Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

Evaluation Pitfalls in Post-Training Quantization of Diffusion Transformers: An Empirical Audit

Zeming Jin *

School of Management Science and Engineering, Beijing Information Science and Technology University, Beijing, 100018, China
*** Corresponding author Email:** jin2004wei@163.com

Abstract: Diffusion transformers are moving toward practical deployment, and post-training quantization (PTQ) has emerged as a standard, retraining-free compression route. The conventions used to evaluate PTQ for diffusion transformers—the Fréchet Inception Distance (FID) of images produced by a fixed sampling pipeline, with a cheap calibration-error statistic as a configuration-search proxy—are typically inherited from prior codebases rather than audited. Auditing this pipeline through a DiT-XL/2 (ImageNet 256×256) case study of 209 controlled experimental records, this paper identifies three systematic pitfalls. First, sampling-pipeline configuration: two latent defects in the legacy pipeline manufactured spurious method-level differences, and controlled A/B reproduction (deviation below 1%) shows that correcting them reverses earlier conclusions on channel equalization. Second, proxy-metric failure: 32 records sharing a bit-identical calibration error span an FID range of 98.74, and its rank correlation with FID is statistically indistinguishable from zero in every experiment family tested. Third, calibration-protocol sensitivity: under the corrected protocol, an error-aware dynamic-programming grouping strategy is statistically indistinguishable from a fixed-interval baseline (273.62±1.78 vs. 274.32±1.80), while group-wise equalization turns from apparently beneficial to harmful (273.62→313.23). The only method-level conclusion that survives the protocol correction—under three seeds, a single model, and the 8-bit weight-and-activation (W8A8) configuration studied—is that grouped quantization outperforms no grouping (388.96→274.32). This work closes with a corrected evaluation-protocol checklist.

Keywords: Post-training Quantization; Diffusion Transformer; Evaluation Audit; Fréchet Inception Distance; Calibration.

1. Introduction

Diffusion Transformers (DiTs) have established transformer backbones as a competitive architecture for high-fidelity image generation, most notably on class-conditional ImageNet synthesis [1]. Their iterative denoising makes inference expensive, and post-training quantization (PTQ) has emerged as a practical compression route, converting a pretrained full-precision model into a low-bit representation with only a small calibration set and no retraining. A growing line of work adapts PTQ to DiTs, including salience-aware calibration, automatic granularity allocation, timestep-aware designs, and hybrid-precision schemes [2]–[7], building on equalization techniques originally developed for large language models [8] and on earlier PTQ methods for U-Net diffusion models [9].

This literature is held together by a shared evaluation convention. Generation quality is reported almost exclusively through FID on a fixed number of generated samples; the sampling pipeline that produces them—sampler type and step count, classifier-free guidance (CFG) scale and its channel-wise application, latent-space clipping, VAE decoder variant, image resizing, and reference statistics—is treated as fixed background infrastructure. Because full FID evaluation is expensive, configuration search is frequently delegated to cheaper proxies, such as the quantization calibration error, under the implicit assumption that the proxy preserves the ranking FID would induce. Screening runs on as few as 1K images are likewise common, although FID values at different sample sizes are not directly comparable and screening-scale rankings need not transfer. These conventions are typically inherited from prior codebases rather than independently validated.

Our own research program followed exactly this convention—and this paper originates from the moment it broke down. In an earlier study on grouped weight-and-activation quantization with channel equalization for DiT-XL/2, we had reported that equalization improved FID in all 21 of 21 tested configurations, that a fine-grained error-aware grouping strategy outperformed fixed-interval grouping, and that a particular group count was optimal. While preparing a revision, we discovered two defects in our legacy sampling pipeline: the sampler's clip_sample option clipped intermediate latents, and classifier-free guidance was applied to all four latent channels rather than only the three noise channels. A controlled audit showed that these defects, not the quantization method, had manufactured several of our central conclusions. Rather than quietly correcting the record, we follow the precedent of evaluation audits in generative modeling [10]–[12] and turn our own pipeline into a case study: we document what went wrong, quantify each failure mode with controlled experiments, and extract a corrected protocol.

This paper makes three contributions. First, we provide a controlled demonstration that sampling-pipeline configuration alone can reverse method-level conclusions in DiT PTQ: our legacy pipeline reproduces its historical FID values to within 0.8%, confirming that the A/B isolates the pipeline itself, and under the corrected pipeline the previously uniform benefit of equalization reverses sign. The audit also exposes our full-precision baseline as an unresolved pipeline discrepancy, standing well above published anchors for the same model and task [17], [18]—the anchor's own protocol sensitivity is part of what this audit examines.

Second, we present first-hand evidence that the calibration-error proxy fails rank preservation against FID: 32 records covering 26 distinct configurations share a bit-identical calibration error while their FID values span 98.74 points, and

within every experiment family with usable variance the rank correlation is statistically indistinguishable from zero.

Third, using multi-seed grids under the corrected pipeline, we map the sensitivity of method-level conclusions to the calibration protocol: an error-aware dynamic-programming (EADP) grouping strategy, previously reported as superior, is statistically indistinguishable from a fixed-interval baseline (273.62±1.78 vs. 274.32±1.80 over three seeds), and group-wise channel equalization (Eq) turns from apparently beneficial to harmful; the single surviving conclusion is that grouped quantization substantially outperforms no grouping. Building on these findings, we distill a corrected evaluation-protocol checklist.

Three scope statements are in order. This paper proposes no new quantization method and claims no state-of-the-art result; every comparison is within-protocol, and we do not rank our configurations against other published methods. Our evidence is drawn from a single model (DiT-XL/2 on ImageNet 256×256), a single sampler configuration (DDIM, 50 steps, CFG scale 4.0), and a single headline metric (FID); the resulting generalization questions remain open. Finally, our stance is deliberately neutral: we audit our own system, not others', presenting defects, mechanisms, and corrections with the same evidentiary standard. The remainder of the paper reviews related work (Section 2), formalizes the audit framework (Section 3), details the three pitfalls (Sections 4–6), draws a method-level implication (Section 7), presents the corrected checklist (Section 8), and concludes (Section 9).

2. Related Work

2.1. Post-training Quantization for Diffusion Models and DiTs

PTQ for diffusion models was first systematized for U-Net backbones; Q-Diffusion [9], for example, combined timestep-aware calibration with block reconstruction. The Diffusion Transformer [1], replacing the U-Net with a transformer operating on latent patches, shifted the quantization target and exposed new activation distributions tied to timestep conditioning. PTQ4DiT [2] provided an early DiT-specific framework addressing salient activation channels and timestep-dependent statistics. Q-DiT [3] proposed automatic allocation of quantization granularity with improved handling of activation outliers across timesteps. ViDiT-Q [4] extended DiT quantization to image and video generation, HTG [5] grouped timesteps hierarchically to reduce calibration burden, TQ-DiT [6] pursued time-aware quantization with Hessian-guided optimization, and HQ-DiT [7] explored hybrid floating-point formats. Equalization techniques in the SmoothQuant lineage [8]—redistributing quantization difficulty between activations and weights via per-channel scaling—were imported from the language-model literature into several of these pipelines. Two features of this practice matter for our audit. First, configuration search is commonly delegated to inexpensive screening signals; Q-DiT [3], for instance, selects its hyperparameters using 500 generated images, and the screening signal's validity vis-à-vis full FID ranking is assumed rather than tested. Second, the sampling pipeline itself is rarely reported as an experimental variable: sampler options, guidance-channel scope, and decoder variants appear, if at all, as fixed implementation details. What matters here is not these works' configurations or headline gains but their shared evaluation convention: FID on a fixed sampling pipeline as the decisive metric, with a calibration-side error statistic guiding configuration search.

TreeQ [17] provides the closest prior acknowledgment of the proxy risk: observing that heuristic proxies ignore inter-layer interactions, it relaxes the proxy requirement from absolute accuracy to local order preservation, reported as effective in its setting. Our audit supplies a worst-case counterexample to that premise: 26 distinct configurations share a bit-identical proxy value while their FID values span 98.74 points, so order preservation itself can fail wholesale rather than approximately; a proxy carrying no ranking signal gives any search procedure that consumes it no way to detect the failure. Recent work continues to advance DiT PTQ: LRQ-DiT [20] combines log-domain quantization with adaptive rotation, and MixDiT [21] accelerates DiT inference with mixed-precision MX-format quantization.

2.2. Audits of Generative-model Evaluation

Our work belongs to a genre that audits the evaluation infrastructure of generative modeling rather than proposing new models. Parmar et al. [10] showed that aliased resizing in common FID implementations systematically distorts the metric and can reorder published conclusions. Stein et al. [11] documented failure modes of generative-model evaluation metrics that disadvantage diffusion models. Obukhov et al. [12] provided torch-fidelity, a reference FID implementation with a reproducibility checklist. Sørensen et al. [16] reviewed quantization for deployable diffusion models and called for robust benchmarks and reproducible evaluation protocols. Dufour et al. [19] quantified a different dimension of evaluation randomness: in a two-factor panel over training and generation seeds on SiT networks, the FID variance attributable to retraining was roughly 3.2× that attributable to resampling ($\sigma_{between} = 0.438$ vs. $\sigma_{within} = 0.137$). Their study is complementary to ours—the FID Lottery isolates randomness along the training- and sampling-seed dimension, whereas we isolate the evaluation-protocol configuration dimension. The recurring lesson of this genre is that conclusions inherit the validity of the pipeline that produced them. Our contribution is domain- and evidence-specific: a complete, controlled "defect → wrong conclusion → corrected reproduction" chain inside a single DiT PTQ research program.

2.3. Failure Mechanisms Adjacent to Our Findings

Three recent works provide mechanistic or motivational context. DMQ [13] analyzed why SmoothQuant-style equivalent scaling can fail for diffusion models: the learned scaling redistributes error onto weights, where it accumulates across the denoising trajectory. This mechanism anticipates our observation that group-wise equalization reverses sign under a corrected sampling pipeline, and it implies that equalization methods transferred from language models to DiTs require re-validation. S²Q-VDiT [14], a PTQ framework for video diffusion transformers, reported that quantization performance is highly sensitive to the choice of calibration data; our calibration-protocol grid supplies the complementary image-domain evidence at the level of a controlled audit. Q-Drift [15], a sampler-side correction for quantized diffusion sampling, reported that similarity metrics against full-precision outputs can degrade even when FID improves, warning that full-precision-similarity proxies and distributional metrics need not agree—directly analogous to our calibration-error finding. None of these works assembles the sampling, proxy-metric, and calibration-protocol failure

modes into a single controlled audit of the DiT PTQ evaluation pipeline—the gap this paper fills.

3. Audit Framework and Data Foundation

3.1. Audit Methodology

Our audit applies four methodological principles throughout. (i) Controlled A/B comparison: pipeline variants differ in exactly one factor at a time (e.g., clip_sample on vs. off), with model weights, calibration caches, random seeds, and evaluation code held fixed, so that any metric difference is attributable to the toggled factor. (ii) Multi-seed repetition: headline comparisons are run over n = 3 or n = 5 independent generation seeds and reported as mean ± standard deviation, so that seed-level noise stays visible. (iii) Bit-exact reproducibility checks: selected configurations are rerun end-to-end to verify determinism; independent reruns of the same configuration under the corrected pipeline produced identical FID values (288.6028 in both runs), and interpolation controls at $\alpha = 0.0$ and $\alpha = 0.5$ reproduced the noEq and Eq results bit-for-bit (§4.2), confirming that observed differences reflect genuine factors rather than nondeterminism. (iv) Statistical testing: rank preservation between proxy statistics and FID is assessed with Spearman's ρ and Pearson's r, each reported with its p-value; families lacking within-family variance are reported as untestable rather than assigned a spurious correlation.

3.2. Case-study Setup

The audited system is our own DiT PTQ pipeline built on DiT-XL/2 for class-conditional ImageNet 256×256 generation [1]. Unless otherwise noted, sampling uses DDIM with 50 steps and CFG scale 4.0, and FID is computed on 10K generated images. Deviations from this default are: (i) FID-1K screening runs (flagged as such); (ii) the 100-image sanity runs and (iii) the CFG 1.0 ablation in Table 4; and (iv) the VAE-variant A/B of §4.4. We refer to the original pipeline as the legacy pipeline (clip_sample=True, CFG applied to all four latent channels) and to the repaired one as the corrected pipeline (clip_sample=False, CFG restricted to the three noise channels). The quantization methods under audit are grouped weight-and-activation quantization (W8A8, with W4A8 as a second bit-width) with two grouping strategies—error-aware dynamic programming (EADP) and fixed-interval—and optional group-wise channel equalization (Eq) with interpolation coefficient α. The data foundation also includes a dedicated full-precision baseline rerun under the corrected pipeline at 10K samples (FID = 49.94; §4.4 and Table 4).

3.3. Data Foundation and Reproducibility

Every number in this paper is drawn from a versioned inventory of 209 unique experimental records, consolidated from 204 scanned result files (387 raw records; 178 summary-level duplicates removed; zero unparsed files). Of these, 154 are primary run records and 55 derived records from an earlier unified summary, used only for cross-checking historical claims. For rank-preservation analysis, the inventory provides 99 record-level (calibration error, FID) pairs, deduplicating to 53 configuration-level pairs. Two auxiliary artifacts accompany the run records: a layer-sensitivity summary (227 layers × 50 timesteps) and an analytic storage-overhead table. Table 1 enumerates the experiment families, their protocols, record counts, and roles in this paper.

Table 1. Experiment families in the audit inventory.

Experiment family	Protocol	Records	Role in this paper
unified summary (derived)	legacy aggregate	55	cross-checking historical claims only
ablation_all_fid	legacy (defective)	31	baseline evidence for Pitfalls A and B
repro_seed42 (E-a–E-d, E-legacy)	legacy vs. corrected A/B	17	controlled A/B reproduction (Pitfall A)
E1	corrected	15	main multi-seed results (Pitfall C)
E2	corrected	40	calibration-seed grid (Pitfalls B and C)
E4a	corrected	10	group-count sweep (Pitfall C)
E2 (smoke)	corrected, small-scale	1	sanity check; excluded from conclusions
sanity_ddim	legacy (defective)	10	FP16 baseline diagnosis (Pitfall A)
stage2	calibration stage, no sampling	10	calibration and latency records
stage3_3ch_fix_100	partially corrected	2	clip_sample fix validation
stage3_cfg	legacy (defective)	4	FP16 baseline diagnosis (Pitfall A)
stage3_full	legacy (defective)	10	legacy FP16 reference at 10K samples
stage3_quick	legacy (defective)	4	FP16 baseline diagnosis (Pitfall A)
Total	—	**209**	—

4. Pitfall A: Sampling-Pipeline Configuration Determines Method-Level Conclusions

This section documents the first and most consequential trap: two defects in the sampling pipeline—not in the quantization method—determined the sign of our method-level conclusions. Unless explicitly marked FID-1K, all FID values below are computed on 10K generated images with DDIM 50 steps and CFG scale 4.0 on DiT-XL/2 (ImageNet 256×256) [1].

4.1. Two Latent-space Sampling Defects

Defect 1: clip_sample=True applied in latent space. The diffusers DDIMScheduler option clip_sample=True, designed for pixel-space models whose data range is exactly [-1, 1], clips the predicted clean sample $\hat{x}_0$ to [-1, 1] at every denoising step. DiT, however, operates in the 4-channel KL-f8 VAE latent space, whose values are not bounded to [-1, 1];

clipping therefore truncates the legitimate latent trajectory at all 50 steps, systematically distorting every generated sample. The defect is silent: sampling completes, images look plausible, and FID degrades as if the model were genuinely worse.

Defect 2: classifier-free guidance applied to all four channels. The official DiT sampling code applies guidance only to the first three channels of the four-channel noise prediction—the remaining channel is consumed by the learned-σ head. The legacy pipeline applied the guidance scaling $\varepsilon_{uncond} + s \cdot (\varepsilon_{cond} - \varepsilon_{uncond})$ to all four channels, amplifying a channel the reference implementation deliberately leaves unguided. Because guidance rescales the per-step update, the damage is not a constant offset shared by all rows of a results table, and it can reorder rows.

Both defects lived in shared sampling code, so every legacy-pipeline configuration—including the FP16 baseline and every quantized variant—inherited them. The controlled A/B reproduction below isolates their effect.

Controlled A/B (seed 42, shared calibration cache). We re-ran two W8A8 configurations under both code versions, holding the calibration cache, the sampling seed, and every other setting fixed. Table 2 shows that fixing the sampling code changes FID by up to 49.5 points in opposite directions, and that the legacy code reproduces the prior-draft values to within 0.83%—confirming that the prior numbers were faithful outputs of a defective pipeline, not bookkeeping errors.

Table 2. A/B reproduction at seed 42: identical calibration cache, identical sampling seed; only the sampling code version differs.

Configuration	Legacy code FID	Fixed code FID (E-b)	Prior-draft value (legacy protocol)	Legacy reproduction deviation
ExpC_W8A8_noEq	318.3923	288.6028	315.7823	0.83%
ExpC_W8A8_Eq	278.9706	328.4740	281.3126	0.83%

Two observations follow. First, the noEq configuration *improved* by 29.8 FID points after the fix (318.3923 → 288.6028), while the Eq configuration *degraded* by 49.5 points (278.9706 → 328.4740): the pipeline correction does not shift all configurations uniformly, it re-ranks them. Second, the prior-draft values (315.7823 and 281.3126) are reproduced within 0.83%, so we treat them throughout exclusively as *falsified historical results*—never as current evidence.

Conclusion reversals. Table 3 places each method-level conclusion of the prior draft next to the corresponding corrected-protocol result: all three headline conclusions of the legacy study are overturned or dissolved.

Table 3. Method-level conclusions: legacy pipeline vs. corrected protocol (within-protocol comparisons only).

Prior-draft conclusion (legacy pipeline)	Corrected-protocol evidence	Verdict
Group-wise channel equalization improves FID (favorable in 21/21 within-protocol comparisons)	W8A8: 313.23 ± 6.83 (Eq) vs. 273.62 ± 1.78 (noEq), n = 3 seeds; A/B at seed 42: 328.4740 vs. 288.6028	**Reversed** — Eq is harmful under the corrected protocol
G = 2 timestep groups is optimal	E4a, 5 seeds: G = 4 FID 292.98–308.38 vs. G = 2 FID 363.23–427.55	**Reversed** — larger *G* is clearly better
Error-aware dynamic programming (EADP) grouping outperforms fixed-interval grouping	273.62 ± 1.78 vs. 274.32 ± 1.80, n = 3 seeds	**Dissolved** — no significant separation

Note: the rows draw on experiment families of unequal strength — the G = 2 optimality reversal rests on E4a (n = 5 seeds), whereas the Eq reversal and EADP ≈ FixedInterval rest on E1 (n = 3 seeds); readers should weigh each row's evidence accordingly.

4.2. Bit-identical Reproducibility of the Corrected Pipeline

The corrected pipeline is deterministic given the sampling seed and calibration cache, verified at the bit level. Two independent re-runs of ExpC_W8A8_noEq (E-c runs 1 and 2) both reproduce the E-b value FID = 288.6028 exactly, and the interpolation ablation E-d confirms implementation-level equivalence: α = 0.0 reproduces the noEq result bit-for-bit (288.6028) and α = 0.5 the Eq result (328.4740). These checks establish that the A/B differences in Table 2 stem solely from the two sampling-code defects, not from run-to-run nondeterminism.

4.3. The FID Sample-size Trap

FID is a biased estimator whose bias is material at small sample sizes [11]; evaluation-toolchain work warns against comparing FID values computed at different sample counts or reference statistics [10], [12]. Our own pipeline exhibits the trap directly: under the legacy pipeline, the FP16 baseline scored FID = 77.3194 at 1K images but FID = 46.7842 at 10K images—a 30.5-point spread from sample count alone, with identical sampling code; under the corrected pipeline, FP16 scores 79.15–79.43 at 1K (two VAE variants; §4.4) and FID = 49.94 at 10K, again a ~29-point spread. Consequently, conclusions drawn at 1K-image screening level—including our legacy ablation series and the calibration-set-size sweep of §6.3—cannot be extrapolated to 10K-level claims, and cross-paper comparisons are meaningless unless the sample count and reference set are reported. We therefore attach a sample count to every FID and treat 1K-level numbers strictly as screening evidence. A complementary dimension of FID uncertainty—variance across training and generation seeds (≈3.2× on SiT)—is documented in [19], while this paper focuses on the protocol-configuration dimension.

4.4. FP16 Baseline Anomaly and VAE Exclusion

Our best FP16 baseline under the legacy pipeline was FID = 46.7842 (10K images), far above published full-precision

anchors for the same model and task: FID 19.11 under the closest published protocol (DiT-XL/2, 50 steps, CFG 4.0, 10K images, ADM toolchain, DDPM solver rather than our DDIM) [17, Tab. 1], and FID 16.39 at CFG 5.0 under a DDIM-50 protocol matching our sampler [18]. We take the TreeQ value as primary anchor because its protocol is closest to ours; the DiT paper's FID 2.27 [1] (250 steps, CFG 1.5, FID-50K) is too far from ours to serve as one. Our FP16 anchor is thus roughly 2.4×–2.9× above published baselines, and we treat this gap as a pipeline anomaly to be diagnosed—not as a generic consequence of protocol differences. Because all quantized results are interpreted relative to the FP16 anchor, we investigated three hypotheses:

1. **VAE decoder variant.** The pipeline's VAE is not the MSE-trained (ft-mse) variant: a tensor-by-tensor comparison against a verified ft-mse checkpoint shows 0/106 differing encoder tensors but 138/138 differing decoder tensors, i.e., the decoder is most likely ft-ema (or original kl-f8), while project documentation incorrectly claimed "sd-vae-ft-mse". The A/B result *excludes* this as the cause: swapping in the verified ft-mse decoder changes FID-1K from 79.15 to 79.43 ($\Delta = 0.28$). The VAE variant matters for reporting accuracy, but not for the 46.78 anomaly.

2. **Sampling defects.** Excluded at both scales. At 1K, the corrected pipeline scores 79.15–79.43 versus 77.3194 for the legacy pipeline (~2 points). At 10K, a dedicated FP16 run fully matched to the E1 and E2 protocol (DDIM 50 steps, CFG 4.0, clip_sample=False, 10K-image FID reference) yields FID = 49.94 versus 46.78 for the legacy pipeline—the sampling fix moves the full-precision baseline by only 3.2 points (6.8%), nowhere near the >2× gap. This exclusion presupposes that the corrected pipeline is free of unknown defects—a presupposition the residual anomaly shows cannot be fully verified—so we state the exclusion as holding under the assumption that the corrected pipeline is sound.

3. **Residual suspects.** With the VAE variant and the sampling defects excluded, the remaining candidates are the FID reference set (our 10K ImageNet validation subset vs. the 50K references used in published evaluations), the DiT checkpoint itself, and systematic differences from the ADM evaluation suite. These cannot be resolved with the data available to this study; we report the anomaly as bounded but open: FID 49.94 under our corrected 10K protocol versus 19.11 [17, Tab. 1] under the closest published protocol.

Table 4 summarizes all FP16 variants, including the completed corrected-pipeline 10K run (FID = 49.94; 2.15 h wall-clock on an RTX 5090).

Table 4. FP16 baseline FID across pipeline variants (all DDIM 50 steps; FID-1K/FID-10K by image count).

Pipeline variant	Images	CFG	FID	Note
Legacy (clip_sample=True, CFG on 4 channels)	100	4.0	194.4639	screening
Legacy, clip_sample fixed only	100	4.0	193.9283	isolates Defect 1: ~0.5-point effect at 100 images
Legacy	1,000	4.0	77.3194	FID-1K
Legacy	10,000	4.0	46.7842	FID-10K; 2.4×–2.9× above published anchors [17, Tab. 1], [18]
Legacy, guidance off	100	1.0	219.4710	CFG ablation
Corrected, current VAE	1,000	4.0	79.15	FID-1K
Corrected, verified ft-mse VAE	1,000	4.0	79.43	FID-1K; $\Delta = 0.28$ vs. current VAE
Corrected, current VAE	10,000	4.0	49.9426	FID-10K, matched to the E1 and E2 protocol; 2.6×–3.0× above published FP anchors

Table 5 makes the incomparability of published full-precision baselines explicit: for the same model and task, published FP FID values scatter from 2.27 to 19.11, driven by four protocol knobs—sampler, step count, CFG scale, and sample count/reference set. Merely lowering CFG to 1.5 moves the published anchors to 2.23–6.28 [17], [18]. Absolute FID values are therefore not comparable across papers; only within-protocol comparisons are meaningful—hence this audit's argument for protocol transparency.

Table 5. Published full-precision baselines for DiT-XL/2 (ImageNet 256×256) are not mutually comparable.

Paper	FP FID	Sampler / steps	CFG	Samples / evaluation
DiT [1]	2.27	DDPM, 250 steps	1.5	FID-50K
PTQ4DiT [2]	4.53	DDPM, 250 steps	none	FID-10K
TQ-DiT [6]	4.62	DDPM, 250 steps	none	FID-10K
Q-DiT [3]	12.40 (100 steps) / 13.47 (50 steps); 5.31 / 5.32 at CFG 1.5	DDIM, 50/100 steps	none / 1.5	FID-10K
RSTR [18]	16.39 (CFG 5.0) / 19.71 (CFG 7.5) / 2.23 (CFG 1.5)	DDIM, 50 steps	1.5 / 5.0 / 7.5	FID-10K, FP16
TreeQ [17]	19.11 (CFG 4.0) / 6.28 (CFG 1.5)	DDPM solver, 50 steps	4.0 / 1.5	FID-10K, ADM toolchain

Summary of Pitfall A. Sampling-pipeline configuration is not an implementation detail; in our case study it decided the sign of every method-level conclusion. The corrected pipeline is bit-reproducible, and the FID sample count and VAE variant must be reported alongside any number. Yet even after all corrections the FP16 anchor remains an open anomaly that

bounds every quantitative conclusion of this paper, including those under the corrected protocol: what we claim is the existence and direction of each pitfall, while specific effect sizes may shift with the protocol.

5. Pitfall B: Proxy Metrics Fail Rank Preservation

5.1. The Community Practice We Inherited

Calibration-error- or reconstruction-loss-based proxies are widely used to select PTQ configurations without paying for sampling and FID evaluation. Q-DiT [3], for instance, searches hyperparameters using 500 generated samples, and related work has acknowledged that MSE-style reconstruction proxies do not guarantee rank preservation with respect to generation quality. The practice is attractive because FID at 10K images is expensive; our audit quantifies what it costs in correctness.

5.2. First-hand Evidence: Bit-identical Proxy, 98.74-point FID Spread

Across our legacy ablation series (ablation_all_fid) and the A/B reproduction series (repro_seed42), **32 records covering 26 distinct configurations share exactly the same calibration error, 1.659664e+11 (exact value 165966437415.6679077148), bit-for-bit**, because they all reuse the same seed-42 calibration cache. Their FID values span **262.71 to 361.45 — a range of 98.74 points**. The proxy cannot distinguish configurations whose generation quality differs by nearly a hundred FID points; Table 6 shows representative pairs.

Table 6. Configurations with bit-identical calibration error (1.659664e+11) but divergent FID (legacy protocol unless noted).

Configuration pair (same calib. error)	FID (config 1)	FID (config 2)	FID gap
ExpC_W3A8_Eq / ExpC_W3A8_noEq	287.4875	300.9183	13.43
ExpC_W4A8_Eq / ExpC_W4A8_noEq	332.4404	353.0350	20.59
ExpC_W8A8_Eq / ExpC_W8A8_noEq	281.3126	315.7823	34.47
ExpB_alpha0.0 / ExpB_alpha1.0	320.1901	262.7111	57.48
ExpC_W3A8_noEq (ablation) / ExpC_W3A8_noEq (fixed code, E-a)	300.9183	272.9117	28.01
ExpC_W8A8_Eq (ablation) / ExpC_W8A8_Eq (fixed code, E-b)	281.3126	328.4740	47.16
ExpC_W8A8_noEq (legacy code) / ExpC_W8A8_noEq (fixed code, E-b)	318.3923	288.6028	29.79

The same pathology appears in the corrected-protocol experiments (family E2): within each calibration seed, **7 of 8 configurations share a bit-identical calibration error** while their FID values spread over 66.90–83.93 points per seed (Table 9, §6). We tested rank preservation statistically across every family offering within-family variance (Table 7).

Table 7. Rank-preservation statistics, calibration error vs. FID.

Experiment family	n	Spearman ρ (p)	Pearson *r* (p)	Note
ablation_all_fid (legacy protocol)	31	0.100 (0.594)	0.133 (0.474)	no rank preservation
repro_seed42	17	not computable	not computable	proxy has zero variance — itself the evidence
E2, seed 0	8	0.082 (0.846)	0.264 (0.527)	
E2, seed 1	8	0.082 (0.846)	0.177 (0.676)	
E2, seed 2	8	−0.082 (0.846)	0.067 (0.874)	
E2, seed 3	8	0.082 (0.846)	0.164 (0.698)	
E2, seed 4	8	0.082 (0.846)	0.107 (0.801)	
All-pairs pool	99	0.159 (0.116)	0.103 (0.308)	no significant correlation
E2 + E4a mixed pool	50	0.238 (0.0962)	0.791 (8.38e-12)	caution: Pearson is driven by E4a's legitimate G = 2 vs. G = 4 differences, not within-family ranking

No family shows significant Spearman rank correlation: within every individual family, $|\rho| \leq 0.100$ with $p \geq 0.116$. The pooled mixed-family correlation ($\rho = 0.238$, $p = 0.096$) and the apparently high Pearson value are aggregation artifacts driven by the legitimate *G*-sweep differences in E4a, reported only as a warning against pooling heterogeneous configurations. The repro_seed42 family is the purest demonstration: with zero proxy variance across 17 records, a rank correlation is mathematically undefined, yet FID spans 272.91–361.45.

The practical consequence is stark. In the legacy study, this proxy was the basis on which configuration families were shortlisted for full FID evaluation: any shortlisting rule applied to Table 6's rows is effectively a coin flip, because the proxy values being compared are literally the same number. A selection procedure can tolerate a noisy proxy; it cannot tolerate one with zero discriminative power over the exact design dimensions (bit-width, equalization, interpolation strength) being selected. This is the sense in which the proxy "fails"—stronger than merely showing a weak correlation.

5.3. Mechanism: The Proxy is Computed Before the Transform it is Supposed to Evaluate

The failure has a concrete structural cause. The calibration-error value is computed from group-wise quantizer statistics

before the equalization transform is applied; equalization therefore cannot change the measured metric, although it changes the channel scaling used during quantized inference. More generally, because the error is accumulated from per-(layer, timestep-group) statistics of the calibration pass, configurations that share a calibration cache—differing in bit-width, grouping strategy, equalization, or even the sampling code version used afterwards—all inherit the same proxy value. The metric measures the calibration set and the cache, not the deployed configuration.

5.4. Side Finding: Calibration-Cache Consistency is Itself a Trap

During the audit we found one systematic exception to within-seed bit-identity: in every E2 seed, ExpC_W3A8_noEq's calibration error differs from the other seven configurations at a relative deviation of $\approx 4\times10^{-6}$ (identical to six significant digits; e.g., seed 0: 164719304816.3294372559 vs. 164719979632.9858703613). The pattern indicates a slightly different calibration path—a cache miss followed by recomputation. The deviation is numerically negligible, but it shows that "same cache" claims must be verified, not assumed; we recommend a cache-consistency check in any proxy-based protocol (§8).

Summary of Pitfall B. In our case study the calibration-error proxy failed rank preservation in every testable family, by construction (pre-transform computation) and by sharing (cache reuse). Any selection, search, or early-stopping decision made on this proxy is unsubstantiated unless rank preservation is explicitly validated on held-out configurations with ground-truth FID.

6. Pitfall C: Calibration Protocol Sensitivity

The third pitfall is that corrected-protocol results remain sensitive to calibration-protocol choices—seed, grouping granularity, calibration-set size—in ways single-run reporting hides. Table 8 gives the main corrected-protocol result (E1: 5 configurations × 3 sampling seeds, FID-10K).

Table 8. E1 main results under the corrected protocol (FID-10K, mean ± std over seeds 0–2).

Configuration	Grouping	Eq	FID mean ± std	Per-seed FID (s0, s1, s2)
Naive_W8A8	none (G = 1)	no	388.96 ± 7.50	394.04, 380.34, 392.50
FixedInterval_G8_W8A8	fixed, G = 8	no	274.32 ± 1.80	274.12, 272.63, 276.21
ErrorAwareDP_G8_W8A8	EADP, G = 8	no	273.62 ± 1.78	272.78, 272.41, 275.66
ErrorAwareDP_G8_Eq_W8A8	EADP, G = 8	yes	313.23 ± 6.83	318.66, 305.56, 315.46
ErrorAwareDP_G8_Eq_W4A8	EADP, G = 8	yes	285.83 ± 5.00	287.95, 280.11, 289.42

6.1. Calibration-seed Sensitivity

E2 repeats 8 configurations across 5 calibration seeds. The calibration error varies across seeds from **1.6314e+11 to 1.6858e+11** (a relative spread of ≈3.3%), while — per Pitfall B — remaining bit-identical across 7 of 8 configurations within each seed. Table 9 shows the proxy's seed-to-seed variation is small beside the FID spread inside each identical proxy value (66.90–83.93 points per seed): protocol noise in the proxy is not the dominant problem; its insensitivity to configuration is.

Table 9. E2: calibration error and FID range per calibration seed (8 configurations each).

Calib. seed	Calib. error (7 configs, bit-identical)	Calib. error (W3A8_noEq, cache miss)	FID range	FID spread
0	1.647200e+11	1.647193e+11	276.82–343.73	66.90
1	1.643736e+11	1.643729e+11	276.88–354.01	77.13
2	1.685842e+11	1.685836e+11	277.24–361.17	83.93
3	1.631403e+11	1.631397e+11	280.09–346.91	66.82
4	1.634515e+11	1.634509e+11	278.48–344.24	65.75

6.2. Grouping Strategy: Sophistication does Not Pay, Grouping Itself Does

Two within-protocol findings survive the corrected protocol. First, **EADP ≈ FixedInterval**: error-aware dynamic-programming grouping scores 273.62 ± 1.78 against 274.32 ± 1.80 for plain fixed-interval grouping at $G = 8$ — a 0.70-point mean difference, smaller than half the per-configuration standard deviation and far from significant at $n = 3$. A paired power analysis (computed from the E1 per-seed records) makes this precise: the three seed-paired differences are -1.33, -0.22, and -0.55 FID points (std = 0.57), so a two-sided paired t-test at $\alpha = 0.05$ with $n = 3$ reaches 80% power only for effects of at least

$$\mathrm{MDE} = \left(t_{0.975,2} + t_{0.80,2}\right) \cdot \frac{\mathrm{std}}{\sqrt{3}} \approx 1.76 \qquad (1)$$

FID points. The observed 0.70-point gap is far below this MDE, so "no significant difference" should be read as "no detectable difference at $n = 3$": if a true difference exists, it is smaller than ≈ 1.76 FID points. We report this as a bounded non-finding rather than as proof of equivalence. The elaborate grouping algorithm buys nothing measurable over a fixed grid. Second, and this is the **only robust positive method-level conclusion of our study**, grouping itself matters enormously: Naive W8A8 without timestep grouping scores 388.96 ± 7.50, versus 274.32 ± 1.80 with fixed-interval grouping — a ~115-point FID reduction, reproduced on every seed. Calibration-protocol sophistication is optional; calibration-protocol adequacy is not. Tightening this interval with additional seeds

is inexpensive (≈2.15 h per configuration) and is a natural next step.

6.3. Calibration-set Size (Screening-level Evidence)

All data in this subsection come from 1K-image, screening-level experiments under the legacy pipeline. They serve only as a historical record and cannot support any corrected-protocol conclusion; the subsection is retained for full disclosure of the audit trail.

The legacy ExpD sweep varied the calibration-set size over {64, 128, 256, 512} images at fixed W8A8: FID 359.67, 356.82, 358.51, and 355.50 respectively—a total spread of ~4.2 points with no monotone trend. Per §4.3, such screening results do not extrapolate to FID-10K conclusions, and we make no claim that 64-image calibration is "sufficient" at evaluation scale. A corrected-protocol calibration-size sweep at 10K is future work.

6.4. Group-count Sweep and Storage Overhead

E4a sweeps the timestep-group count under the corrected protocol: $G = 2$ scores FID 363.23–427.55 across 5 seeds, $G = 4$ scores 292.98–308.38 — here the calibration-error difference between G values is legitimate (different grouping counts produce different quantizer statistics), and the proxy *does* order the two groups correctly ($G = 2$: 2.78e+11–2.92e+11; $G = 4$: 1.75e+11–1.84e+11). Together with E1 ($G = 8 \approx 274$), this contradicts the legacy $G = 2$-optimality conclusion (Table 3).

Table 10 gives the analytic storage overhead as a function of G, under the authoritative code-level enumeration of 112 quantized linear layers (28 blocks × [attn.qkv, attn.proj, mlp.fc1, mlp.fc2]). Two 140-layer figures circulated in earlier materials and must not be conflated: the stage2 layer-statistics field counts 28 additional attn.proj_drop dropout entries, and the prior draft's prose incorrectly included 28 adaLN-modulation linears; neither is quantized. The authoritative count for storage accounting is 112.

Table 10. Calibration-parameter storage overhead vs. group count G (code-actual enumeration, 112 layers; W8 model size 675.1 MiB; FP32 storage).

G	Params (weight scales / act. / Eq)	noEq total	+Eq total	noEq (% of model)	+Eq (% of model)
1	290,304 / 224 / 225,792	1.108 MiB	1.970 MiB	0.164%	0.292%
2	290,304 / 448 / 451,584	1.109 MiB	2.832 MiB	0.164%	0.419%
4	290,304 / 896 / 903,168	1.111 MiB	4.556 MiB	0.165%	0.675%
8	290,304 / 1,792 / 1,806,336	1.114 MiB	8.005 MiB	0.165%	1.186%
16	290,304 / 3,584 / 3,612,672	1.121 MiB	14.902 MiB	0.166%	2.207%
32	290,304 / 7,168 / 7,225,344	1.135 MiB	28.697 MiB	0.168%	4.251%

Summary of Pitfall C. Under the corrected protocol, multi-seed reporting is what separates signal from noise: it confirms one robust positive conclusion (grouping ≫ no grouping), dissolves one claimed algorithmic advantage (EADP), and bounds the rest (calibration seed, calibration-set size) as secondary effects at the current screening level of evidence.

7. Method-Level Implication: Equalization Transfer Failure

The starkest consequence of Pitfall A is methodological. Group-wise channel equalization (Eq) — a SmoothQuant-style [8] technique that migrates activation outliers into weights via per-channel scaling — was our prior draft's flagship component, favorable in 21/21 within-protocol comparisons under the legacy pipeline. Under the corrected protocol, Eq is *harmful* at W8A8: FID degrades from 273.62 ± 1.78 to 313.23 ± 6.83, a +14.4% regression reproduced across all three seeds and independently in the seed-42 A/B (288.6028 → 328.4740). The method conclusion was determined entirely by the evaluation pipeline: same algorithm, same checkpoint, same calibration cache—opposite verdict.

We do not read this as "equalization never works for DiT", and we make no such counter-claim. A plausible mechanism, consistent with DMQ [13], is that equalization-style scaling amplifies weight error, which then accumulates across the denoising trajectory—exactly the regime (50 steps) where our FID penalty would grow. Under the legacy pipeline, the sampling defects apparently contributed errors large and correlated enough to mask—or be partially cancelled by—the equalization transform, producing a spurious benefit. We state this mechanism explicitly as a hypothesis, consistent with our data but not verified by controlled experiments. Distinguishing "Eq helps under defect X" from "Eq helps" is impossible from legacy data alone and requires dedicated mechanism experiments beyond the scope of this audit; the transfer of SmoothQuant-class methods from LLMs to DiT must therefore be re-validated under a verified sampling protocol before any method-level claim is made.

Two implications follow. First, the legacy defects did not merely add noise; they added *structured* error correlated with the very design dimension Eq manipulates (per-channel activation scaling), so the spurious benefit was stable across seeds and ablation rows—21/21 favorable comparisons—and therefore highly convincing internally. Stability of a within-protocol result is not evidence of its external validity: a biased pipeline produces biased results reproducibly. Second, published cross-paper comparisons typically pool numbers produced by different sampling stacks, CFG channel conventions, VAE decoders, and FID sample counts; our A/B shows that two of these factors alone move FID by tens of points in configuration-dependent directions (Table 2), so a method-level margin smaller than the pipeline-induced spread cannot support a ranking claim across papers—only within a shared protocol.

We also decline to propose a replacement proxy metric. Q-Drift [15] warns that full-precision-similarity-style proxies can move *against* FID; combined with our §5 finding that the

calibration-error proxy fails rank preservation by construction, the responsible conclusion is narrower: proxies must earn trust per protocol, by demonstrated rank preservation against ground-truth FID on held-out configurations—not inherit it from convention. Designing a validated, sampling-aware proxy for DiT PTQ remains an open problem; the checklist in §8 is the interim safeguard.

8. Recommendations: A Corrected Evaluation Protocol Checklist

The constructive contribution of this audit is a checklist for evaluating DiT PTQ (and diffusion PTQ more broadly). Each item is grounded in a specific failure we measured; none requires new methods—they are reporting and verification practices, in the spirit of prior evaluation audits [10]–[12], [16].

Table 11. Corrected evaluation protocol checklist, with evidence pointers.

#	Checklist item	Failure it prevents	Evidence
C1	Report scheduler type, step count, and clip_sample (or equivalent clipping); justify against the model's native space (pixel vs. latent)	Silent latent-trajectory truncation	§4.1, Table 2
C2	Verify CFG channel handling against the reference implementation (incl. learned-σ channel layout)	Guidance applied to non-noise channels; config-dependent distortion	§4.1, Tables 2–3
C3	Report the exact VAE variant/weights (hash-level identity), not the intended name	Documentation–artifact mismatch; irreproducible decoding	§4.4
C4	Attach FID sample count and reference-set size to every number; never mix FID-1K screening with FID-10K conclusions	~30-point sample-size bias	§4.3, Table 4
C5	Report multi-seed mean ± std ($n \geq 3$) for every configuration	Single-run protocol noise masquerading as method effects	§6, Table 8
C6	Validate proxy-metric rank preservation (Spearman ρ vs. ground-truth FID) on held-out configurations before using the proxy for selection	Proxy-driven selection of worse models	§5, Tables 6–7
C7	Verify calibration-cache consistency (bit-level or hash-level) across configurations sharing a cache	Silent cache miss/recompute; "same proxy" assumptions	§5.4, Table 9
C8	Report calibration-artifact storage overhead vs. model size	Hidden deployment costs; enumeration discrepancies	§6.4, Table 10

C1–C4 are properties of the *evaluation harness*, C5 of the *reporting discipline*, and C6–C8 of the *calibration protocol*; our case study shows failures of all three kinds in a single pipeline, each sufficient on its own to flip a method-level conclusion.

Two usage notes accompany the checklist. First, the cheapest verifications come first: C1–C3 and C7–C8 are one-time metadata or consistency checks costing minutes, whereas C4–C6 require additional sampling runs; the cheap checks would have caught both sampling defects and the cache anomaly before any FID was computed. Second, the checklist is intentionally scoped to what we measured: it prescribes no particular sampler, guidance scale, or calibration-set size—§6 shows that several such choices are secondary once the harness is correct—and it complements, rather than replaces, the evaluation-suite guidance of prior audits [10]–[12]. We will adopt C1–C8 in our own revised study, and we invite reviewers of future DiT PTQ work to request the same disclosures.

9. Conclusion

We audited the evaluation pipeline of our own DiT PTQ study and found three systematic traps, each verified with controlled experiments on 209 deduplicated result records. (i) **Sampling-pipeline configuration decides method conclusions**: two latent-space sampling defects (clip_sample=True; CFG on all four channels) reversed the verdict on equalization, group count, and grouping algorithm, while the legacy code reproduced our prior numbers to within 0.83%—the pipeline was faithful; the conclusions were artifacts. (ii) **The calibration-error proxy fails rank preservation**: 32 records spanning 26 configurations share a bit-identical proxy value while FID ranges over 98.74 points; no family shows significant Spearman correlation, and the failure is structural—the proxy is computed before the transform it is meant to evaluate. (iii) **Calibration-protocol sensitivity is real but secondary**: multi-seed reporting isolates the study's one robust positive finding—timestep grouping improves FID by ~115 points over no grouping ($388.96 \pm 7.50 \rightarrow 274.32 \pm 1.80$)—while the elaborate EADP grouping matches a fixed interval grid (273.62 ± 1.78 vs. 274.32 ± 1.80). We distill the findings into an eight-item protocol checklist (Table 11).

Limitations. Our audit is deliberately narrow: a single model (DiT-XL/2 on ImageNet 256×256); a single quality metric (FID), whose own pitfalls bound what any FID-based audit can conclude; protocol-specific findings whose magnitudes (though not their existence) may not transfer to other samplers, models, or datasets; and a bounded but open FP16 baseline anomaly (§4.4)—our corrected 10K baseline (FID = 49.94) remains 2.6×–3.0× above the closest published anchors (19.11 [17, Tab. 1]; 16.39) even after the VAE variant and both sampling defects are excluded, with the residual most plausibly attributable to reference-set and evaluation-suite differences this study cannot isolate. All absolute effect sizes reported here are therefore protocol-dependent; what generalizes is the existence and direction of the pitfalls. All comparisons are within-protocol; we claim no state-of-the-art result.

Data availability. The complete set of 209 deduplicated experimental records, per-run metrics, analysis scripts, and A/B reproduction artifacts supporting this study are available from the corresponding author upon reasonable request.

References

[1] Peebles, W., & Xie, S. (2023). Scalable Diffusion Models with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV51070.2023.00449.

[2] Wu, J., Wang, H., Shang, Y., Shah, M., & Yan, Y. (2024). PTQ4DiT: Post-training Quantization for Diffusion Transformers. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).

[3] Chen, L., Meng, Y., Tang, C., Ma, X., Jiang, J., Wang, X., Wang, Z., & Zhu, W. (2025). Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Zhao, T., Fang, T., Liu, E., Rui, W., Soedarmadji, W., Li, S., Lin, Z., Dai, G., Yan, S., Yang, H., Ning, X., & Wang, Y. (2024). ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation. arXiv preprint arXiv:2406.02540.

[5] Ding, N., Han, J., Tian, Y., Xu, C., Han, K., & Tang, Y. (2025). Post-Training Quantization for Diffusion Transformer via Hierarchical Timestep Grouping. arXiv preprint arXiv:2503.06930.

[6] Hwang, Y., Lee, H., & Kang, J. (2025). TQ-DiT: Efficient Time-Aware Quantization for Diffusion Transformers. arXiv preprint arXiv:2502.04056.

[7] Liu, W., & Zhang, S. Q. (2024). HQ-DiT: Efficient Diffusion Transformer with FP4 Hybrid Quantization. arXiv preprint arXiv:2405.19751.

[8] Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In Proceedings of the International Conference on Machine Learning (ICML). https://doi.org/10.48550/arXiv.2211.10438.

[9] Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., & Keutzer, K. (2023). Q-Diffusion: Quantizing Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV51070.2023.00244.

[10] Parmar, G., Zhang, R., & Zhu, J.-Y. (2022). On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR52688.2022.01137.

[11] Stein, G., Cresswell, J. C., Hosseinzadeh, R., Sui, Y., Ross, B. L., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, J. E. T., & Loaiza-Ganem, G. (2023). Exposing Flaws of Generative Model Evaluation Metrics and their Unfair Treatment of Diffusion Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).

[12] Obukhov, A., Seitzer, M., Lin, E. Y.-J., Kyl, J., Zhydenko, S., & willylulu. (2021). torch-fidelity (Version 0.3.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4957738.

[13] Lee, D., Hur, J., Shon, H., Lee, J. Y., & Kim, J. (2025). DMQ: Dissecting Outliers of Diffusion Models for Post-Training Quantization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[14] Feng, W., Qin, H., Yang, C., Li, X., Yang, H., Li, Y., An, Z., Huang, L., Magno, M., & Xu, Y. (2025). S²Q-VDiT: Accurate Quantized Video Diffusion Transformer with Salient Data and Sparse Token Distillation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).

[15] Ryu, S., Salzmann, M., & Javed, S. (2026). Q-Drift: Quantization-Aware Drift Correction for Diffusion Model Sampling. arXiv preprint arXiv:2603.18095.

[16] Sørensen, M., Holm, F., Kristensen, A., Jensen, S., Madsen, L., Rasmussen, E., & Aline, C. (2025). Quantization as a Foundation for Deployable High Performance Diffusion Models within the Landscape of Large Scale Generative AI. TechRxiv preprint. https://doi.org/10.36227/techrxiv.175624440.05705133.

[17] Yang, K., Yang, K., Wu, B., Zhang, X., Yang, Q., Qin, H., Zhang, H., & Zhang, Y. (2025). TreeQ: Pushing the Quantization Boundary of Diffusion Transformer via Tree-Structured Mixed-Precision Search. arXiv preprint arXiv:2512.06353.

[18] Sun, R., Yang, T., Niu, W., & Sun, J. (2025). RSTR: Reducing SpatioTemporal Redundancy in Diffusion Transformers. arXiv preprint arXiv:2512.14096.

[19] Dufour, N., Efros, A. A., & Pérez, P. (2026). The FID Lottery: Quantifying Hidden Randomness in Generative Model Evaluation. arXiv preprint arXiv:2606.20536.

[20] Yang, L., Lin, H., Zhao, T., et al. (2025). LRQ-DiT: Log-Rotation Post-Training Quantization of Diffusion Transformers for Image and Video Generation. arXiv preprint arXiv:2508.03485.

[21] Kim, D., Hwang, J., Oh, C., & Park, J. (2025). MixDiT: Accelerating Image Diffusion Transformer Inference with Mixed-Precision MX Quantization. arXiv preprint arXiv:2504.08398.