RESTORING MINORITY GENERATION IN LOW-BIT DIFFUSION MODELS WITH A PLUG-IN REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient deployment of diffusion models has motivated extensive research on low-bit quantization, which can substantially reduce memory and computation while maintaining generation quality under standard sampling. However, we argue that this preservation is not guaranteed to extend to the task of *minority generation*, which specifically targets generation from low-density but valid regions of the data distribution. We show that quantization error becomes significant in clean estimates at high noise levels, on which existing minority samplers heavily rely yet standard sampling places only a small weight. Based on this analysis, we propose a lightweight plug-in repair that aligns the high-noise clean estimates of quantized models with those of their full-precision counterparts. Our repair combines selective bit allocation with a quantization-consistent low-rank adapter that remains on the deployed quantization grid, and is activated only when computing these estimates, leaving the ordinary sampling path unchanged. Experiments across diverse diffusion architectures, bit widths, and minority samplers show that our repair substantially improves minority generation with negligible overhead while retaining the efficiency of low-bit inference. Code is available here.

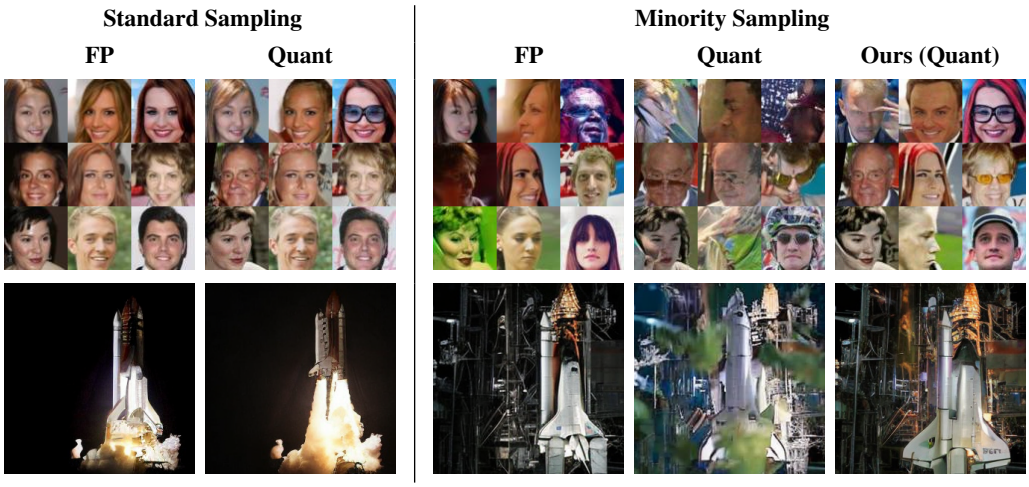

Figure 1: **Quantization preserves standard generation but can degrade minority generation.** Top: ADM on CelebA. Bottom: DiT-XL/2 on ImageNet ("space shuttle"). Under W6A6 quantization, standard samples stay comparable to full precision while minority samples lose rare attributes and fidelity, which our repair restores.

1 INTRODUCTION

Real-world datasets often exhibit long-tailed distributions, with most samples concentrated in high-density regions and *minority samples* occupying low-density yet valid regions of the data manifold (Sehwag et al., 2022; Um et al., 2024). Although infrequent, minority samples can contain distinctive attributes that are important for applications such as classification, anomaly detection, medical diagnosis, and creative AI (Du et al., 2022; 2023; Um et al., 2024; Han et al., 2023). Gen-

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.

erating such samples can expand the coverage of rare cases available for training and evaluation. However, diffusion models are trained to reproduce the data distribution and therefore naturally generate high-likelihood majority samples more frequently, making the collection of minority samples inefficient and computationally demanding (Sehwag et al., 2022; Um et al., 2024). Recent minority samplers address this majority-oriented behavior by steering the reverse process toward low-density regions (Um et al., 2024; Um & Ye, 2024; 2025; Um et al., 2025; Park & Um, 2026).

At the same time, the practical deployment of diffusion models remains computationally demanding because generating a single sample requires repeated evaluations of a large denoising network (Ho et al., 2020; Rombach et al., 2022). Low-bit quantization (Esser et al., 2020; Li et al., 2021) has consequently attracted substantial attention as a means of reducing the memory footprint and latency of each denoising step. Recent diffusion quantization methods achieve aggregate generation quality, typically measured by metrics such as FID (Heusel et al., 2017), comparable to that of their full-precision counterparts under standard sampling (Huang et al., 2024). However, aggregate performance under standard sampling does not reveal whether the minority-generation capability of the full-precision model is also preserved after quantization. Despite rapid progress in both diffusion quantization and minority generation, their interaction remains largely unexplored.

In this work, we take the first step toward studying minority generation with low-bit quantized diffusion models. We uncover an overlooked failure mode: a quantized model that closely preserves the generation quality of its full-precision counterpart under standard sampling can nevertheless suffer substantial degradation under minority sampling, suppressing rare attributes and reducing sample fidelity, as illustrated in Fig. 1. This failure stems from how existing minority generation methods locate low-density regions: most of them rely on clean estimates obtained via Tweedie's formula (Robbins, 1992) at high noise levels to guide sampling toward such regions (Um et al., 2024; Um & Ye, 2024; 2025; Park & Um, 2026). Our analysis reveals that Tweedie's formula amplifies quantization error in the denoiser prediction at high noise levels while attenuating it at low noise levels (see Fig. 2), which can misguide quantized models when steering toward low-density regions. We further show that standard sampling is less affected, as the reverse update places only a small weight on the clean estimate at high noise levels, whereas minority guidance depends on it directly.

Motivated by this finding, we propose a lightweight plug-in repair for already-quantized diffusion models that restores accurate clean estimates at high noise levels. Guided by a shared clean-prediction objective, our method combines selective bit allocation, which refines the activation grids of the few quantizers that dominate the high-noise error, with a quantization-consistent low-rank adapter (Hu et al., 2022) that corrects the weights while remaining on the deployed quantization grid. The repaired model thus runs on the original low-bit kernels, and we show that reducing this objective bounds the error of the minority score itself. For minority samplers built on high-noise clean estimates, the repair is activated only when computing them, while ordinary reverse updates retain the original quantized model. We further find that the repair generalizes to minority samplers that do not rely on such estimates, such as Boost-and-Skip (Um et al., 2025).

Experiments across diverse diffusion architectures, bit widths, and minority samplers show that our repair substantially restores minority generation quality with negligible overhead, while retaining the efficiency of low-bit inference. Our contributions are summarized as follows:

- We present the first study of minority generation with low-bit quantized diffusion models, revealing that Tweedie's formula amplifies quantization error at high noise levels, which can misguide minority generation even when standard generation is well preserved.
- We propose a lightweight plug-in repair that aligns the high-noise clean estimates of quantized models with their full-precision counterparts, combining selective bit allocation with a quantization-consistent low-rank adapter that remains on the deployed quantization grid.
- Across diverse diffusion architectures, bit widths, and minority samplers, we demonstrate substantial recovery of minority generation while preserving low-bit efficiency, and further show that the repair generalizes to samplers that do not rely on high-noise clean estimates without retraining.

2 Related Work

Minority generation with diffusion models. Minority generation targets low-density yet valid samples whose rare attributes are valuable for data augmentation, anomaly detection, medical analysis, and creative applications (Sehwag et al., 2022; Um et al., 2024). Early approaches such as Mi-

nority Guidance (MG) employ trained classifiers to guide sampling toward low-density regions (Sehwag et al., 2022; Um et al., 2024), and later methods extend this through self-guidance (SGMS), prompt optimization, representation-based guidance, and minority-preference fine-tuning (Um & Ye, 2024; 2025; Park & Um, 2026; Kim et al., 2026). Most of these methods rely on a minority score, typically estimated from clean predictions reconstructed under strong noise, to guide sampling or optimization. A notable exception is Boost-and-Skip (BnS), which avoids minority-score estimation by modifying the initial noise and sampling schedule (Um et al., 2025).

Quantization of diffusion models. Low-bit quantization has been actively studied to reduce the memory and computational costs of diffusion models. Existing methods address diffusion-specific challenges such as timestep-varying activation distributions, temporal-feature distortion, quantization-error accumulation, and activation outliers (Shang et al., 2023; Li et al., 2023; So et al., 2023; Huang et al., 2024; Lee et al., 2025). Recent studies further extend quantization to diffusion transformers (Wu et al., 2024; Chen et al., 2025; Zhao et al., 2024; Hwang et al., 2025). These approaches achieve aggregate generation quality close to full precision under standard sampling. However, their calibration and evaluation primarily target the ordinary sampling trajectory, leaving their ability to preserve low-density generation and the high-noise clean predictions used for minority estimation largely unexplored.

3 PRELIMINARIES

Diffusion models. Diffusion-based generative models (Ho et al., 2020; Song et al., 2021) are characterized by a fixed forward process that gradually corrupts data (with Gaussian noise) and a learned reverse process that gradually removes it. The forward diffusion process is defined as $q_{\alpha_t}(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I})$, where $\alpha_t = \prod_{i \le t}(1-\beta_i)$. A diffusion model learns to reverse this process and generates through

$$\boldsymbol{x}_{t-1} = \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) + \boldsymbol{\Sigma}_\theta^{1/2}(\boldsymbol{x}_t, t)\boldsymbol{z}, \qquad \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right), \quad (1)$$

with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \beta_t \boldsymbol{I}$ (Ho et al., 2020). Here, $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ denotes the noise predicted by the denoiser at timestep t. Standard sampling approximately follows the learned data distribution and thus tends to favor high-density regions (Sehwag et al., 2022; Um et al., 2024).

Minority generation. *Minority samples* are instances in low-density regions of the data distribution (Um et al., 2024; Um & Ye, 2024). Unlike standard samplers, which tend to generate high-density samples, *minority generation* aims to draw valid instances from these low-density regions. This amounts to sampling from $\mathcal{S}_\tau = \{\boldsymbol{x} \in \mathcal{M} : p_\theta(\boldsymbol{x}) < \tau\}$, where $\mathcal{M}$ denotes the data manifold, p_θ the density learned by the model, and τ a density threshold (Um & Ye, 2024; Um et al., 2025).

Existing minority-sampling methods mostly rely upon a *minority score* that measures how atypical a sample is under the diffusion prior (Um et al., 2024; Um & Ye, 2024; 2025). A common construction uses the model's clean reconstruction behavior at high noise levels. Specifically, given $\boldsymbol{x}_t \sim q_{\alpha_t}(\cdot \mid \boldsymbol{x}_0)$, Tweedie's formula gives (Robbins, 1992):

$$\hat{\boldsymbol{x}}_0^\theta(\boldsymbol{x}_t) \coloneqq \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right). \quad (2)$$

This estimate tends to favor high-density modes especially under strong noise corruption, so minority samples typically incur larger reconstruction errors (Um et al., 2024). This motivates the following reconstruction-based minority score:

$$\mathcal{L}(\boldsymbol{x}_0; s) \coloneqq \mathbb{E}_{q_{\alpha_s}(\boldsymbol{x}_s \mid \boldsymbol{x}_0)}\left[d\left(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0^\theta(\boldsymbol{x}_s)\right)\right], \quad (3)$$

where d denotes a discrepancy in pixel, latent, or feature space, such as LPIPS (Zhang et al., 2018) or squared ℓ_2 distance (Um & Ye, 2025). The minority-score-based methods discussed in Sec. 2 differ in how they construct and apply guidance but share a reliance on denoiser-derived clean predictions.

Quantized diffusion models. Quantization reduces the memory footprint and latency of neural networks by representing their weights and activations with low-bit integers (Esser et al., 2020; Li et al., 2021). For diffusion models, post-training quantization (PTQ) is common, as it requires only a small calibration set instead of retraining the costly denoiser (Shang et al., 2023; Li et al., 2023).

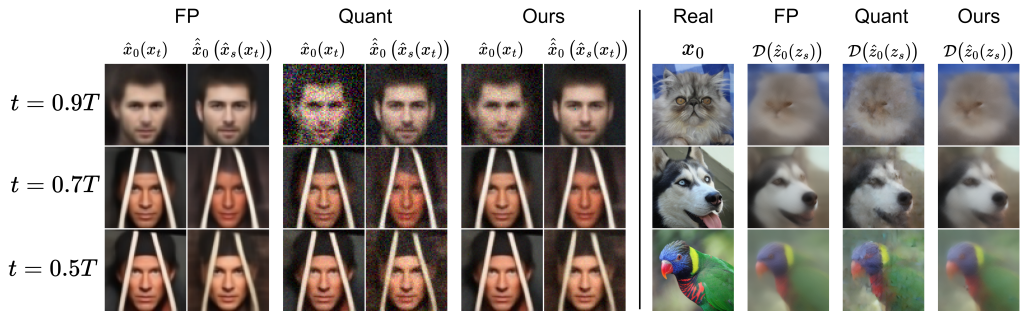

Figure 2: **Quantization corrupts the clean predictions underlying minority scores.** *Left.* SGMS scores a sample online from its Tweedie estimate $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ and a second estimate obtained by re-corrupting and reconstructing it. *Right.* MG scores a real sample $\boldsymbol{x}_0$ offline against the decoded clean prediction $D(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_s))$ at a fixed corruption level. In both, quantization visibly corrupts the reconstructions and our repair brings them close to full precision.

A uniform quantizer projects a tensor onto a b-bit grid as

$$Q(\boldsymbol{v}) = \Delta \left(\mathrm{clip} \left(\lfloor \boldsymbol{v}/\Delta \rceil + z_q, 0, 2^b - 1 \right) - z_q \right), \tag{4}$$

where Δ and z_q denote the scale and zero point that define the grid. They are calibrated per tensor or per channel and are omitted from Q for brevity. Rounding introduces an error of at most $\Delta/2$ and values outside the calibrated range incur a larger clipping error. We write Q_ℓ^{w} and Q_ℓ for the weight and activation quantizers of layer ℓ so that each quantized layer computes $Q_\ell^{\mathrm{w}}(\boldsymbol{W}_\ell)\, Q_\ell(\boldsymbol{a}_\ell)$. We denote the resulting quantized denoiser by $\boldsymbol{\epsilon}_{\hat{\theta}}$ and it replaces $\boldsymbol{\epsilon}_\theta$ in Eq. (1) during generation. A key challenge for diffusion models is that the denoiser is evaluated repeatedly across timesteps whose activation distributions vary substantially. Existing PTQ methods therefore calibrate quantizers on states collected along the standard sampling trajectory and maintain timestep-aware activation quantizers (Shang et al., 2023; Li et al., 2023; So et al., 2023; He et al., 2023; Huang et al., 2024). This achieves strong performance under standard sampling (Huang et al., 2024). Whether it also preserves minority generation remains unexplored and is the question we investigate next.

4 Why Quantization Harms Minority Generation

We analyze why quantization degrades minority generation while largely preserving standard sampling. We first show how the quantization error of the denoiser propagates to the Tweedie estimate across noise levels, and then that minority scores rely on this estimate precisely where the error is amplified, whereas standard sampling does not.

Quantization error. To isolate the effect of quantization, we decompose the prediction of the quantized denoiser $\boldsymbol{\epsilon}_{\hat{\theta}}$ into its full-precision counterpart and a quantization-induced error:

$$\boldsymbol{\epsilon}_{\hat{\theta}}(\boldsymbol{x}_t, t) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \boldsymbol{\delta}(\boldsymbol{x}_t, t), \tag{5}$$

where $\boldsymbol{\delta}$ captures the combined error from weight and activation quantization and depends on both $\boldsymbol{x}_t$ and t. We use this decomposition to trace how $\boldsymbol{\delta}$ propagates to the clean estimate.

Tweedie estimates amplify quantization error at high noise levels. Let $\hat{\boldsymbol{x}}_0^{\hat{\theta}}$ and $\hat{\boldsymbol{x}}_0^{\theta}$ denote the Tweedie estimates obtained from the quantized and full-precision denoisers, respectively. Since Eq. (2) is linear in the denoiser prediction, substituting Eq. (5) gives

$$\hat{\boldsymbol{x}}_0^{\hat{\theta}}(\boldsymbol{x}_t) - \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_t) = -\sqrt{\frac{1-\alpha_t}{\alpha_t}}\, \boldsymbol{\delta}(\boldsymbol{x}_t, t). \tag{6}$$

Note that the factor $\sqrt{(1-\alpha_t)/\alpha_t}$ exceeds one whenever $\alpha_t < 0.5$ and grows without bound as $\alpha_t \to 0$. Hence, the quantization error is amplified in the clean estimate at high noise levels. Conversely, the factor vanishes as $\alpha_t \to 1$, so the same error is attenuated at low noise levels.

Minority scores operate in the high-noise regime. Minority scores are intentionally evaluated under strong corruption. Existing methods use $s = 0.8T$ for cosine schedules and $s = 0.5T$ for

linear schedules (Um et al., 2024; Um & Ye, 2024), both corresponding to $\alpha_s \approx 0.1$, so we define the high-noise regime as

$$\mathcal{T}_{\text{hn}} = \{t \mid \alpha_t \leq 0.1\}, \tag{7}$$

where the amplification factor in Eq. (6) is at least 3 and grows rapidly as α_t decreases.

Scoring at lower noise levels would avoid this amplification, but the resulting clean estimates stay close to $\boldsymbol{x}_t$ and carry little information about low-density regions (Um et al., 2024; Um & Ye, 2024). Minority scores are therefore inherently exposed to the amplified quantization error. Fig. 2 shows this on the reconstructions themselves, where the clean predictions of the quantized model are visibly corrupted at high noise levels. Fig. 3 quantifies the effect on ADM CelebA at W6A6. The clean-prediction error grows with t and is largest in $\mathcal{T}_{\text{hn}}$, and the SGMS scores (Um & Ye, 2024) built from it are inflated by about $2\times$ relative to full precision with a concordance correlation (CCC) of only 0.31. Our repair reduces the clean-prediction error by $2.5\times$ on average over $\mathcal{T}_{\text{hn}}$ and restores the scores to the diagonal with CCC 0.82.

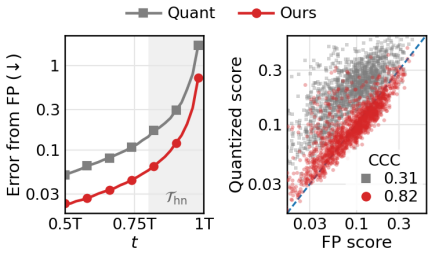

Figure 3: ADM CelebA at W6A6. (a) Clean-prediction error relative to full precision across t. (b) Per-sample SGMS scores against full precision. Gray: quantized. Red: repaired.

Why standard sampling is less affected. Standard samplers also rely on $\hat{\boldsymbol{x}}_0$ at every step, yet quantized models preserve generation quality under standard sampling. To see why, we substitute Tweedie's formula into Eq. (1) and rewrite the mean of each reverse update as (Ho et al., 2020)

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \lambda_t\, \hat{\boldsymbol{x}}_0^\theta(\boldsymbol{x}_t) + \gamma_t\, \boldsymbol{x}_t, \qquad \lambda_t = \frac{\sqrt{\alpha_{t-1}}\,\beta_t}{1-\alpha_t}, \quad \gamma_t = \frac{\sqrt{1-\beta_t}\,(1-\alpha_{t-1})}{1-\alpha_t}. \tag{8}$$

Since γ_t does not depend on the denoiser, the quantization error enters the update only through $\hat{\boldsymbol{x}}_0$. Combining with Eq. (6), we obtain

$$\boldsymbol{\mu}_{\hat{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \lambda_t\big(\hat{\boldsymbol{x}}_0^{\hat{\theta}}(\boldsymbol{x}_t) - \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_t)\big) = -\frac{\beta_t}{\sqrt{(1-\beta_t)(1-\alpha_t)}}\,\boldsymbol{\delta}(\boldsymbol{x}_t, t), \tag{9}$$

where the net factor is bounded by $\sqrt{\beta_t/(1-\beta_t)}$ for all t. Crucially, at high noise levels, the small weight λ_t offsets the amplification in Eq. (6). Conversely, λ_t approaches one at low noise levels, where $\hat{\boldsymbol{x}}_0$ remains accurate since the amplification factor vanishes. Hence, in either regime, quantization error has only a marginal impact on each reverse update. Fig. 6 in Appendix A confirms this across all t, where the update error stays below 0.02 in pixel units while the clean-prediction error exceeds 1 in $\mathcal{T}_{\text{hn}}$. In contrast, minority scores are computed directly from $\hat{\boldsymbol{x}}_0$ at high noise levels, and are thus fully exposed to the amplified error.

5 Method

In this section, we develop a lightweight repair that restores accurate high-noise clean estimates in low-bit quantized diffusion models. Guided by Sec. 4, it targets the clean-prediction discrepancy of Eq. (6) rather than the noise prediction itself, and both of its components are driven by this single quantity. We first identify the few activation quantizers that dominate this discrepancy and allocate additional bits only to them (Sec. 5.1). We then train a low-rank weight correction (Hu et al., 2022) that remains on the deployed quantization grid, with an objective that directly reduces the discrepancy and thereby bounds the error of the minority score (Sec. 5.2).

5.1 Bit Allocation Guided by Clean Predictions

A few activation sites dominate the clean-prediction error. Layers are known to differ widely in their sensitivity to quantization (Dong et al., 2019; Wang et al., 2019). Since minority-score estimation depends on the high-noise clean prediction, we measure sensitivity with respect to its discrepancy from the full-precision reference,

$$\mathcal{D}_{\text{clean}} = \mathbb{E}\left[\left\|\hat{\boldsymbol{x}}_0^{\hat{\theta}}(\boldsymbol{x}_t) - \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_t)\right\|_2^2\right] = \mathbb{E}\left[\frac{1-\alpha_t}{\alpha_t}\left\|\boldsymbol{\epsilon}_{\hat{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right\|_2^2\right], \tag{10}$$

where the expectation runs over calibration samples, timesteps $t \in \mathcal{T}_{\text{hn}}$, and forward corruption $\boldsymbol{x}_t \sim q_{\alpha_t}(\cdot \mid \boldsymbol{x}_0)$, and the second equality follows from Eq. (6). We write d_{clean} for the per-sample term inside the expectation. Restoring activation quantizers to full precision one at a time reveals that this error is concentrated in a small subset of the network, as Fig. 4 shows for ADM CelebA at W6A6. Four of the 190 sites close 89% of the gap to a model with all activations restored, whereas six further sites add less than 5%. Additional precision is therefore needed only at a handful of sites.

Sensitivity from a single backward pass. Such an exhaustive scan requires separate evaluations for every quantizer and must be repeated for each model and bit configuration, and residual magnitude alone is not a reliable substitute, since a moderate residual at an early site can propagate through the entire network. We instead estimate the effect of each quantization residual $\boldsymbol{e}_\ell = Q_\ell(\boldsymbol{a}_\ell) - \boldsymbol{a}_\ell$ on the per-sample loss d_{clean} inside Eq. (10),

$$S_\ell = \mathbb{E}\left[\left\langle \frac{\partial d_{\text{clean}}}{\partial Q_\ell(\boldsymbol{a}_\ell)}, \boldsymbol{e}_\ell \right\rangle\right], \tag{11}$$

which to first order is the error removed by eliminating $\boldsymbol{e}_\ell$ and is large only when the residual is both sizable and aligned with the direction that increases the clean-prediction error. A single backward pass per calibration batch scores all quantizers at once, and the resulting ranking closely follows that of the exhaustive scan (Appendix B).

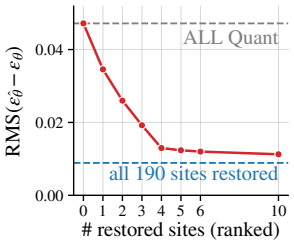

Figure 4: **A few activation sites dominate the error.** Noise-prediction RMS over $\mathcal{T}_{\text{hn}}$ as activation quantizers are restored in ranked order, ADM CelebA at W6A6.

Selective bit allocation. We select the K quantizers with the largest scores and increase their activation bit width by one. We set K at the knee of the cumulative error curve in saliency order, beyond which the marginal reduction drops by an order of magnitude. This curve matches the exhaustive one of Fig. 4 over the sites that matter (Appendix B), giving $K = 4$ for ADM at W6A6. The refined grids keep the calibrated ranges and halve the spacing at the selected sites, and the allocation is fixed.

5.2 Quantization-Consistent Minority Repair

Selective bit allocation reduces the input-dependent rounding error at sensitive activation sites, but with the quantized weights left unchanged it can be insufficient to restore the clean prediction. We therefore add a low-rank correction to the weights that stays on the deployed grid.

Correction confined to the deployed grid. A separate full-precision adapter branch breaks low-bit inference, and merging a full-precision update into the quantized weights before requantizing discards most of it, because updates smaller than the grid spacing are rounded away. We therefore confine the correction itself to the deployed grid. Recall that Q_ℓ^{w} is the frozen weight quantizer of layer ℓ with its calibrated scale and zero point. A low-rank term $\boldsymbol{B}_\ell \boldsymbol{A}_\ell$ is added to the quantized weights and the sum is projected back onto the same grid,

$$\boldsymbol{W}_\ell^{\text{repair}} = Q_\ell^{\text{w}}\left(Q_\ell^{\text{w}}(\boldsymbol{W}_\ell) + \boldsymbol{B}_\ell \boldsymbol{A}_\ell\right), \tag{12}$$

so that $\boldsymbol{W}_\ell^{\text{repair}}$ is a b-bit tensor on exactly the grid the deployed model uses, differing from $Q_\ell^{\text{w}}(\boldsymbol{W}_\ell)$ by an integer number of grid steps at every entry. On the activation side, the adapter learns a per-channel scale and offset with a per-timestep factor, applied to the input of each activation quantizer. At deployment this folds into per-channel scales and zero points of the quantizer and leaves its bit width unchanged (Appendix C). Gradients pass through the rounding with the straight-through estimator. All adapter parameters are initialized so that training starts from the deployed model, and only the low-rank factors and the activation affine parameters are updated while the base weights and weight quantizers stay frozen.

Objective for minority-signal fidelity. A minority sampler consumes the high-noise clean prediction, not the noise prediction. We therefore train the adapter on the clean-prediction discrepancy of Eq. (10), now taken over all timesteps and evaluated for the repaired denoiser $\boldsymbol{\epsilon}_{\hat{\theta},\phi}$, where ϕ collects the low-rank factors and the activation affine parameters. By Eq. (6) this equals a noise-prediction error scaled by the squared amplification, which gives

$$\mathcal{L}_{\text{adapt}} = \mathbb{E}\left[w_t \left\| \boldsymbol{\epsilon}_{\hat{\theta},\phi}(\boldsymbol{x}_t, t) - \text{sg}\left[\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right] \right\|_2^2\right], \qquad w_t = \min\left(\frac{1-\alpha_t}{\alpha_t}, w_{\max}\right), \tag{13}$$

where sg denotes stop-gradient and the expectation runs over calibration samples, all timesteps, and forward corruption. The cap $w_{\max}$ only down-weights the most amplified timesteps. The objective is

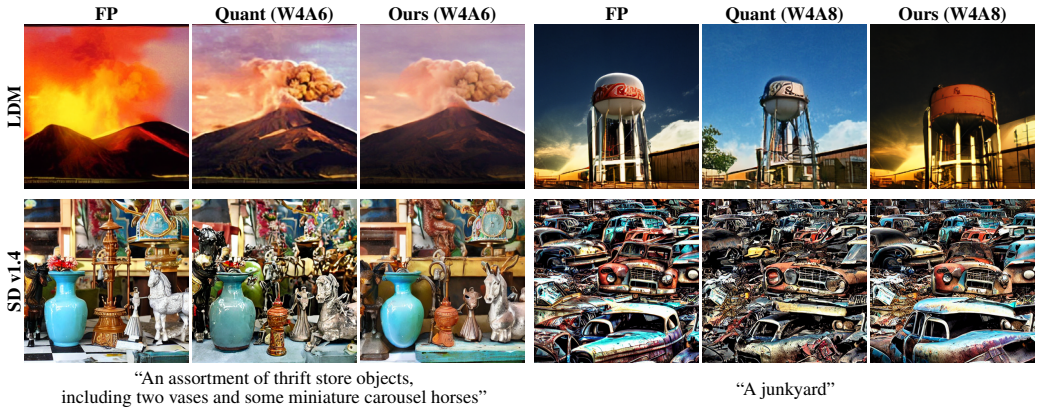

Figure 5: Qualitative comparison of minority generation across architectures and quantization settings. Rows show LDM under SGMS and Stable Diffusion under BnS with full-precision, quantized and repaired models, with bit configurations indicated above. ImageNet classes are "volcano" and "water tower" for LDM, and the MS-COCO prompt is given below each triplet for Stable Diffusion.

thus a distillation in clean-prediction space rather than noise-prediction space, and it spends adapter capacity where the minority signal is most sensitive.

The link to the minority score is direct. Whenever the discrepancy d of a sampler is C-Lipschitz in its second argument, the score gap between the repaired and full-precision models obeys

$$\left|\mathcal{L}_{\hat{\theta},\phi}(\boldsymbol{x}_0;s)-\mathcal{L}_{\theta}(\boldsymbol{x}_0;s)\right| \le C\,\mathbb{E}_{q_{\alpha_s}(\boldsymbol{x}_s|\boldsymbol{x}_0)}\left[\left\|\hat{\boldsymbol{x}}_0^{\hat{\theta},\phi}(\boldsymbol{x}_s)-\hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_s)\right\|_2\right], \tag{14}$$

so the gap is controlled by the clean-prediction gap at the scoring level s, one of the per-timestep terms that $\mathcal{L}_{\text{adapt}}$ drives down. Appendix D gives the proof and extends the bound to the two-stage score of SGMS and the ordinal labels of MG.

Plug-in deployment. Since Eq. (12) returns weights on the deployed grid, the repaired denoiser is itself a b-bit model that runs on the unchanged low-bit kernels, either fused as a second b-bit weight set or rebuilt from the stored low-rank factors by a single projection (Appendix C.1). We store only the factors, whose size is reported with the results. For score-based samplers, only the minority-signal path uses the repaired weights and the refined activation grids, so standard sampling is unchanged, and Table 4 confirms that the repair adds no measurable latency.

Extension to guidance-free minority samplers. A natural question is whether the repair is limited to samplers built on minority scores. We find that the same repair, without retraining, also improves guidance-free minority samplers such as Boost-and-Skip (Um et al., 2025), where the repaired denoiser replaces the quantized one throughout the reverse process (Table 2). Since the repaired denoiser is itself a b-bit model, this incurs no loss in low-bit efficiency.

We attribute this benefit to the variance-boosted initialization of BnS. Its initial inputs $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0},\gamma^2\boldsymbol{I})$ with $\gamma > 1$ have larger magnitudes than the noisy states on which the quantizers are calibrated, inducing additional clipping error and hence a larger $\boldsymbol{\delta}$ at the highest noise levels, where the amplification in Eq. (6) is also the largest. By aligning the high-noise clean estimates with their full-precision counterparts, the repair mitigates this error as Appendix E verifies across γ.

6 Experiments

Models and datasets. We evaluate our method on diverse diffusion architectures: the pixel-space ADM (Dhariwal & Nichol, 2021) for unconditional generation, the latent-space LDM-4 (Rombach et al., 2022) and the transformer-based DiT-XL/2 (Peebles & Xie, 2023) for class-conditional generation, and Stable Diffusion v1.4 (Rombach et al., 2022) for text-to-image generation. We use four benchmarks: CelebA 64×64 and LSUN Bedrooms 256×256 for ADM, ImageNet 256×256 for LDM-4 and DiT-XL/2, and MS-COCO 512×512 for Stable Diffusion.

Quantization baselines. We build on the official implementations of recent diffusion quantization methods: TFMQ (Huang et al., 2024) for ADM, DMQ (Lee et al., 2025) for LDM-4 and Stable Diffusion, and TQ-DiT (Hwang et al., 2025) for DiT-XL/2. Bit configurations follow each implementation and are reported in the tables, where WxAy denotes x-bit weights and y-bit activations.

Table 1: Minority generation across backbones, quantization methods and bit widths using the minority-score-based samplers SGMS and MG. Size for Ours includes the stored adapter.

Model	Bits	Method	cFID↓	sFID↓	Prec.↑	Rec.↑	Size (MB)
ADM CelebA 64×64	FP	Full Precision + SGMS	19.80	8.90	0.851	0.551	1179.7
	W6A6	TFMQ + SGMS	27.89	9.38	0.775	0.511	221.2
		TFMQ + SGMS + Ours	**19.46**	**7.93**	**0.781**	**0.548**	240.7
	W4A8	TFMQ + SGMS	16.07	8.21	0.774	0.574	147.4
		TFMQ + SGMS + Ours	**15.56**	**7.48**	**0.788**	**0.584**	166.9
ADM LSUN Bedrooms 256×256	FP	Full Precision + SGMS	23.31	17.17	0.801	0.400	2210.1
	W6A6	TFMQ + SGMS	36.84	23.73	0.652	0.224	414.4
		TFMQ + SGMS + Ours	**28.33**	**20.69**	**0.752**	**0.304**	441.4
	W4A8	TFMQ + SGMS	28.60	18.58	0.701	0.280	276.2
		TFMQ + SGMS + Ours	**25.42**	**17.49**	**0.782**	**0.348**	303.2
LDM-4 ImageNet 256×256	FP	Full Precision + SGMS	56.40	8.21	0.966	0.181	1560.8
	W4A6	DMQ + SGMS	58.94	8.79	**0.960**	0.174	195.1
		DMQ + SGMS + Ours	**52.14**	**5.39**	0.955	**0.195**	218.5
	W4A8	DMQ + SGMS	58.27	9.44	**0.962**	0.165	195.1
		DMQ + SGMS + Ours	**55.40**	**8.58**	0.961	**0.175**	218.5
DiT-XL/2 ImageNet 256×256	FP	Full Precision + MG	37.26	13.72	0.902	0.302	1784.0
	W6A6	TQ-DiT + MG	42.04	15.07	0.858	0.255	334.5
		TQ-DiT + MG + Ours	**38.38**	**14.15**	**0.872**	**0.317**	352.0
	W8A8	TQ-DiT + MG	37.72	13.90	0.887	0.324	446.0
		TQ-DiT + MG + Ours	**37.24**	**13.65**	**0.896**	**0.335**	463.5

Minority samplers. We consider the score-based MG and SGMS and the guidance-free BnS, and for each compare the full-precision model, the quantized baseline, and the quantized with our repair.

Evaluation. For unconditional and class-conditional generation, we follow prior work (Um et al., 2024; Um & Ye, 2024; Um et al., 2025) and report cFID (Parmar et al., 2022), sFID (Nash et al., 2021), and precision and recall (Kynkäänniemi et al., 2019), computed against a reference set of real minority samples. For text-to-image generation, we follow the minority protocol of Um & Ye (2025). We measure text alignment and human preference with CLIPScore (Hessel et al., 2021), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023), and report precision, recall, density, and coverage (Naeem et al., 2020) against all MS-COCO validation images. To quantify rarity, we further report the log-likelihood (LL) of each sample under the full-precision Stable Diffusion v1.4, computed with the probability-flow ODE (Song et al., 2021), where lower LL indicates rarer samples. Further details are provided in Appendix F.

Implementation. We apply our repair on top of each quantized baseline with its weights and weight quantizers frozen, and the refined grids and adapter are used only where enabled, so standard sampling runs on the stock quantized model. The adapter is trained once per model and bit width on random training images and shared across samplers; following Sec. 5, it is enabled only for minority-score estimation in MG and SGMS and throughout the reverse process in BnS. All compared methods share the initial noise and conditioning; details are in Appendix G.

6.1 Results

Qualitative comparisons. Fig. 5 compares full-precision, quantized and repaired models on LDM under SGMS and Stable Diffusion under BnS. At W4A6 the quantized models preserve minority characteristics but show noticeable artifacts and structural distortions, which our repair substantially reduces; at W4A8 the degradation is milder, and the repair still corrects residual artifacts and incoherent details. Results for all architectures and samplers are in Appendix L.

Quantitative evaluation. Table 1 evaluates SGMS, which builds self-guidance from the reconstruction discrepancy, on ADM and LDM, and MG, which guides DiT with a classifier trained on reconstruction-based minority levels. Our repair consistently improves quality and diversity of the quantized models at comparable precision. Two further patterns are worth noting: the degradation is milder at 8-bit activations, and the repaired model occasionally scores below full precision in cFID, both of which Appendix H explains from Eq. (6). Table 2 evaluates BnS on Stable Diffusion v1.4 under the protocol of Um & Ye (2025), where a minority sampler is judged by how far it lowers LL while keeping alignment and preference scores. BnS was reported to fail with the probability-flow ODE, yet we find that it transfers to the DDIM sampler with a small boost ($\gamma \leq 1.06$), and Ap-

Table 2: Minority generation without minority scores using BnS on Stable Diffusion v1.4 and MS-COCO. CLIP denotes CLIPScore, Pick denotes PickScore and IR denotes ImageReward. Den./Cov. denote density/coverage and LL denotes log-likelihood with lower values indicating rarer samples.

Bits	Method	CLIP↑	Pick↑	IR↑	Prec.↑	Rec.↑	Den.↑	Cov.↑	LL↓
FP	Full Precision + BnS	0.3143	21.310	0.261	0.676	0.711	0.734	0.1625	0.7636
W4A6	DMQ + BnS	0.3113	20.982	0.085	0.610	**0.728**	0.643	0.1369	0.7285
	DMQ + Ours + BnS	**0.3139**	**21.032**	**0.178**	**0.698**	0.701	**0.760**	**0.1439**	**0.7195**
W4A8	DMQ + BnS	0.3138	21.284	0.260	0.669	0.587	0.759	0.1519	0.7883
	DMQ + Ours + BnS	**0.3148**	**21.303**	**0.263**	**0.682**	**0.601**	**0.767**	**0.1534**	**0.7876**

Table 3: Component ablation on CelebA with TFMQ and SGMS.

Bits	Grid	Adapter	cFID↓	sFID↓	Prec.↑	Rec.↑
W6A6	×	×	33.39	19.35	0.758	0.514
	✓	×	30.54	18.32	0.779	0.558
	×	✓	29.65	18.27	0.768	0.561
	✓	Plain	27.57	18.19	0.772	0.565
	✓	✓	**25.69**	**17.97**	**0.780**	**0.577**
W4A8	×	×	22.31	18.24	0.765	0.578
	✓	×	22.17	17.71	0.786	0.588
	×	✓	21.91	17.66	0.790	0.579
	✓	Plain	21.67	17.62	0.794	0.578
	✓	✓	**21.34**	**17.51**	**0.797**	**0.592**

Table 4: Sampling latency on LDM-4 ImageNet with SGMS and DMQ.

NVIDIA RTX A6000 48GB			
Method	Time (s/image)	Speedup	Peak Mem.
Full precision	1.290	1.00×	2.45GB
W4A8	0.486	2.65×	1.43GB
W4A8 + Ours	0.487	2.65×	1.76GB
Jetson Orin Nano 8GB			
Method	Time (s/image)	Speedup	Peak Mem.
Full precision	17.888	1.00×	2.44GB
W4A8	5.155	3.47×	1.42GB
W4A8 + Ours	5.213	3.43×	1.78GB

pendix G.5 shows that both follow from the analysis of Um et al. (2025). Since BnS has no score branch, the same adapter is applied throughout its trajectory without retraining. The repair improves every alignment and preference metric at both bit widths and lowers LL relative to the stock quantized model, so its gains do not trade rarity for quality and extend beyond minority-score-based sampling. BnS results on DiT are in Appendix I.

6.2 Ablation Study

Component ablation. Table 3 evaluates activation-grid refinement and the quantization-consistent adapter under different bit-width settings. Grid refinement protects the activation quantizers most responsible for high-noise clean-prediction errors, while the adapter corrects the residual errors throughout the quantized denoiser, and combining them yields the most consistent improvements. Appendix J reports the same ablation on LDM. The Plain rows isolate the effect of the objective by training the same adapter with uniform noise-prediction distillation ($w_t \equiv 1$ in Eq. (13)). The clean-prediction weighting improves every metric over this baseline at both bit widths, with the largest margin at W6A6 where the amplified error is greatest, confirming that spending adapter capacity where the minority signal is most sensitive matters beyond adding an adapter alone.

Inference speed and trade-offs. Table 4 reports end-to-end SGMS latency for LDM-4 on an RTX A6000 and a Jetson Orin Nano 8GB. Because the repaired weights lie on the deployed grid, the score branch runs on the same low-bit kernels as the base model with no additional forward pass, so latency is unchanged and only peak memory grows by the second weight set. The W4A8 speedup over full precision is thus retained on both devices. Details of the int8 runtime and storage overhead are given in Appendices C and C.1, and trade-off curves over the minority-specific strength for CelebA and MS-COCO in Appendix K.

7 Conclusion

We showed that a quantized diffusion model can preserve standard generation yet fail at minority generation, because minority scores depend on high-noise clean predictions where the Tweedie map amplifies denoiser error that the reverse update otherwise attenuates. Our plug-in repair refines the few activation grids responsible for most of this error and trains a quantization-consistent adapter to match the full-precision clean prediction, leaving the base weights and the standard sampling path untouched while restoring minority generation across diverse models, bit widths and samplers.

Limitations. The repair adds little storage, keeping only the low-rank factors, but accelerated inference holds the base and repaired weights together, raising peak memory by one low-bit model. It is also trained per model and bit configuration, and is evaluated only on image diffusion models.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods: an LLM assistant helped write parts of the code for the adapter training and the selective bit allocation. We have not used generative AI tools for formulating mathematical claims, assisting with proofs, designing the methodology or experiments, or interpreting results, and generating synthetic data is not applicable to this work. Additionally, we used generative AI tools for editing the manuscript for readability and flow. We have reviewed all AI-assisted work. All LLM-assisted code was executed by the authors, reviewed line by line and tested for correctness before use, and all edited text was checked by the authors for faithfulness to the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Anonymized source code for our method, including the selective bit allocation and the adapter training, is available at the link given in the abstract. Full implementation details, including the quantized baselines, bit allocation, adapter training hyperparameters and sampler settings, are given in Appendix G, and the evaluation protocol in Appendices F and G.6. The proof of Eq. (14) is provided in Appendix D.

REFERENCES

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28306–28315. IEEE, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 293–302, 2019.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022.

Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgO66VKDS.

Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score : A new metric to evaluate the uncommonness of synthesized images. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JTGimap_-F.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. *Advances in neural information processing systems*, 36:13237–13249, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

Younghye Hwang, Hyojin Lee, and Joonhyuk Kang. Tq-dit: Efficient time-aware quantization for diffusion transformers. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Hyunsoo Kim, Jungmyung Wi, Soobin Um, Donghyun Kim, and Suhyun Kim. Self-improving diffusion classifiers with minority preference optimization. In *European Conference on Computer Vision*, pp. 285–303. Springer, 2026.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Dongyeun Lee, Jiwan Hur, Hyounguk Shon, Jae Young Lee, and Junmo Kim. Dmq: Dissecting outliers of diffusion models for post-training quantization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18510–18520. IEEE, 2025.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17489–17499. IEEE, 2023.

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. URL https://openreview.net/forum?id=POWv6hDd9XH.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

Sol Park and Soobin Um. Beyond generative priors: Minority sampling with jepa-guided diffusion. *arXiv preprint arXiv:2605.24631*, 2026.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11400–11410. IEEE, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in statistics*, pp. 388–394. Springer, 1992.

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.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 11492–11501, 2022.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1972–1981. IEEE, 2023.

Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. *Advances in neural information processing systems*, 36:48686–48698, 2023.

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*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *European Conference on Computer Vision*, pp. 414–430. Springer, 2024.

Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20926–20936. IEEE, 2025.

Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 21632–21670, 2024.

Soobin Um, Beomsu Kim, and Jong Chul Ye. Boost-and-skip: A simple guidance-free diffusion for minority generation. *arXiv preprint arXiv:2502.06516*, 2025.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8604–8612. IEEE, 2019.

Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Ptq4dit: Post-training quantization for diffusion transformers. *Advances in neural information processing systems*, 37: 62732–62755, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Enshu Liu, Rui Wan, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, et al. Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. *arXiv preprint arXiv:2406.02540*, 2024.

A Quantization Error Along the Reverse Process

Fig. 6 decomposes the effect of quantization along the reverse process on ADM CelebA at W6A6, plotting the per-dimension error of the quantized model relative to full precision for the noise prediction $\boldsymbol{\delta}$, the clean prediction $\hat{\boldsymbol{x}}_0$ and the reverse-update mean $\boldsymbol{\mu}$ across all timesteps. The raw error $\boldsymbol{\delta}$ is largest at low noise levels, where the noise is barely observable in $\boldsymbol{x}_t$ and its prediction is ill-conditioned, and settles near 0.05 for $t \geq 0.5T$. The Tweedie map of Eq. (6) amplifies it by more than an order of magnitude in $\mathcal{T}_{\text{hn}}$, which is the error that minority scores consume. The reverse update of Eq. (9) instead scales it by a factor that stays near 0.01 over most of the trajectory and rises only in the final steps where β_t of the cosine schedule is clipped, so the update error remains below 0.02 in pixel units throughout. The rise near $t = 0$ reflects the larger $\boldsymbol{\delta}$ there rather than the schedule, and the two curves coincide at $t = 1$ where $\lambda_t = 1$.

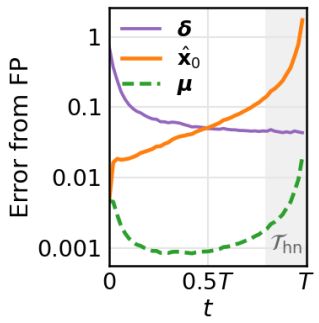

Figure 6: Quantization error along the reverse process on ADM CelebA at W6A6. Per-dimension error of the quantized model relative to full precision for the noise prediction $\boldsymbol{\delta}$, the clean prediction $\hat{\boldsymbol{x}}_0$ and the reverse-update mean $\boldsymbol{\mu}$. The clean-prediction error is $\boldsymbol{\delta}$ amplified by $\sqrt{(1-\alpha_t)/\alpha_t}$, whereas the update error stays within $\sqrt{\beta_t/(1-\beta_t)}$ of $\boldsymbol{\delta}$ and below 0.02 for all t.

B Saliency Ranking versus Exhaustive Scan

Sec. 5.1 replaces the exhaustive scan of Fig. 4, which restores one activation quantizer to full precision at a time and re-evaluates $\mathcal{D}_{\text{clean}}$, with the first-order saliency of Eq. (11) obtained from a single backward pass. Table 5 compares the two rankings on ADM CelebA at W6A6 over the 190 candidate sites. The two criteria select the same four sites in the same order, namely the input convolution and the three highest-resolution skip connections of the decoder, which are exactly the sites that carry concatenated features into the last output blocks. Over the sites that matter the agreement is also close in rank: the Spearman correlation between the two rankings is 0.79 within the exhaustive top-10 and 0.55 within the top-20. Over all 190 sites it drops to 0.37 (Kendall τ 0.26, $p = 4 \times 10^{-7}$), but this is expected rather than a failure of the criterion. Beyond the top-10 the exhaustive gains are on the order of 10^{-6}, i.e. at the level of evaluation noise, so the tail ordering of either criterion carries no information. Since K is set at the knee of the cumulative curve, which lies well inside the top-10, the saliency ranking reproduces the bit allocation of the exhaustive scan at a small fraction of its cost.

C Latency Measurement and Low-Bit Runtime

The released DMQ code evaluates the quantized model through fake quantization, i.e. an fp32 GEMM preceded by a dozen elementwise quantize–dequantize operations per layer, which reproduces the numerics of low-bit inference but not its speed. Table 4 therefore uses an int8 runtime that we built for DMQ on LDM-4 and that runs the same kernels, modules, sampler and protocol on the RTX A6000 and the Jetson Orin Nano. We describe it here because it is also what makes the

Table 5: Top-4 activation sites on ADM CelebA (W6A6) under the exhaustive scan and the saliency of Eq. (11). Both criteria select the same sites in the same order. Spearman correlation between the full rankings is 0.79 within the exhaustive top-10, 0.55 within the top-20 and 0.37 over all 190 sites.

Rank	Site	Saliency rank	Exhaustive rank
1	`input_blocks.0.0`	1	1
2	`output_blocks.15.0.skip_connection`	2	2
3	`output_blocks.14.0.skip_connection`	3	3
4	`output_blocks.13.0.skip_connection`	4	4

overhead of the repair vanish: all three components of Sec. 5 fold into the same static representation as the base model, so the repaired model runs exactly the same kernels the same number of times.

Weights. The AdaRound codes of the deployed model are computed once at load time and stored in an int8 container, with the per-channel equivalent scaling of DMQ absorbed into the codes. We precompute the per-channel weight scales and the column sums of the codes, which the epilogue uses for zero-point correction. The power-of-two channel scales that DMQ applies to skip-connection inputs are folded into the corresponding weight columns, with a per-row rescaling whenever folding would leave the int8 range.

Activations. DMQ divides the timesteps into 20 buckets, each with its own scale and zero point per quantizer. The released code re-uploads several hundred tensors from CPU to GPU on every forward; we keep all 20 tables resident on the GPU and change a single bucket index per forward. Quantization happens inside a Triton kernel that reads the fp16 input, divides by the equivalent scale and the step size, rounds with the same mode as PyTorch, clamps and writes int8 directly, so no fp32 intermediate is materialized. For 3×3 convolutions the input is quantized once per element into NHWC int8 and a second kernel gathers those bytes into im2col columns, so each element is quantized once rather than once per tap and the im2col buffer is a quarter of its fp32 size.

GEMM and epilogue. Products are int8$\times$int8 with int32 accumulation on tensor cores. Dequantization is done entirely in the GEMM epilogue in a single kernel: zero-point correction from the precomputed column sums, activation scale times weight scale, bias, and for split layers the accumulation of both segments, written directly as fp16. A linear layer thus costs two kernels and a 3×3 convolution three. Since DMQ does not quantize attention activations for this model, attention and normalization run in plain fp16.

Backward of the score branch. SGMS needs $\partial\boldsymbol{\epsilon}/\partial\boldsymbol{x}_t$ through the UNet, which the released code obtains by differentiating the fake-quantization graph with autograd. We instead give each layer a custom autograd function that saves only its fp16 input. The backward computes the input gradient with the dequantized weights and then applies the straight-through mask in one kernel, recomputing the mask from the saved input and the resident bucket table. In the low-memory layout used on the Jetson, no fp16 weight copy is kept for the backward: convolutions dequantize on the fly with a small kernel and linear layers convert int8 weights to fp16 inside the GEMM kernel.

Launch overhead. After these changes the GPU work per forward is small, but at batch size 1 a forward issues about 1,500 small kernels and the CPU launch time exceeds the GPU time. Two further steps remove it. First, the UNet runs in pure fp16 without autocast, with only the time embedding and the output head kept in fp32, which eliminates roughly 250 casting copies introduced by autocast and the fp32 GroupNorm wrapper of the LDM code. Second, every UNet call in the protocol has fixed shapes, timesteps and parameter sets, so on one warm-up image we capture the 20 generation steps and the 20 score-branch events as CUDA graphs and only replay them afterwards. A score-branch graph contains the gradient-enabled forward, the re-corruption, the second forward and the backward. The 40 graphs share one memory pool, which is possible because they are replayed in a fixed sequential order.

Why the repair adds no cost. The low-rank correction is added to the dequantized weights and rounded back onto the AdaRound grid as in Eq. (12), so at deployment it is nothing but a second set of int8 codes. The activation affine of Sec. 5.2 folds away completely: its per-channel scale into the weight columns, its per-bucket scalar into the activation scale table and its shift into the bias, where for 3×3 convolutions the shift becomes a per-pixel bias map precomputed once per resolution. The

K sites with refined 9-bit activation grids do not fit int8 tensor cores, so the same GEMM kernel runs in fp16-operand mode for those layers only. Selecting the repaired set is a single flag that swaps weight pointers: the sampler enables it inside the score branch at every step and the generation path never touches the second set. Since graphs are captured per set, a score-branch event is simply the replay of a different graph. Between the base and the repaired runtime the kernels, their count and the activation buffers are identical, which is why the two W4A8 rows of Table 4 coincide within timing noise.

Peak memory. The int8 runtime is compact by construction: one byte per weight, fp16 activations quantized straight to int8, a backward that stores only layer inputs, and resident bucket tables. The full-precision row is smaller than its parameter count suggests because this UNet has 395M parameters and, at batch size 1 on 64×64 latents with single-head attention over at most 1,024 tokens, activations are smaller than weights; the released code also checkpoints its transformer blocks and recomputes intermediates in the backward, so its peak is the fp32 weights plus a transient rise during the score-branch backward. The int8 runtime instead keeps its working set resident (graph pool, bucket tables and, in the default layout, an fp16 weight copy for the backward and per-bucket weight copies for the folded skip scales), trading memory for launch-free replay. The repair adds one more static block, the second code set and its scale tables, and no transient memory, since both sets share the same activation buffers and graph pool. The low-memory layout on the Jetson drops the backward weight copy and the per-bucket copies in favor of on-the-fly dequantization and in-kernel column scaling at a cost of a few percent in speed.

Jetson deployment. The patched model is exported from the server as a bundle without a checkpoint. On the board the module structure is instantiated on the meta device and tensors are streamed one at a time from a memory-mapped file to the GPU, so host memory peaks at the size of a single tensor. On devices with less than 16 GB we disable Triton autotuning and fix one GEMM configuration, since the benchmark buffers of autotuning exceed the unified memory, and graph capture falls back to eager execution if it runs out of memory. The speedup is larger on the Jetson than on the A6000 for a structural reason. At batch size 1 the A6000 is never saturated, so much of the int8 gain is hidden behind the remaining launch and synchronization cost; the Orin is compute- and bandwidth-bound, so the int8 tensor-core throughput and the four-fold reduction in weight traffic translate directly into wall-clock time.

C.1 Storage Cost

At deployment we store the base quantized weights once and, for the adapted layers, either the low-rank factors or the requantized repaired weights of Eq. (12). Because the repaired weights lie on the deployed grid, both forms execute on the unchanged low-bit kernels. Since every weight-quantized module is adapted, the fused form is a second b-bit weight set of the same size as the base model; the latency measurements of Table 4 use this form, and the peak-memory difference there reflects the second set held alongside the base model. The factor form is far smaller: Table 11 reports the base quantized size, the low-rank factors in FP16 (they can also be stored in INT8 at half the size), the activation affine parameters, the trainable parameter count and the resulting storage overhead. The activation-grid refinement adds only one scale and zero point per selected site.

D Score Fidelity Bound

This appendix makes precise the relation between the adapter objective of Eq. (13) and the minority score of Eq. (3) that Eq. (14) summarizes in the main text. Throughout, $\|\cdot\|$ denotes the ℓ_2 norm and all expectations over $\boldsymbol{x}_t$ are taken under the forward law $q_{\alpha_t}(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ of Section 3.

Notation. Fix a data sample $\boldsymbol{x}_0$ and a timestep t. Write $\hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_t)$ for the Tweedie estimate of the full-precision denoiser and $\hat{\boldsymbol{x}}_0^{\phi}(\boldsymbol{x}_t)$ for that of the repaired denoiser $\boldsymbol{\epsilon}_{\hat{\theta},\phi}$, evaluated with the selected activation grids and the weights of Eq. (12). Since Eq. (2) is linear in the noise prediction, the two estimates differ by

$$\hat{\boldsymbol{x}}_0^{\phi}(\boldsymbol{x}_t) - \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_t) = -\sqrt{\frac{1-\alpha_t}{\alpha_t}}\,\boldsymbol{\delta}_{\phi}(\boldsymbol{x}_t, t), \qquad \boldsymbol{\delta}_{\phi}(\boldsymbol{x}_t, t) = \boldsymbol{\epsilon}_{\hat{\theta},\phi}(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}_{\theta}(\boldsymbol{x}_t, t), \tag{15}$$

which is Eq. (6) with the repaired denoiser in place of the deployed one. We define the per-sample clean-prediction discrepancy at timestep t as

$$\mathcal{D}_t(\boldsymbol{x}_0;\phi) = \mathbb{E}_{\boldsymbol{x}_t}\left[\left\|\hat{\boldsymbol{x}}_0^{\phi}(\boldsymbol{x}_t) - \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_t)\right\|^2\right] = \frac{1-\alpha_t}{\alpha_t}\,\mathbb{E}_{\boldsymbol{x}_t}\left[\|\boldsymbol{\delta}_\phi(\boldsymbol{x}_t,t)\|^2\right]. \tag{16}$$

With $\phi = 0$ and the deployed activation grids, Eq. (12) returns $Q_\ell^{\mathrm{w}}(\boldsymbol{W}_\ell)$ because the quantizer is idempotent on its own grid, and $\mathcal{D}_t(\boldsymbol{x}_0;0)$ is the summand of Eq. (10). The sensitivity measure of the main text is therefore $\mathcal{D}_{\mathrm{clean}} = \mathbb{E}_{\boldsymbol{x}_0}\,\mathbb{E}_{t\in\mathcal{T}_{\mathrm{hn}}}\left[\mathcal{D}_t(\boldsymbol{x}_0;0)\right]$.

The adapter objective of Eq. (13) draws t uniformly from $\{1,\dots,T\}$ and can be written in the same terms. Since $w_t\|\boldsymbol{\delta}_\phi\|^2 = c_t\,\frac{1-\alpha_t}{\alpha_t}\|\boldsymbol{\delta}_\phi\|^2$ with

$$c_t = \min\left(1,\ w_{\max}\,\frac{\alpha_t}{1-\alpha_t}\right) \in (0,1], \tag{17}$$

we have

$$\mathcal{L}_{\mathrm{adapt}}(\phi) = \frac{1}{T}\sum_{t=1}^{T} c_t\,\mathbb{E}_{\boldsymbol{x}_0}\left[\mathcal{D}_t(\boldsymbol{x}_0;\phi)\right], \tag{18}$$

and $c_t = 1$ for every timestep whose amplification $(1-\alpha_t)/\alpha_t$ does not exceed $w_{\max}$. The objective is thus a sum of per-timestep clean-prediction discrepancies in which only the most amplified timesteps are down-weighted.

Finally, for a scoring level s and a discrepancy d, the minority scores of the two models are

$$\mathcal{L}_\theta(\boldsymbol{x}_0;s) = \mathbb{E}_{\boldsymbol{x}_s}\left[d\big(\boldsymbol{x}_0,\hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_s)\big)\right], \qquad \mathcal{L}_{\hat\theta,\phi}(\boldsymbol{x}_0;s) = \mathbb{E}_{\boldsymbol{x}_s}\left[d\Big(\boldsymbol{x}_0,\hat{\boldsymbol{x}}_0^{\phi}(\boldsymbol{x}_s)\Big)\right]. \tag{19}$$

Assumption. There is a constant C such that for every $\boldsymbol{x}_0$ and every pair $\boldsymbol{a},\boldsymbol{b}$,

$$|d(\boldsymbol{x}_0,\boldsymbol{a}) - d(\boldsymbol{x}_0,\boldsymbol{b})| \le C\,\|\boldsymbol{a}-\boldsymbol{b}\|. \tag{20}$$

Proposition 1. Under Eq. (20), for every $\boldsymbol{x}_0$ and every scoring level s,

$$\left|\mathcal{L}_{\hat\theta,\phi}(\boldsymbol{x}_0;s) - \mathcal{L}_\theta(\boldsymbol{x}_0;s)\right| \le C\,\mathbb{E}_{\boldsymbol{x}_s}\left[\left\|\hat{\boldsymbol{x}}_0^{\phi}(\boldsymbol{x}_s) - \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x}_s)\right\|\right] \le C\sqrt{\mathcal{D}_s(\boldsymbol{x}_0;\phi)}. \tag{21}$$

Averaging over data,

$$\mathbb{E}_{\boldsymbol{x}_0}\left|\mathcal{L}_{\hat\theta,\phi}(\boldsymbol{x}_0;s) - \mathcal{L}_\theta(\boldsymbol{x}_0;s)\right| \le C\sqrt{\mathbb{E}_{\boldsymbol{x}_0}\left[\mathcal{D}_s(\boldsymbol{x}_0;\phi)\right]} \le C\sqrt{\frac{T}{c_s}\,\mathcal{L}_{\mathrm{adapt}}(\phi)}. \tag{22}$$

Proof. Both scores in Eq. (19) are expectations under the same law $q_{\alpha_s}(\boldsymbol{x}_s \mid \boldsymbol{x}_0)$, so

$$\left|\mathcal{L}_{\hat\theta,\phi} - \mathcal{L}_\theta\right| = \left|\mathbb{E}_{\boldsymbol{x}_s}\left[d(\boldsymbol{x}_0,\hat{\boldsymbol{x}}_0^{\phi}) - d(\boldsymbol{x}_0,\hat{\boldsymbol{x}}_0^{\theta})\right]\right| \le \mathbb{E}_{\boldsymbol{x}_s}\left|d(\boldsymbol{x}_0,\hat{\boldsymbol{x}}_0^{\phi}) - d(\boldsymbol{x}_0,\hat{\boldsymbol{x}}_0^{\theta})\right| \le C\,\mathbb{E}_{\boldsymbol{x}_s}\left\|\hat{\boldsymbol{x}}_0^{\phi} - \hat{\boldsymbol{x}}_0^{\theta}\right\|, \tag{23}$$

where the first inequality is Jensen's inequality for the absolute value and the second is Eq. (20). The square root is concave, so Jensen's inequality gives $\mathbb{E}\|\boldsymbol{v}\| \le \sqrt{\mathbb{E}\|\boldsymbol{v}\|^2}$ for any random vector $\boldsymbol{v}$, and Eq. (16) turns this into the second inequality of Eq. (21). For Eq. (22), apply Jensen's inequality once more to the outer expectation over $\boldsymbol{x}_0$. Then observe that Eq. (18) is a sum of nonnegative terms that contains $\frac{c_s}{T}\,\mathbb{E}_{\boldsymbol{x}_0}[\mathcal{D}_s(\boldsymbol{x}_0;\phi)]$, so $\mathbb{E}_{\boldsymbol{x}_0}[\mathcal{D}_s(\boldsymbol{x}_0;\phi)] \le \frac{T}{c_s}\mathcal{L}_{\mathrm{adapt}}(\phi)$. □

Constants for the discrepancies used in practice. The bound holds for any d that satisfies Eq. (20). The samplers we evaluate use the following, and none of them requires the clean predictions to be clipped or otherwise bounded.

Distance in ℓ_2. For $d(\boldsymbol{x}_0,\boldsymbol{a}) = \|\boldsymbol{x}_0 - \boldsymbol{a}\|$ the triangle inequality gives $C = 1$.

LPIPS. LPIPS passes both inputs through a fixed convolutional extractor ψ, normalizes each channel vector as $\boldsymbol{u} \mapsto \boldsymbol{u}/(\|\boldsymbol{u}\| + \varepsilon)$ with $\varepsilon > 0$, and averages the squared differences of the normalized features with fixed nonnegative weights. Convolutions and ReLUs are globally Lipschitz, the normalization is globally Lipschitz with a constant of order $1/\varepsilon$, and the normalized features lie in the

unit ball, on which the squared difference is 2-Lipschitz. The map $\boldsymbol{a} \mapsto d(\boldsymbol{x}_0, \boldsymbol{a})$ is therefore globally Lipschitz in its second argument with a finite constant that depends only on ψ and ε, so Eq. (20) holds without any restriction on the range of the clean predictions.

Latent-space samplers. For LDM, DiT, and Stable Diffusion the denoiser and the adapter act on latents, and the score compares $\boldsymbol{x}_0$ with the decoded clean prediction $D(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_s))$. The decoder D is a fixed network of Lipschitz layers, so $d \circ D$ satisfies Eq. (20) on the latent domain with C replaced by $C\,\mathrm{Lip}(D)$, and Proposition 1 holds with the latent clean-prediction discrepancy, which is what $\mathcal{L}_{\mathrm{adapt}}$ controls for these models.

Two-stage score of self-guidance. Self-guided minority sampling scores the current state through $d(\boldsymbol{a}, \hat{\boldsymbol{x}}_0(\tilde{\boldsymbol{x}}_s))$, where $\boldsymbol{a} = \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ is the current Tweedie estimate and $\tilde{\boldsymbol{x}}_s = \sqrt{\alpha_s}\,\boldsymbol{a} + \sqrt{1-\alpha_s}\,\boldsymbol{\epsilon}$ is its re-corruption. Both arguments now depend on the model. Suppose d satisfies Eq. (20) in each argument and the full-precision Tweedie map $\boldsymbol{x} \mapsto \hat{\boldsymbol{x}}_0^{\theta}(\boldsymbol{x})$ at level s is L_s-Lipschitz. Write $\boldsymbol{a}^{\phi}, \boldsymbol{a}^{\theta}$ for the two current estimates and $\tilde{\boldsymbol{x}}^{\phi}, \tilde{\boldsymbol{x}}^{\theta}$ for their re-corruptions with the same $\boldsymbol{\epsilon}$. Adding and subtracting $d(\boldsymbol{a}^{\theta}, \hat{\boldsymbol{x}}_0^{\phi}(\tilde{\boldsymbol{x}}^{\phi}))$ and $d(\boldsymbol{a}^{\theta}, \hat{\boldsymbol{x}}_0^{\theta}(\tilde{\boldsymbol{x}}^{\phi}))$ gives

$$\begin{aligned}\left| d(\boldsymbol{a}^{\phi}, \hat{\boldsymbol{x}}_0^{\phi}(\tilde{\boldsymbol{x}}^{\phi})) - d(\boldsymbol{a}^{\theta}, \hat{\boldsymbol{x}}_0^{\theta}(\tilde{\boldsymbol{x}}^{\theta})) \right| &\le C\|\boldsymbol{a}^{\phi} - \boldsymbol{a}^{\theta}\| + C\|\hat{\boldsymbol{x}}_0^{\phi}(\tilde{\boldsymbol{x}}^{\phi}) - \hat{\boldsymbol{x}}_0^{\theta}(\tilde{\boldsymbol{x}}^{\phi})\| + CL_s\|\tilde{\boldsymbol{x}}^{\phi} - \tilde{\boldsymbol{x}}^{\theta}\| \\ &= C\left(1 + L_s\sqrt{\alpha_s}\right)\|\boldsymbol{a}^{\phi} - \boldsymbol{a}^{\theta}\| + C\|\hat{\boldsymbol{x}}_0^{\phi}(\tilde{\boldsymbol{x}}^{\phi}) - \hat{\boldsymbol{x}}_0^{\theta}(\tilde{\boldsymbol{x}}^{\phi})\|. \end{aligned} \tag{24}$$

The first term is the clean-prediction gap at the current timestep t and the second is the gap at level s evaluated at an input of that noise level. Both are gaps of the form controlled by Eq. (18), so the two-stage score inherits the same control, with the factor $1 + L_s\sqrt{\alpha_s}$ accounting for the propagation of the first estimate into the re-corrupted input.

Ordinal labels of minority guidance. Minority guidance assigns to $\boldsymbol{x}_0$ the label k when $\mathcal{L}_{\theta}(\boldsymbol{x}_0; s)$ falls in the k-th bin $[\tau_{k-1}, \tau_k)$ of a fixed partition. By Eq. (21) the repaired model assigns the same label whenever the score gap is smaller than the distance of $\mathcal{L}_{\theta}(\boldsymbol{x}_0; s)$ to the nearest boundary,

$$C\sqrt{\mathcal{D}_s(\boldsymbol{x}_0; \phi)} < \min_k \left|\mathcal{L}_{\theta}(\boldsymbol{x}_0; s) - \tau_k\right|. \tag{25}$$

Bounding the clean-prediction gap therefore bounds the set of samples whose label can change, and only samples whose full-precision score lies close to a bin boundary are at risk.

Scope of the bound. Proposition 1 bounds the value of the score, not the direction of the guidance gradient, and it is loose when C is large. Its role is to show that the adapter objective controls the quantity the samplers consume, rather than to certify a numerical margin.

E Clean-Prediction Error under Boosted Initialization

Fig. 7 measures the clean-prediction error of the quantized and repaired DiT-XL/2 relative to full precision at $t = T$ as a function of the boost γ of BnS, on 256 fixed initial noises and class labels. For both bit widths the error grows rapidly with γ, confirming that boosted inputs fall outside the calibrated activation ranges and incur clipping error. The repair reduces this error throughout the range, by about 74% at $\gamma = 1.4$ and 25% at $\gamma = 2.4$ for W6A6, even though the adapter never sees boosted inputs during training. The relative reduction shrinks as γ grows, which is consistent with the adapter being trained at $\gamma = 1$, and matches the larger γ that the repaired rows tolerate in Table 10.

F Evaluation Protocol

Minority references. For ADM, LDM-4 and DiT, cFID, sFID, precision and recall are computed against a fixed reference set of real minority samples, following the protocol of MG, SGMS, and BnS (Um et al., 2024; Um & Ye, 2024; Um et al., 2025): the real images with the highest AvgkNN distance in ResNet-50 feature space with k=5. The references contain 10,000 CelebA images, 50,000 LSUN Bedrooms images and 128,161 ImageNet images (Um et al., 2024; Um & Ye, 2024; Um et al., 2025), where the ImageNet reference selects the top 10% of each class. The same reference set is used for every row of a given dataset, so the metrics of the full-precision, quantized and repaired models are directly comparable.

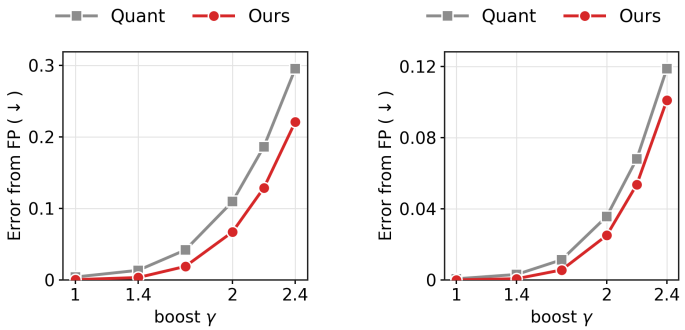

Figure 7: Clean-prediction error at $t = T$ relative to full precision as a function of the BnS boost γ on DiT-XL/2. Left: W6A6. Right: W8A8.

Generated samples. For all scenarios we evaluate with at least 10K generated samples. ImageNet samples are balanced across the 1,000 classes and ImageNet cFID is computed over the pooled samples, and MS-COCO uses one image per prompt. Within each setting, all compared rows share the initial noise, the ancestral noise and the conditioning, as described in Appendix G.4.

Text-to-image. MS-COCO has no reference minority set of this kind, and its metrics and the reading of LL are described in Appendix G.6.

G Implementation Details

This appendix provides the full configuration used in our experiments. Sec. G.1 describes the quantized baselines, Sec. G.2 the selective bit allocation, Sec. G.3 the adapter and its training, and Sec. G.4 the integration with each minority sampler. All experiments are run on a single NVIDIA RTX A6000.

G.1 Quantized Baselines

We use the official post-training quantization implementations of TFMQ (Huang et al., 2024) for ADM, DMQ (Lee et al., 2025) for LDM-4 and Stable Diffusion v1.4, and TQ-DiT (Hwang et al., 2025) for DiT-XL/2, with their released calibration procedures and hyperparameters. Each baseline keeps a small set of layers in full precision by design: TFMQ leaves the final output layer of ADM in full precision, and for LSUN Bedrooms additionally the activations of the input convolution and the time-embedding layers; DMQ and TQ-DiT follow their released scopes. TQ-DiT quantizes DiT-XL/2 (SD-VAE f=8, four-channel latents, linear schedule, 100 respaced ancestral DDPM steps) with activation quantizers that hold a separate table for each of ten timestep groups, calibrated on 6,400 full-precision 100-step trajectories. The resulting quantized models are treated as fixed deployed models: their weights, quantization scales and zero points are never updated by our repair. Table 6 summarizes the configurations.

Table 6: Quantized baselines. N_{w} and N_{act} denote the number of weight-quantized modules and activation quantizers in each model. N_{act} exceeds N_{w} because modules whose input is a concatenation of two branches (the skip connections of the UNet decoder) carry one quantizer per branch, and because the attention matmuls $\boldsymbol{QK}^{\top}$ and $\boldsymbol{AV}$ have no weights but quantize both operands where the baseline includes them (SD v1.4 and DiT).

Model	**Dataset**	**Quantizer**	**Bits**	N_{w}	N_{act}
ADM	CelebA 64	TFMQ	W6A6 / W4A8	174	190
ADM	LSUN Bedrooms 256	TFMQ	W6A6 / W4A8	181	199
LDM-4	ImageNet 256	DMQ	W4A6 / W4A8	257	269
DiT-XL/2	ImageNet 256	TQ-DiT	W6A6 / W8A8	114	225
SD v1.4	MS-COCO	DMQ	W4A6 / W4A8	257	392

G.2 Selective Bit Allocation

Sensitivity estimation. The saliency of Eq. (11) is computed on calibration images from the training split (Table 8; for LDM-4 the ranking uses two images from each of five classes). For each image we draw timesteps from the high-noise regime $\mathcal{T}_{\mathrm{hn}}$ of Eq. (7) with one forward corruption per timestep, evaluate d_{clean} against the full-precision denoiser, and accumulate $\langle \partial d_{\mathrm{clean}} / \partial Q_\ell(\boldsymbol{a}_\ell), \boldsymbol{e}_\ell \rangle$ for every activation quantizer in a single backward pass. Scores are averaged over images and timesteps.

Selection and grid refinement. Quantizers are restored greedily in ranked order and we stop at the knee of the cumulative error curve, beyond which the marginal reduction drops by an order of magnitude. The selected sites receive one additional activation bit (A6→A7, A8→A9). The refined grid keeps the calibrated range of the deployed quantizer and doubles the number of levels, i.e. $\Delta_\ell^{\mathrm{repair}} = \Delta_\ell / 2$ with the zero point rescaled accordingly, so no additional calibration or training is involved. For DiT, the refinement is applied to every timestep-group table of the selected site. The allocation is fixed after this step; Table 7 lists K for every configuration.

Table 7: Selective bit allocation. K is the number of activation quantizers whose bit width is increased by one.

Model	Bits	K	Refined act. bits
ADM CelebA	W6A6	4	7
ADM CelebA	W4A8	4	9
ADM LSUN	W6A6	4	7
ADM LSUN	W4A8	4	9
LDM-4	W4A6	6	7
LDM-4	W4A8	6	9
DiT-XL/2	W6A6	3	7
DiT-XL/2	W8A8	5	9
SD v1.4	W4A6	10	7
SD v1.4	W4A8	6	9

G.3 Adapter Architecture and Training

Adapted layers. Low-rank factors are attached to every weight-quantized module of the baseline, including all convolution and linear layers, the attention query/key/value and output projections, embedding projections and skip connections (174 modules for ADM on CelebA and 114 for DiT-XL/2). Layers that the baseline keeps in full precision receive no adapter. For a module with weight $\boldsymbol{W}_\ell \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$, where convolution kernels are flattened along the input dimension, we use rank $r = 16$ with $\boldsymbol{B}_\ell \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ and $\boldsymbol{A}_\ell \in \mathbb{R}^{r \times d_{\mathrm{in}}}$. Following the standard LoRA parameterization the update is scaled by α / r with $\alpha = 16$, so the scale is one. $\boldsymbol{A}_\ell$ is initialized with Kaiming-uniform and $\boldsymbol{B}_\ell$ with zeros, so that Eq. (12) returns the deployed weights at initialization.

Forward pass during training. Each training forward evaluates the repaired denoiser exactly as it is deployed: adapted layers use $Q_\ell^{\mathrm{w}}(Q_\ell^{\mathrm{w}}(\boldsymbol{W}_\ell) + \boldsymbol{B}_\ell \boldsymbol{A}_\ell)$, i.e. round-to-nearest of the sum onto the frozen integer grid, the selected activation sites use the refined grids of Sec. G.2, and all other quantizers are those of the baseline. Gradients pass through the rounding with the straight-through estimator, and only the adapter parameters ϕ, i.e. the low-rank factors $\{\boldsymbol{A}_\ell, \boldsymbol{B}_\ell\}_\ell$ and the activation affine parameters described next, receive gradients.

Learned activation affine. In addition to the low-rank factors, the adapter ϕ includes a per-channel affine map $\boldsymbol{a}_\ell \mapsto s_t \, (\boldsymbol{\gamma}_\ell \odot \boldsymbol{a}_\ell + \boldsymbol{\beta}_\ell)$ applied to the input of every activation quantizer, where s_t is a per-timestep scalar (one per timestep for ADM, LDM and SD, one per timestep group for DiT). All entries are initialized to the identity ($\boldsymbol{\gamma}_\ell = \mathbf{1}$, $\boldsymbol{\beta}_\ell = \mathbf{0}$, $s_t = 1$) and trained jointly with the low-rank factors. Since the map precedes a uniform quantizer, it is equivalent to a learned per-channel scale and offset of that quantizer, and it can be folded into the quantizer's scale and zero point at deployment; the bit width and the number of levels are unchanged. These parameters amount to two vectors per activation quantizer plus T scalars and are negligible next to the low-rank factors (Table 11). Base weights and weight quantizers remain frozen throughout.

Data and corruption. The minority reference sets used for evaluation follow the protocol of MG and SGMS (Um et al., 2024; Um & Ye, 2024), i.e. the real images with the highest AvgkNN distance in ResNet-50 feature space (k=5), as described in Appendix F. Calibration images are sampled uniformly at random from the training split of each dataset, independently of any minority statistic: 10,000 images with horizontal flips for ADM on CelebA, 10,000 images for ADM on LSUN Bedrooms, 256 images for LDM-4 and DiT-XL/2 on ImageNet (two per class over 128 classes), and 256 image–caption pairs from the MS-COCO training split at 512×512 for Stable Diffusion; no validation prompts are used. For latent models the images are encoded once with the frozen VAE encoder and training operates on latents. Timesteps are drawn uniformly from the full range as in Eq. (13), with one timestep per micro-batch since the activation tables of the baselines are timestep-dependent, and a fresh forward corruption $\boldsymbol{x}_t \sim q_{\alpha_t}(\cdot \mid \boldsymbol{x}_0)$ is sampled at every iteration. Conditional models use the paired class label or caption, and the target is the conditional prediction without classifier-free guidance.

Objective and optimization. We minimize $\mathcal{L}_{\text{adapt}}$ of Eq. (13) with $w_{\max} = 10$; in the implementation the loss is evaluated equivalently in clean-prediction space as $\min\big(1, w_{\max}\alpha_t/(1-\alpha_t)\big)\|\hat{\boldsymbol{x}}_0^{\hat{\theta},\phi} - \hat{\boldsymbol{x}}_0^{\theta}\|_2^2$ (cf. Eq. (17)). The target $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is computed by the full-precision model with gradients detached. We train for 5,000 iterations with an effective batch size of 16 (micro-batch $\times$ accumulation in Table 8) using AdamW without weight decay, learning rate 5×10^{-4} and cosine annealing to 5×10^{-5}. For DiT the affine parameters use a learning rate of 1×10^{-4} and gradients are clipped at norm 1. Stable Diffusion is trained with mixed precision. For ADM and LDM-4 we take the adapter at the final iteration. For DiT and Stable Diffusion we keep the checkpoint with the smallest noise-prediction gap to the full-precision model on 16 fixed calibration images with fixed noise (evaluated at t=80 of the 100-step schedule for DiT), which selects iteration 750 for SD W4A8 and 3,250 for SD W4A6. For DiT this gap decreases from 0.0188 to 0.0084 at W8A8 and from 0.0643 to 0.0221 at W6A6. Every sampler verifies at load time that the fused adapter reproduces this gap, which guards against mismatches between the training-time and deployed quantization.

Table 8: Adapter training configuration. N_{cal} is the number of calibration images, and Batch is micro-batch $\times$ gradient accumulation. All configurations use rank r=16, $w_{\max}$=10, 5,000 iterations and cosine annealing from 5×10^{-4} to 5×10^{-5}.

Model	Bits	Calibration data	N_{cal}	Batch
ADM CelebA	W6A6 / W4A8	CelebA train + flip	10,000	8×2
ADM LSUN	W6A6 / W4A8	LSUN Bedrooms train	10,000	2×8
LDM-4	W4A6 / W4A8	ImageNet train, 2/class, 128 classes	256	8×2
DiT-XL/2	W6A6 / W8A8	ImageNet train, 2/class, 128 classes	256	8×2
SD v1.4	W4A6 / W4A8	COCO train, 512^2, paired captions	256	2×8

G.4 Integration with Minority Samplers

All compared rows share the number of denoising steps, the CFG scale, the initial noise and the ancestral noise. The minority-specific strength (w_{mg} for SGMS, w for MG and γ for BnS) is the one quantity selected per row, by the sweep described for each sampler below, and the same sweep is applied to the full-precision row. Every row is thus compared at its own best operating point, and the only other difference between full-precision, quantized and repaired models is the denoiser (or, for MG, the denoiser that produced the labels).

SGMS. At each step the reverse update is computed from the noise prediction of the stock quantized model, with the repair disabled and without gradients. For LDM-4 this is the classifier-free-guided prediction from the batched unconditional and conditional forward. The minority score is then computed in a separate branch. A gradient-enabled conditional forward of the repaired denoiser at $(\boldsymbol{x}_t, t)$ produces $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$, which is re-corrupted to the scoring level s ($0.8T$ for cosine and $0.5T$ for linear schedules, as in SGMS) and reconstructed by a second repaired forward without gradients. The discrepancy $J = d(\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{x}}_0')$ is back-propagated to $\boldsymbol{x}_t$ and the resulting gradient $\hat{\boldsymbol{g}}$ is added to the reverse update as $\boldsymbol{x}_{t-1} \leftarrow \boldsymbol{x}_{t-1} + w_{\text{mg}}\tilde{\beta}_t\hat{\boldsymbol{g}}$. This recomputation of $\boldsymbol{\epsilon}(\boldsymbol{x}_t, t)$ is inherent to SGMS, whose reference implementation also re-evaluates the denoiser with gradients enabled for the score. Since the score uses the conditional rather than the guided prediction, the two forwards cannot be

shared regardless of the repair, and the repair therefore adds no forward pass. The repaired weights and refined activation grids are used only inside this branch, and in the fused implementation the generation forward and the score branch read separate weight buffers. The discrepancy d and the scoring interval follow the SGMS reference configuration for each model. For ADM on CelebA and LSUN, d is LPIPS in pixel space and the score is evaluated every 5 of the 250 DDPM steps, with the reverse update in between using the stock quantized model alone. For LDM-4, d is the squared ℓ_2 distance in latent space and the score is evaluated at every one of the 20 DDIM steps. The strength w_{mg} is selected per row by a sweep, and the selected values and the sampler settings are listed in Table 10. The repaired rows at the lowest bit widths select a larger w_{mg} than the stock quantized rows. This is expected, since the repair removes the quantization noise that inflates the reconstruction discrepancy of the stock model (Fig. 2), so the same guidance effect requires a larger scale on the cleaner score, and all rows are compared at their own best setting under the matched sweep of Fig. 8.

MG. MG builds its guidance from minority scores of real training images, so the repair enters MG the same way it enters SGMS, wherever a minority score is estimated. (i) *Scoring.* We use 128,161 ImageNet training images sampled uniformly across the 1,000 classes. Each image is encoded, corrupted to t=60 of the 100-step schedule ($0.6T$, $\alpha_t \approx 0.024$) and reconstructed with a single conditional Tweedie step without classifier-free guidance, and the score is the LPIPS-VGG distance between the decoded original and the decoded reconstruction. All rows share the same corruption noise. Scores are binned per class into 25 equal-mass quantile bins to form ordinal levels. (ii) *Classifier.* We adapt the ADM encoder-UNet to $4\times32\times32$ noisy latents, add the class embedding to the timestep embedding and train the resulting 30.7M-parameter classifier $p(l \mid \boldsymbol{z}_t, t, c)$ with $t \sim \mathcal{U}\{0, \ldots, 999\}$ using AdamW (3×10^{-4}, linear decay, weight decay 0.05), batch size 256 and 20k iterations, holding out 5% of the images with the same split for every row. (iii) *Guided sampling.* Following the ADM conditional-mean formulation, the reverse mean is shifted by $w\,\boldsymbol{\Sigma}_\theta \nabla_{\boldsymbol{z}_t} \log p(l{=}24 \mid \boldsymbol{z}_t, t, c)$ at every one of the 100 steps with CFG scale 1.5, and the guidance gradient passes through the classifier. The guidance scale w is selected per row from $\{1, 2, 3, 3.5, 5, 7.5\}$ on 300 samples as the smallest cFID subject to precision ≥ 0.85, giving w=3.5 (FP), 3.5 (W8A8), 5 (W8A8+Ours), 3 (W6A6) and 3 (W6A6+Ours), and the selected setting is then sampled at 10k images (10 per class). Each row runs all three stages with its own denoiser.

Table 9 reports how faithfully each denoiser reproduces the full-precision labels, which is the quantity that Eq. (14) controls. The first row re-runs the full-precision model with independent corruption noise and thus gives the agreement attainable from a single draw. Quantization at W6A6 reduces the rank correlation with the full-precision score to 0.48, well below this noise floor, and the repair restores it to 0.95; at W8A8 the repair raises it from 0.94 to 0.99. The classifiers inherit this gap: on the held-out split at t=100, the rank correlation of the predicted label with the full-precision label is 0.83 (FP), 0.81 (W8A8), 0.39 (W6A6), 0.83 (W8A8+Ours) and 0.82 (W6A6+Ours), and the ratio of the predicted probability of the top bin to the rest rises from $1.4\times$ at W6A6 to $7.4\times$ with the repair.

Table 9: Fidelity of MG minority labels on DiT-XL/2 with respect to the full-precision labels. FP (re-draw) is the full-precision model scored with independent corruption noise.

Labeling model	**Spearman ρ vs. FP**	**Top-bin retention**
FP (re-draw)	0.788	0.333
W8A8	0.940	0.659
W8A8 + Ours	0.992	0.872
W6A6	0.484	0.185
W6A6 + Ours	0.950	0.731

BnS. BnS involves no training. The initial state is drawn as $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \gamma^2 \boldsymbol{I})$ and the sampler otherwise runs unchanged. We do not use the timestep skipping of Um et al. (2025), for reasons given below, so that the trajectory is identical to the standard sampler except for its initial state and the comparison isolates the effect of the denoiser. Since BnS has no score branch, the repaired row attaches the adapter and refined grids for the entire trajectory, updating the timestep-group index of the activation tables at every step. The DiT adapter is the one trained for MG, and boosted initial states ($\gamma > 1$) are never seen during training. The FP anchor with γ=1 is standard sampling. For each bit width we sweep γ with 2k index-paired samples (batch 16, fixed seed) shared across rows,

select the γ with the best cFID for each row, and resample the selected settings with 10k images. The selected values on DiT are γ=5.5 (FP), 2.2 (W8A8), 2.4 (W8A8+Ours), 1.4 (W6A6) and 2.0 (W6A6+Ours). The quantized models tolerate far less boosting than full precision before quality collapses, and the repair extends the usable range in both bit configurations.

G.5 Boosting under Stochastic and Deterministic Sampling

BnS was designed for stochastic samplers, and Um et al. (2025) show that it fails with the probability-flow ODE (their Fig. 2e and Sec. A.3). We find that boosting does transfer to DDIM, but with a much smaller γ, and both regimes follow from their analysis. Under the reverse SDE, Proposition A.2 of Um et al. (2025) shows that the total-variation distance between a boosted and a standard trajectory contracts,

$$\mathrm{TV}\big(\boldsymbol{x}^{\mathrm{SDE}}_{t_{\min}}, \boldsymbol{y}^{\mathrm{SDE}}_{t_{\min}}\big) \le \rho\, \mathrm{TV}(\boldsymbol{x}_{t_{\max}}, \boldsymbol{y}_{t_{\max}}), \qquad \rho < 1, \tag{26}$$

and in the Gaussian case (their Proposition 3.2) the boost reaches the output only through the factor $\alpha(T_{\mathrm{skip}})^2$,

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{SDE}}_0 = \boldsymbol{\Sigma}_0 + \alpha(T_{\mathrm{skip}})^2\, \boldsymbol{\Sigma}_0^2 \boldsymbol{\Sigma}^{-2}_{T_{\mathrm{skip}}} \left(\gamma^2 \boldsymbol{I} - \boldsymbol{\Sigma}_{T_{\mathrm{skip}}}\right). \tag{27}$$

Both results say the same thing. A stochastic sampler erases most of the initial perturbation, so a large γ is both tolerated and required. Eq. (27) is the idealized limit for the optimal score, in which the boost would vanish entirely without timestep skipping once $\alpha(T) \approx 0$, and Um et al. (2025) add skipping for this reason. The learned denoiser is not the optimal score at boosted inputs it never saw during training, and Um et al. (2025) show that the boosted trajectory departs from the standard one over the first steps before the contraction takes hold (their Fig. 3) and that boosting alone already shifts the samples (their Table 2a). On DiT this residual effect is sufficient at the γ values we select, between 1.4 and 5.5, so we omit skipping and keep the trajectory otherwise identical to the standard sampler.

Under the probability-flow ODE the same propositions give the opposite picture. The total-variation distance is preserved exactly, $\mathrm{TV}(\boldsymbol{x}^{\mathrm{ODE}}_{t_{\min}}, \boldsymbol{y}^{\mathrm{ODE}}_{t_{\min}}) = \mathrm{TV}(\boldsymbol{x}_{t_{\max}}, \boldsymbol{y}_{t_{\max}})$, and the Gaussian covariance (their Proposition A.1) becomes

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{ODE}}_0 = \boldsymbol{\Sigma}_0\, \boldsymbol{\Sigma}_T^{-1}\, \gamma^2 \boldsymbol{I} \;\approx\; \gamma^2 \boldsymbol{\Sigma}_0 \quad \text{for } \alpha(T) \approx 0, \tag{28}$$

with no $\alpha(T)$ attenuation at all. The boost is transmitted to the output one-to-one. This is exactly why a large γ leaves the data manifold under the ODE, and it is also why a small γ is enough. On Stable Diffusion with DDIM (η=0) the selected values lie in $[1.02, 1.06]$, a 4 to 12% increase in output variance by Eq. (28), which already shifts generation toward low-density regions (lower LL in Table 2) without the off-manifold failure that Um et al. (2025) observe at γ=1.1. Skipping is unnecessary here as well, since there is no $\alpha(T)$ factor to keep away from zero. In both regimes the repaired rows select a larger γ than the stock quantized rows. We attribute this to the boosted state being an input outside the quantization calibration range, so that a denoiser whose high-noise clean prediction is closer to full precision stays on the manifold under a stronger boost.

Table 10: Sampler configurations. Steps gives the sampler and number of denoising steps, CFG the classifier-free guidance scale (– for unconditional models), and s the scoring level. Per-row values are listed in the order FP / Quant / Quant+Ours for each bit configuration. For SGMS, d is the score discrepancy and "every k" the interval at which the score is evaluated.

Model	Sampler	Steps	CFG	s	Minority hyperparameters
ADM CelebA	SGMS	DDPM 250	–	$0.8T$	d=LPIPS, every 5; w_{mg}: 0.4 / 0.4 / 1.6 (W6A6), 0.4 / 0.4 / 0.5 (W4A8)
ADM LSUN	SGMS	DDPM 250	–	$0.5T$	d=LPIPS, every 5; w_{mg}: 0.3 / 0.2 / 0.2 (W6A6), 0.3 / 0.2 / 0.2 (W4A8)
LDM-4	SGMS	DDIM 20, η=0	3	$0.5T$	latent ℓ_2^2, every 1; w_{mg}: 3 / 3 / 30 (W4A6), 3 / 2 / 3 (W4A8)
DiT-XL/2	MG	DDPM 100	1.5	$0.6T$	25 bins, LPIPS-VGG, l=24; w: 3.5 / 3.5 / 5 (W8A8), 3.5 / 3 / 3 (W6A6)
DiT-XL/2	BnS	DDPM 100	1.5	n/a	γ: 5.5 / 2.2 / 2.4 (W8A8), 5.5 / 1.4 / 2.0 (W6A6)
SD v1.4	BnS	DDIM 50, η=0	7	n/a	γ: 1.06 / 1.02 / 1.04 (W4A6), 1.06 / 1.03 / 1.05 (W4A8)

G.6 Evaluation Protocol on MS-COCO

Text-to-image minority generation has no reference minority set of the kind used for ADM, LDM-4 and DiT, so we adopt the protocol of Um & Ye (2025). Rarity is measured by the log-likelihood of

each generated image under a fixed full-precision Stable Diffusion v1.4, computed exactly with the probability-flow ODE of Song et al. (2021) and reported in bits per dimension. Lower LL indicates a rarer sample. Quality and alignment are measured by CLIPScore, PickScore and ImageReward, and distributional fidelity and diversity by precision, recall, density and coverage against all 40,504 val2014 images with k=5 and 10k generated samples, which is why the absolute coverage values are small. Following Um & Ye (2025) we do not report FID, since it measures closeness to the full real distribution and therefore penalizes exactly the low-density generation that a minority sampler is meant to produce.

Two points matter when reading LL. First, LL is evaluated under the full-precision model for every row, so quantized and repaired samples are scored by the same likelihood regardless of which denoiser produced them. Second, artifacts also lower LL, so a stock quantized model can reach a lower LL than full precision (Table 2) without producing rarer content. LL is therefore read together with the preference scores, and the sweeps of Fig. 8 compare rows at matched LL, where a higher PickScore at the same LL means rarer samples of better quality rather than a worse model that happens to be less likely. Under this reading the repaired model dominates the stock quantized model at both bit widths and approaches full precision at W4A8.

Table 11: Storage overhead of the repair. Base is the quantized model size, Factors the low-rank factors stored in FP16, Affine the activation affine parameters, Params the number of trainable parameters, and Overhead the ratio (Factors+Affine)/Base. The fused form instead duplicates Base. Sizes in MB (10^6 bytes), computed from the adapter checkpoints.

Model	Bits	Base	Factors	Affine	Params	Overhead
ADM CelebA	W6A6	221.2	18.9	0.55	9.72M	8.8%
ADM CelebA	W4A8	147.4	18.9	0.55	9.72M	13.2%
ADM LSUN	W6A6	414.4	26.3	0.72	13.53M	6.5%
ADM LSUN	W4A8	276.2	26.3	0.72	13.53M	9.8%
LDM-4	W4A6 / W4A8	195.1	22.6	0.82	11.72M	12.0%
DiT-XL/2	W6A6	334.5	16.6	0.90	8.75M	5.2%
DiT-XL/2	W8A8	446.0	16.6	0.90	8.75M	3.9%
SD v1.4	W4A6 / W4A8	415.6	32.1	1.11	16.60M	8.0%

H Effect of Bit Width and Gains over Full Precision

Two patterns in Tables 1, 2 and 12 deserve comment. The degradation of minority generation is milder at 8-bit activations than at 6-bit, and the repaired model, and in one case the stock quantized model, reaches a lower cFID than the full-precision sampler. Both follow from the analysis of Sec. 4.

H.1 Why 8-bit activations degrade less

Eq. (6) factorizes the clean-prediction error into the amplification $\sqrt{(1-\alpha_t)/\alpha_t}$ and the denoiser error $\boldsymbol{\delta}$. The amplification depends only on the noise schedule and the scoring level, so it is identical for every bit configuration of a given model. What changes with the bit width is $\boldsymbol{\delta}$. Writing $\boldsymbol{\delta} = \boldsymbol{\delta}_{\mathrm{w}} + \boldsymbol{\delta}_{\mathrm{a}}$ for the parts due to weight and activation quantization and expanding the latter to first order in the activation residuals $\boldsymbol{e}_\ell = Q_\ell(\boldsymbol{a}_\ell) - \boldsymbol{a}_\ell$, as in Eq. (11),

$$\mathcal{D}_{\mathrm{clean}} = \frac{1-\alpha_t}{\alpha_t}\,\mathbb{E}\big[\|\boldsymbol{\delta}_{\mathrm{w}} + \boldsymbol{\delta}_{\mathrm{a}}\|_2^2\big], \qquad \boldsymbol{\delta}_{\mathrm{a}} \approx \sum_\ell \boldsymbol{J}_\ell\, \boldsymbol{e}_\ell, \qquad \boldsymbol{J}_\ell = \frac{\partial \boldsymbol{\epsilon}_{\hat{\theta}}}{\partial Q_\ell(\boldsymbol{a}_\ell)}. \tag{29}$$

For the uniform quantizer of Eq. (4) with a calibrated range R_ℓ, the step is $\Delta_\ell = R_\ell/(2^b-1)$ and the rounding residual is uniform on $[-\Delta_\ell/2, \Delta_\ell/2]$ inside the range, so

$$\mathbb{E}\big[\|\boldsymbol{\delta}_{\mathrm{a}}\|_2^2\big] \approx \sum_\ell \frac{\Delta_\ell^2}{12}\,\mathrm{tr}\big(\boldsymbol{J}_\ell \boldsymbol{J}_\ell^\top\big) \;\propto\; 2^{-2b}, \qquad \frac{\Delta_\ell(8)}{\Delta_\ell(6)} = \frac{63}{255} \approx \frac{1}{4}. \tag{30}$$

Raising the activations from 6 to 8 bits therefore divides the activation-induced clean-prediction error by about sixteen, while the amplification in front of Eq. (29), the clipping error and $\boldsymbol{\delta}_{\mathrm{w}}$ are

unchanged. Through Eq. (14) the score gap shrinks by the square root of this factor. This is what the tables show: the cFID gap to full precision falls from 13.5 to 5.3 on LSUN Bedrooms and from 4.8 to 0.5 on DiT with MG, and on CelebA the W4A8 stock model already matches full precision. On LDM-4, where both configurations keep 4-bit weights, the two stock rows are close in cFID (58.94 vs. 58.27), indicating that $\boldsymbol{\delta}$ is dominated there by the term $\boldsymbol{\delta}_{\mathrm{w}}$ of Eq. (29), which extra activation bits do not remove. For the same reason the repair has the most to recover at the lower activation bit width, consistent with Table 3, where its margin is largest at W6A6, and with Fig. 8, where the three curves nearly coincide at W4A8.

H.2 Why the repaired model can exceed full precision

In several rows the repaired model reaches a lower cFID than the full-precision sampler (ADM CelebA W6A6, LDM-4 at both bit widths, DiT with BnS at W8A8), as does the stock model on CelebA at W4A8. This should be read in light of what the metric measures. FID (Heusel et al., 2017) is the Fréchet distance between Gaussian fits to the feature distributions of generated and reference images, and cFID (Parmar et al., 2022) is its variant with standardized image resizing. Being a distance between distributions rather than a per-sample quality score, it is lowered not only by sharper samples but also by samples that spread more evenly over the reference; a mild increase in diversity can therefore reduce FID even when per-sample fidelity, as measured by precision, decreases. Small favorable FID differences of quantized over full-precision models are indeed reported under standard sampling in prior PTQ work (Li et al., 2023; Huang et al., 2024), where the quantization error acts as a slight stochastic perturbation of the denoiser that broadens the generated distribution without visibly harming individual samples.

The same effect operates here, in a setting that rewards it more. Our reference set consists of real minority samples, so the metric rewards spreading mass over low-density regions, which is precisely what added variance promotes (Um et al., 2025). After repair the residual error of Eq. (29) is small but nonzero and no longer distorts the score (CCC 0.82 in Fig. 3); at this magnitude it acts as a mild input-dependent perturbation of the guidance direction rather than a systematic bias, and it surfaces as gains in recall and coverage rather than precision (LDM-4 W4A6 recall 0.195 vs. 0.181 with precision 0.955 vs. 0.966; DiT BnS W8A8 recall 0.411 vs. 0.363 for the stock model). The stock CelebA W4A8 row, whose $\boldsymbol{\delta}$ is already small by Eq. (30), is the same effect without the repair. A second contributing factor is that every row is compared at its own best minority strength under a matched sweep (Appendix G.4), and a denoiser whose clean predictions differ slightly from full precision can tolerate a different, often larger, strength before fidelity collapses: the repaired CelebA W6A6 row selects $w_{\mathrm{mg}} = 1.6$ against 0.4 for full precision, and its lower cFID comes with recall matching full precision and lower precision, i.e. from reaching further into the reference rather than from higher fidelity. Fig. 8 makes this explicit: at matched precision the repaired curve attains higher recall than full precision over most of the sweep, so the sign of a single-point difference reflects the operating point. These margins are small relative to the degradation they reverse, and we do not read them as the low-bit model being a better minority sampler; the claim of this paper is restoration, and the excess is a property of a distribution-level metric evaluated against a minority reference.

I BnS on DiT

Table 12 reports BnS on DiT-XL/2 under the protocol of Appendix G.4. The adapter is the one trained for MG and is attached for the whole trajectory without any additional training. The picture is the same as on Stable Diffusion. Quantization raises cFID by about four points at both bit widths and, at W8A8, drops recall from 0.445 to 0.363. The repair recovers most of this at W6A6 and all of it at W8A8, where cFID (26.65) and sFID (8.56) end up below the full-precision sampler. Precision moves slightly in the other direction, which is consistent with the larger boost γ that the repaired rows select (Table 10). A quantized model that produces cleaner high-noise predictions tolerates more boosting, and BnS converts that headroom into diversity.

Table 12: BnS on DiT-XL/2 (ImageNet 256×256) with 10k samples per row. Each row uses the γ selected for it in Table 10.

Model	Bits	Method	cFID↓	sFID↓	Prec.↑	Rec.↑
DiT-XL/2 ImageNet 256×256	FP	Full Precision + BnS	27.51	9.93	0.834	0.445
	W6A6	TQ-DiT + BnS	31.74	9.57	**0.870**	0.406
		TQ-DiT + BnS + Ours	**29.11**	**8.91**	0.864	**0.408**
	W8A8	TQ-DiT + BnS	31.06	10.51	**0.860**	0.363
		TQ-DiT + BnS + Ours	**26.65**	**8.56**	0.854	**0.411**

Table 13: Component ablation on LDM ImageNet with DMQ and SGMS.

Bits	Grid	Adapter	cFID↓	sFID↓	Prec.↑	Rec.↑
W4A6	×	×	60.47	14.84	0.932	0.161
	✓	×	58.41	12.12	0.957	0.157
	×	✓	57.68	11.63	0.951	0.166
	✓	✓	55.14	10.57	0.956	0.172
W4A8	×	×	61.75	13.29	0.948	0.144
	✓	×	60.91	13.86	0.961	0.159
	×	✓	59.75	13.19	0.955	0.163
	✓	✓	57.17	12.78	0.962	0.171

J Component Ablation on LDM

Table 13 repeats the component ablation of Table 3 on LDM-4 ImageNet with DMQ and SGMS, using 3K samples per setting. Each component improves cFID over the stock quantized model, and their combination gives the best cFID, sFID and recall at both bit widths.

K TRADE-OFF CURVES

Fig. 8 compares full-precision, quantized and repaired models along a sweep of the minority-specific strength, so that each row is read at its own operating points rather than at a single hand-picked scale. The left two panels sweep the SGMS guidance scale w_{mg} on ADM CelebA and plot precision against recall, and the right two sweep the BnS boost γ on Stable Diffusion and plot PickScore against the log-likelihood of the samples. In every panel the repaired curve lies above the stock quantized one, i.e. it reaches higher recall at the same precision, or a higher PickScore at the same rarity. On CelebA at W6A6 the repair lifts recall by about 0.05 over the baseline across the whole sweep and exceeds the full-precision curve at matched precision, meeting it only at the highest precision; at W4A8 the three curves nearly coincide, and the repair keeps a small margin over the baseline. On COCO at W4A6 the baseline loses roughly 0.2 PickScore at every likelihood level and the repair recovers about half of it, while at W4A8 the repaired curve lies on top of the full-precision one and the baseline falls below both at the high-likelihood end. The ordering is the one Sec. 4 predicts. The lower the activation precision, the larger the amplified clean-prediction error, the further the stock quantized curve falls, and the more the repair has to recover.

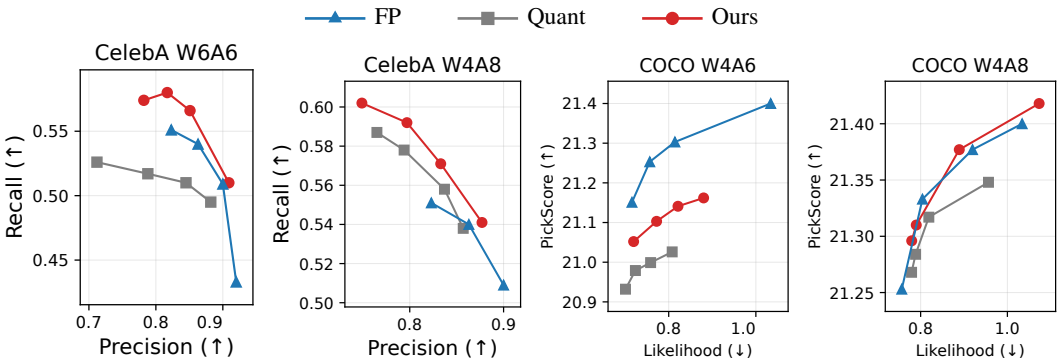

Figure 8: Trade-off curves for ADM on CelebA (precision–recall under SGMS at W6A6 and W4A8, left) and Stable Diffusion on MS-COCO (PickScore–log-likelihood under BnS at W4A6 and W4A8, right).

L ADDITIONAL QUALITATIVE RESULTS

This appendix provides additional qualitative comparisons complementing Fig. 5. Each figure follows the same layout: full-precision, quantized and repaired models for two bit configurations, with the configuration indicated above each column group. ADM and LDM results use SGMS, DiT results use MG and BnS, and Stable Diffusion results use BnS.

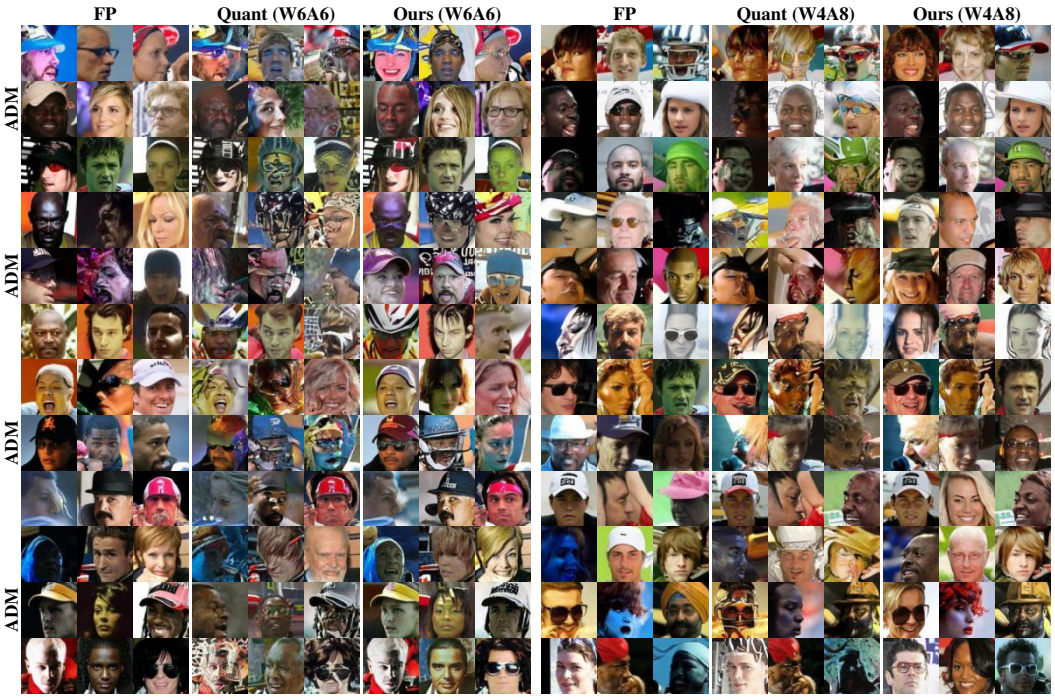

Figure 9: Additional minority samples from ADM on CelebA (64×64) under SGMS. Left: TFMQ W6A6. Right: TFMQ W4A8.

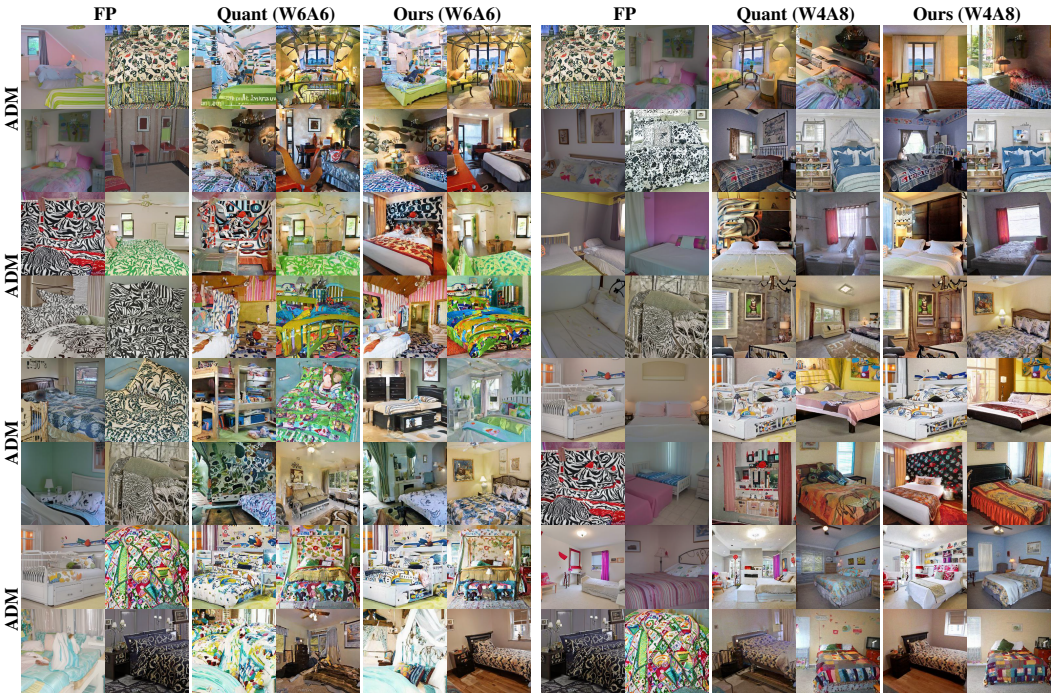

Figure 10: Additional minority samples from ADM on LSUN Bedrooms (256×256) under SGMS. Left: TFMQ W6A6. Right: TFMQ W4A8.

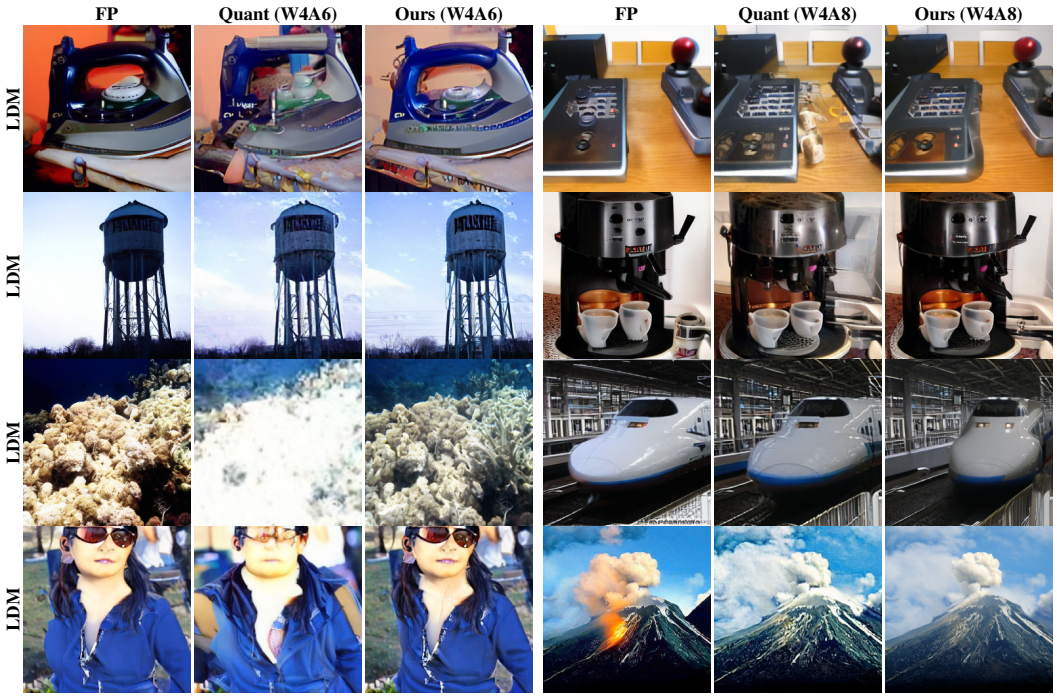

Figure 11: Additional minority samples from LDM-4 on ImageNet (256 × 256) under SGMS. Left: DMQ W4A6. Right: DMQ W4A8.



Figure 12: Additional minority samples from DiT-XL/2 on ImageNet (256 × 256) under MG. Left: TQ-DiT W6A6. Right: TQ-DiT W8A8.

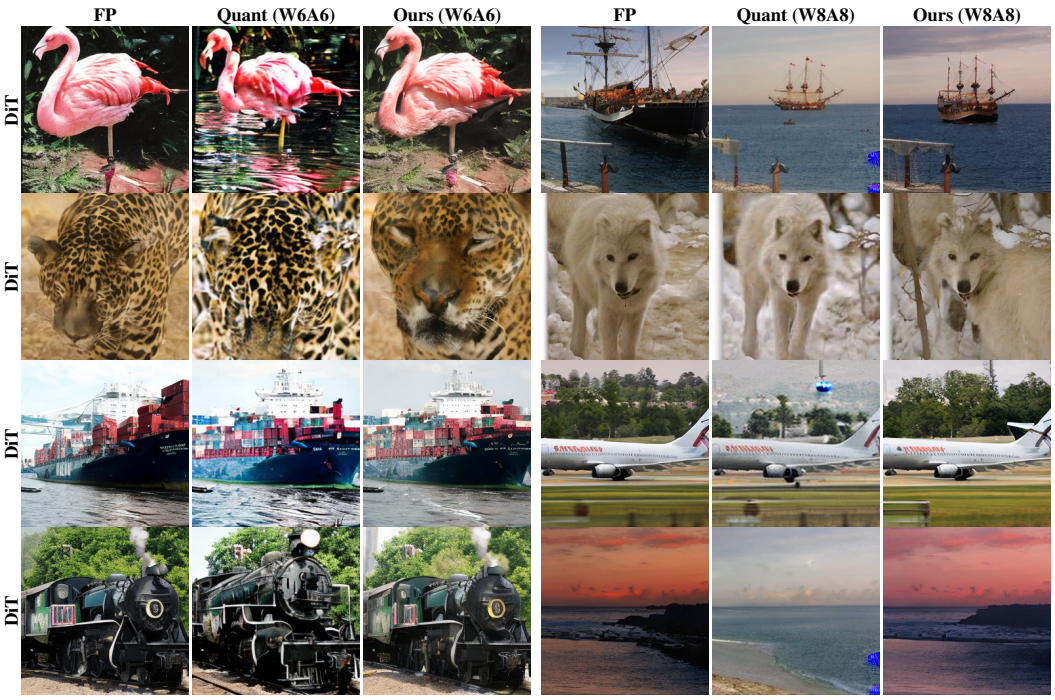

Figure 13: Additional minority samples from DiT-XL/2 on ImageNet (256×256) under BnS. Left: TQ-DiT W6A6. Right: TQ-DiT W8A8.

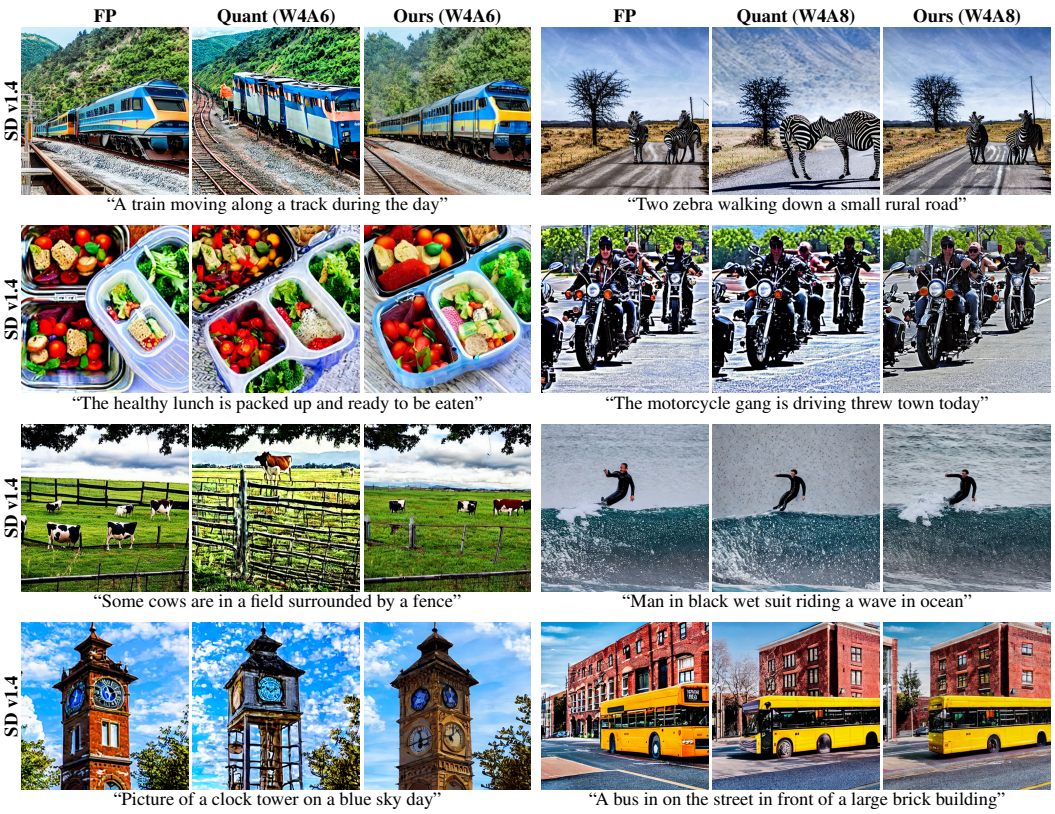

Figure 14: Additional minority samples from Stable Diffusion v1.4 on MS-COCO under BnS. Left: DMQ W4A6. Right: DMQ W4A8. The prompt is given below each triplet, and γ is the value selected for each row in Table 10.