) on CIFAR-10 with 100-step DDIM ($\eta = 1$) under W4A8–W4A5. The dashed line marks the finite-variance boundary $\nu = 2$.

Figure A-3 shows that the fitted degrees of freedom remain separated from $\nu = 2$ across the displayed timesteps, without exhibiting near-boundary amplification of the variance multiplier $\widehat{\nu}_t/(\widehat{\nu}_t - 2)$. Fluctuations at large $\widehat{\nu}_t$ have a limited effect on this multiplier, which converges to one. This diagnostic addresses the stability of the fitted variance; the separate stepwise budget condition $B_{t,s}^2 \widehat{v}_t \leq \sigma_{t,s}^2$ is checked during sampling.

B Reproducibility Details

This section reports the backbone reconstruction, corrector training, calibration, and evaluation settings used in our experiments. Unless otherwise specified, the full-precision teacher and quantized denoiser remain frozen throughout corrector training. All reproduced runs use a single NVIDIA RTX 2080 Ti GPU and an Intel i9-9900KF CPU.

B.1 Backbone Reconstruction and SPEC Training

Backbone reconstruction. We construct the quantized backbones following the Q-Diffusion reconstruction protocol Li et al. (2023). We use asymmetric uniform quantization for weights and symmetric uniform quantization for activations, with per-channel weight quantization and per-tensor activation quantization. The input and output tensors of the denoising network are retained in full precision.

The BRECQ-style Li et al. (2021) reconstruction hyperparameters are summarized in Table B-1.

Table B-1: Hyperparameters used for BRECQ-style weight and activation reconstruction of the quantized diffusion backbone.

Parameter	Weight	Activation
Optimization iterations	20,000	5,000
Warm-up iterations	4,000	0
Calibration batch size	32 for CIFAR-10; 16 for LSUN	
Optimizer	Adam	Adam
Learning rate	1×10^{-3}	4×10^{-4}
LR scheduler	None	Cosine
Weight decay	0	0
Reconstruction loss	MSE + rounding ($\lambda = 0.01$)	L_p ($p = 2.4$)
Temperature scheduler	`LinearTempDecay`	–
Asymmetric calibration	True	False

The corresponding Q-Diffusion baseline and SPEC use the same reconstructed backbone, isolating the sampling-time correction in their comparison.

Two-stage SPEC training. Training follows the two-stage strategy in Section 3.2. Stage I optimizes the shared encoder and temporal-condition branch using $\mathcal{L}_{\text{time}}$. We then freeze these parameters, recollect temporally aligned trajectories, and optimize only the quantization-error branch on $\boldsymbol{e}_t^a$ using $\mathcal{L}_{\text{corr}}$.

Table B-2 lists the default training hyperparameters used for the main CIFAR-10 experiments.

Table B-2: Predictive-corrector training and residual-calibration hyperparameters.

Parameter	Stage I	Stage II
Optimized module	Encoder + temporal branch	Error branch
Objective	$\mathcal{L}_{\text{time}}$	$\mathcal{L}_{\text{corr}}$
Epochs	20	50
Batch size	256	64
Optimizer	AdamW	AdamW
Learning rate	1×10^{-3}	1×10^{-3}
Weight decay	1×10^{-4}	1×10^{-4}
LR scheduler	Cosine	Cosine
Gradient clipping	1.0	1.0
Collected / Student-t fitting trajectories	500 / 200	
Trajectory split	Validation if $\texttt{traj_id} \bmod 5 = 0$ ($\approx 20\%$)	
Temporal refresh steps	8 (uniformly sampled)	
Stage-II auxiliary-loss weight	$\lambda_{\text{geo}} = 0.5$	

B.2 Calibration and Inference

Student-t residual fitting. After flattening the channel and spatial dimensions, let $r^c_{i,t,k}$ denote residual coordinate k of calibration sample i at timestep t, where $k = 1, \ldots, d$. The probability density function of the zero-location Student-t distribution is

$$p(r; \nu_t, s_t) = \frac{\Gamma\left(\frac{\nu_t+1}{2}\right)}{\Gamma\left(\frac{\nu_t}{2}\right)\sqrt{\nu_t \pi}\, s_t} \left(1 + \frac{r^2}{\nu_t s_t^2}\right)^{-\frac{\nu_t+1}{2}}. \tag{B.1}$$

We estimate the timestep-wise degrees of freedom and scale parameters by maximizing the elementwise marginal composite likelihood

$$(\widehat{\nu}_t, \widehat{s}_t) = \underset{\nu>2,\ s>0}{\arg\max} \sum_{i=1}^{N_t} \sum_{k=1}^{d} \log p\left(r^c_{i,t,k}; \nu, s\right). \tag{B.2}$$

The constraint $\nu > 2$ ensures finite fitted variance. This marginal composite likelihood does not assume coordinate independence.

Complete procedure. Algorithm 1 summarizes the two-stage SPEC training, post-training residual-variance calibration, and corrected sampling procedure.

Algorithm 1 Training, calibration, and DDIM sampling of SPEC with optional t-VR.

Input: Frozen denoisers F_θ and Q_θ, nominal DDIM sequence, refresh set $\mathcal{R}$, stochasticity parameter η, and t-VR flag $\chi_{\mathrm{VR}} \in \{0, 1\}$
Output: Trained SPEC predictor, calibrated $\{\widehat{v}_t\}$ if $\chi_{\mathrm{VR}} = 1$ and $\eta > 0$, and sample $\boldsymbol{x}_0$

Training Stage I: Temporal-condition prediction
Collect $\boldsymbol{c}_t = (\boldsymbol{x}_t^q, \boldsymbol{\epsilon}_t^q, t)$ along quantized trajectories; obtain $\delta_t^\star$ using Eq. (8)
Optimize the shared encoder and P_ω^t using $\mathcal{L}_{\mathrm{time}}$, with P_ω^e frozen and inactive
Training Stage II: Full-resolution error prediction
Freeze the shared encoder and P_ω^t; recollect temporally aligned trajectories $\boldsymbol{x}_t^a$
Construct $\boldsymbol{e}_t^a$ using Eq. (12); optimize only P_ω^e using $\mathcal{L}_{\mathrm{corr}}$
Freeze the complete SPEC predictor
if $\chi_{\mathrm{VR}} = 1$ and $\eta > 0$ **then**
  Post-training residual-variance calibration
  Run complete SPEC with the original Gaussian noise schedule to collect $\boldsymbol{x}_t^c$
  Compute $\boldsymbol{r}_t^c$ using Eq. (15); fit $(\widehat{\nu}_t, \widehat{s}_t)$ with $\widehat{\nu}_t > 2$ and $\widehat{s}_t > 0$
  Set $\widehat{v}_t = \widehat{\nu}_t \widehat{s}_t^2 / (\widehat{\nu}_t - 2)$
end if
Inference
Sample $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for each nominal transition $t \to s$ **do**
  Compute the original generalized-DDIM coefficients $A_{t,s}$, $B_{t,s}$, and $\sigma_{t,s}$ using Eqs. (A.6) and (A.4), with $\eta = 0$ for the final transition and the prescribed η otherwise
  Compute $\boldsymbol{\epsilon}_t^q = Q_\theta(\boldsymbol{x}_t, t)$ and $\boldsymbol{c}_t = (\boldsymbol{x}_t, \boldsymbol{\epsilon}_t^q, t)$
  if $t \in \mathcal{R}$ **then**
    Predict $\widetilde{\delta}_t = P_\omega^t(\boldsymbol{c}_t)$; round the offset and clip the adjusted timestep to obtain $\widetilde{t}$
    Set $\widetilde{\boldsymbol{\epsilon}}_t^q = Q_\theta(\boldsymbol{x}_t, \widetilde{t})$
  else
    Set $\widetilde{t} = t$ and $\widetilde{\boldsymbol{\epsilon}}_t^q = \boldsymbol{\epsilon}_t^q$
  end if
  Predict $\widehat{\boldsymbol{e}}_t = P_\omega^e(\boldsymbol{x}_t, \widetilde{\boldsymbol{\epsilon}}_t^q, \widetilde{t})$; set $\boldsymbol{\epsilon}_t^{\mathrm{corr}} = \widetilde{\boldsymbol{\epsilon}}_t^q - \widehat{\boldsymbol{e}}_t$
  if $\chi_{\mathrm{VR}} = 1$, $\sigma_{t,s}^2 > 0$, and $B_{t,s}^2 \widehat{v}_t \le \sigma_{t,s}^2$ **then**
    Compute $\widetilde{\sigma}_{t,s}$ using Eq. (18)
    Compute $\widetilde{B}_{t,s}$ using Eq. (19)
  else
    Set $\widetilde{B}_{t,s} = B_{t,s}$ and $\widetilde{\sigma}_{t,s} = \sigma_{t,s}$
  end if
  Draw $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; update $\boldsymbol{x}_s = A_{t,s}\boldsymbol{x}_t + \widetilde{B}_{t,s}\boldsymbol{\epsilon}_t^{\mathrm{corr}} + \widetilde{\sigma}_{t,s}\boldsymbol{z}_t$
end for
return $\boldsymbol{x}_0$

Table B-3: Available offline preparation cost records for SPEC with t-VR.

Setting	Teacher FW	Accounted GPU-h
CIFAR-10, W8A8, DDIM-100, $\eta = 1$	6.72×10^4	0.36
ImageNet-256 (LDM), W4A8, DDIM-20, $\eta = 1$	2.56×10^3	1.47

GPU-h includes (i) trajectory collection, comprising quantized-UNet and full-precision teacher evaluations and state storage; (ii) training of the lightweight modules; and (iii) Student-t fitting for residual-variance calibration. Teacher FW counts only full-precision teacher evaluations for residual labeling and is computed from the trajectory count and the number of labeled states per trajectory.

Offline preparation cost. Table B-3 reports preparation separately from the inference-time costs in Table 4. Teacher evaluations are used for residual labeling, and all preparation is performed offline; the teacher is not required at deployment.

B.3 Evaluation Protocol

For all reproduced methods, evaluation metrics are computed from the final RGB images using the same preprocessing, sample counts, and reference statistics within each setting. We follow the standard FID protocol Heusel et al. (2017) and report Inception Score (IS) on CIFAR-10 and

CelebA Salimans et al. (2016). We use 50,000 generated images for CIFAR-10 and CelebA, 3,000 for each LSUN dataset, and 10,000 for ImageNet, consistent with Section 4.1.

Table B-2 defines the trajectory-level validation split used during corrector development. For PER evaluation, we freeze the trained predictor and all settings, then use the complete SPEC sampler to generate sampling test trajectories with fresh noise and seeds disjoint from those used for training and model selection. At each collected state $\boldsymbol{x}_t^c$, the denominator uses the original quantization discrepancy before temporal correction, $Q_\theta(\boldsymbol{x}_t^c, t) - F_\theta(\boldsymbol{x}_t^c, t)$, and the numerator uses the residual after complete SPEC correction, $\boldsymbol{\epsilon}_t^{\mathrm{corr}}(\boldsymbol{x}_t^c) - F_\theta(\boldsymbol{x}_t^c, t)$. Both quantities share the same state and full-precision reference at the nominal timestep t.

C Additional Empirical Evidence

C.1 Ablations and Robustness Checks

Table C-1: Ablation of the training objectives and auxiliary-loss weight for 100-step DDIM ($\eta = 0$) on CIFAR-10 (32×32). The default auxiliary-loss weight is $\lambda_{\mathrm{geo}} = 0.5$.

Method	$\mathcal{L}_{\mathrm{err}}$	$\mathcal{L}_{\mathrm{dir}}$	$\mathcal{L}_{\mathrm{mag}}$	FID	IS
Q-Diffusion†	–	–	–	3.80	**9.36**
MSE	✓	–	–	3.63	8.54
MSE + Direction	✓	✓	–	3.59	8.67
MSE + Magnitude	✓	–	✓	3.61	9.03
Ours ($\lambda_{\mathrm{geo}} = 0.5$)	✓	✓	✓	3.58	9.27
Ours ($\lambda_{\mathrm{geo}} = 0.3$)	✓	✓	✓	3.60	9.19
Ours ($\lambda_{\mathrm{geo}} = 0.8$)	✓	✓	✓	**3.57**	9.21

Correction objectives. Table C-1 shows modest FID gains from the directional and magnitude objectives over reconstruction-only training. FID varies from 3.57 to 3.60 across $\lambda_{\mathrm{geo}} \in \{0.3, 0.5, 0.8\}$; the default value 0.5 gives 3.58 FID and 9.27 IS.

Table C-2: Candidate encoder configurations for 100-step DDIM ($\eta = 0$) on CIFAR-10 (32×32).

Encoder	Params	FP32 Size	FID
Pixel-wise MLP ($6 \to 32 \to 32 \to 32$)	≈ 10.0K	≈ 39.1KiB	3.92
$\mathrm{Conv}_{3\times3} + \mathrm{GN} + \mathrm{SiLU}$	≈ 9.3K	≈ 36.4KiB	3.77
Ours: $[\mathrm{Conv}_{3\times3} + \mathrm{GN} + \mathrm{SiLU}] \times 3$	≈ 27.9K	≈ 109.1KiB	3.58
$[\mathrm{Conv}_{5\times5} + \mathrm{GN} + \mathrm{SiLU}] \times 3$	≈ 63.8K	≈ 249.1KiB	3.54

Encoder architecture. The three-block 3×3 encoder provides the selected accuracy–efficiency trade-off; the 5×5 variant improves FID by only 0.04 while more than doubling parameter storage.

Multi-seed stability. Table C-3 reports the individual FIDs of SPEC + t-VR on CIFAR-10 using 100-step DDIM with $\eta = 0.5$ for seeds 1234, 2345, and 3456. We report the arithmetic mean and sample standard deviation across the three runs, using $n - 1 = 2$ in the variance denominator. The corresponding entries in Table 4 report these means, rounded to two decimal places.

Table C-3: Per-seed FID results for SPEC + t-VR on CIFAR-10 with 100-step DDIM ($\eta = 0.5$). The final column reports the mean $\pm$ sample standard deviation over three runs.

Bitwidth	Seed 1234	Seed 2345	Seed 3456	Mean $\pm$ SD
W8A8	3.89	3.92	3.90	3.90 ± 0.02
W4A8	5.44	5.51	5.53	5.49 ± 0.05

C.2 Transfer Test with a Higher-Order Singlestep Sampler

To examine transfer beyond the DDIM-family scope, we apply SPEC to third-order singlestep DPM-Solver++ on CIFAR-10 without solver-matched retraining, using 100-step deterministic sampling and 50,000 generated images. Table C-4 shows that SPEC increases FID from 4.93 to 9.93 under W8A8 and from 7.01 to 12.16 under W4A8, while slightly improving IS. The two metrics capture different aspects: IS summarizes classifier predictions on generated images, whereas FID compares generated and real feature statistics. In contrast to the gains under deterministic DDIM, these results highlight the dependence of correction performance on the sampler's update structure.

Table C-4: Quantization results for deterministic 100-step third-order singlestep DPM-Solver++ on CIFAR-10.

Method	W8A8		W4A8	
	FID	IS	FID	IS
Q-Diffusion†	**4.93**	9.71	**7.01**	9.13
Ours	9.93	**9.88**	12.16	**9.30**

The negative transfer suggests a mismatch in stage-wise residual coupling. DDIM advances the state using the current prediction, whereas each third-order singlestep update constructs two intermediate stages and combines data predictions through stage differences. Prediction errors therefore affect both the intermediate states and the final update. SPEC optimizes same-input reconstruction, direction, and magnitude at individual evaluation points. Differences between the residuals at the starting and intermediate stages are not explicitly constrained by this objective. In data-prediction space, these residual differences are exactly the errors in the corresponding prediction differences relative to full-precision evaluations on the same stage states. Consequently, correction at individual evaluation points may reduce their errors while changing the stage-wise relationships, potentially offsetting the benefit in the higher-order update.

Offset rounding, sparse temporal refresh, and the gap between recollected training states and solver intermediate states can further change the correction response. Solver-matched training would need to cover intermediate states and stage-wise prediction differences or the complete outer-step update.