JOINT BRANCH-SPACE TRANSFORM CODING FOR DIFFUSION ACTIVATION QUANTIZATION WITH CLASSIFIER-FREE GUIDANCE

Anonymous authors
Paper under double-blind review



Figure 1: GCBT is a plug-and-play activation transform built on top of existing post-training quantizers in CFG-guided diffusion models. By applying a branch-wise rotation prior to quantization, it requires minimal modifications to model parameters or base pipelines. On PixArt-Σ (top) and Sana-1.6B (bottom), GCBT demonstrates superior visual fidelity over naive PTQ and SVDQuant.

ABSTRACT

Post-training quantization for diffusion models increasingly exploits timestep, feature, and layer structure. While recent work has begun incorporating CFG structure into diffusion quantization, activation quantization still operates independently across conditional and unconditional coordinates, leaving cross-activation structure unexploited. We show that matched CFG activations form a strongly correlated two-dimensional source and that, under a fixed bit budget, the choice of branch coding basis materially affects quantization fidelity. Motivated by this observation, we introduce **branch-space transform coding**, which rotates matched CFG branches via an offline derived 2×2 orthogonal matrix, requiring minimal modifications to model parameters or the quantization pipeline. We further derive the **Guidance-Correlation Branch Transform (GCBT)**, which jointly incorporates the CFG guidance direction and cross-branch second moments. Under an equal-rate quantization-noise surrogate, GCBT admits a closed-form per-layer solution without gradient optimization or angle search. Applied on top of existing diffusion PTQ methods, GCBT yields statistically significant fidelity gains in most evaluated comparisons with no statistically significant degradation, while leaving the underlying host quantization pipeline unchanged.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021) achieve high generation quality at the cost of repeated, computationally heavy denoising steps. This has motivated extensive work on post-training quantization (PTQ) (Chen et al., 2025; Li et al., 2023; Shao et al., 2026; Li et al., 2025; Zhao et al., 2025; Shafi & Suma, 2026) to reduce inference overhead via low-bit representations of weights and activations. Recent methods increasingly exploit activation structure during diffusion inference, exploiting (i) variation across denoising timesteps (Chen et al., 2025; Zhao et al., 2025), (ii) architectural and feature heterogeneity within the denoising model (Li et al., 2023; Chen et al., 2025), and (iii) activation outliers (Li et al., 2025; 2026). These advances raise a natural question: *what additional structural information in diffusion inference can be exploited to improve activation quantization?*

One such structure is the paired computation in classifier-free guidance (CFG) (Ho & Salimans, 2022). CFG plays a central role in modern diffusion pipelines by strengthening the influence of the condition on generation. At each denoising step, CFG evaluates the same model weights with and without condition, producing matched predictions p_c and p_u and then combining them using a guidance scale s. We denote the corresponding activations as X_c and X_u, and refer to the two-dimensional axis formed by this matched pair as the ***CFG branch space*.**

While recent work has begun incorporating CFG into post-training calibration (Shao et al., 2026; Shafi & Suma, 2026), activation quantization still typically operates on X_c and X_u independently in separate branch coordinates. This choice implicitly fixes the branch basis, overlooking the fact that both activations arise from matched CFG passes through the same model weights.

We find two properties of this branch space that are relevant to activation quantization. First, matched X_c and X_u activations are strongly correlated, exhibiting cross-branch correlation coefficients of a median uncentered correlation of 0.86 (vs. 0.33 for mismatched pairs) across DiT-XL/2 (Peebles & Xie, 2023) layers (Figure 2). Second, the two CFG branches are coupled at the model output by the known guidance direction $g_s = [s, 1-s]^\top$, which varies with the guidance scale s. As illustrated in Figure 3, standard PTQ quantizes activations independently along native branch axes, imposing a structural mismatch between the quantization grid, principal activation variance, and the guidance vector g_s. Under a fixed activation-bit budget, this directional mismatch limits quantization fidelity.

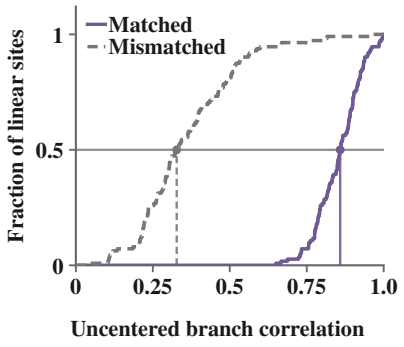


Figure 2: **Matched CFG branches are strongly correlated.** Matched CFG branches exhibit substantially stronger cross-branch correlation than deliberately mismatched pairs.

Motivated by these observations, we introduce **branch-space transform coding**. Before activation quantization, we apply a 2×2 orthogonal branch rotation matrix (R) across each matched conditional/unconditional pair, quantize the transformed activations on top of the existing low-bit quantizer, and invert the transform afterward. The procedure does not change model parameters, nominal activation precision, or the number of quantized values; it changes only the branch coordinates in which activation quantization is performed. Branch-space coding is therefore complementary to existing timestep-, feature-, layer-, and precision-aware diffusion PTQ methods.

To determine the optimal rotation matrix R, we derive the ***Guidance-Correlation Branch Transform (GCBT)***, which jointly incorporates the CFG guidance direction and per-layer cross-branch second moments. Under an equal-rate quantization-noise surrogate, this formulation yields a closed-form per-layer solution without gradient optimization or angle search. Across diverse diffusion PTQ backbones, models, and datasets, GCBT consistently improves reconstruction and perceptual fidelity, establishing CFG branch space as a complementary axis for activation quantization.

Our work makes the following contributions:

- **CFG Branch Space as a Design Axis.** To our knowledge, we are the first to treat the CFG branch axis as an explicit coding space, demonstrating that cross-branch structure between matched activations directly governs quantization fidelity.

- **Branch-Space Transform Coding & GCBT.** We introduce a lightweight branch-space transform coding framework that incorporates the CFG guidance direction and cross-branch second moments into the coding basis.
- **Closed-Form Global Optimum.** We derive a guidance-weighted distortion objective over 2×2 orthogonal rotations and prove that its global optimum admits a closed-form per-layer solution under an equal-rate quantization-noise surrogate.
- **Broad PTQ Compatibility.** We apply GCBT on top of existing activation quantizers without modifying the underlying quantization method, achieving consistent improvements in reconstruction and perceptual fidelity.

2 Related Work

Post-Training Quantization for Diffusion Models. Diffusion PTQ increasingly exploits structure specific to iterative generation. Q-Diffusion (Li et al., 2023) addresses timestep-dependent activation distributions and architectural outliers, while Q-DiT (Chen et al., 2025) and ViDiT-Q (Zhao et al., 2025) further account for spatial, temporal, layer, and channel variation in Diffusion Transformers. SVDQuant (Li et al., 2025) handles activation outliers through smoothing and low-rank compensation, and DeltaQuant (Li et al., 2026) exploits spatiotemporal activation structure for low-bit video diffusion. These methods adapt quantization to diffusion-specific activation structure, but retain the native conditional/unconditional branch coordinates during activation quantization.

Rotation-Based Quantization. Rotary quantization has been widely studied beyond diffusion models. QuaRot (Ashkboos et al., 2024) applies orthogonal rotations to reduce activation outliers, while SpinQuant (Liu et al., 2025) learns rotation matrices for improved low-bit quantization. Related methods use Hadamard-based incoherence processing (Tseng et al., 2024) or learnable linear transformations (Sun et al., 2025; Hu et al., 2025) to improve the quantizability of weights and activations. In diffusion models, TR-DQ (Shao et al., 2026) introduces timestep-dependent feature rotations, while OrbitQuant (Lee et al., 2026) quantizes in a normalized rotated basis. These approaches transform feature or hidden-state coordinates; in contrast, our transform acts across the two-dimensional CFG branch axis while leaving token and feature coordinates unchanged.

CFG-Aware Quantization. Recent diffusion PTQ methods have also begun to exploit the relation between CFG branches. TR-DQ (Shao et al., 2026) uses redundancy across CFG executions, while guidance-aware methods such as GAMP (Shafi & Suma, 2026) calibrate quantization with respect to the guided prediction and CFG sensitivity. These methods incorporate CFG structure while retaining the native conditional/unconditional branch basis. Our work studies a complementary variable: the *branch coding basis* itself. We treat the CFG branch axis as an explicit quantization space and select its basis from the CFG branch direction and cross-branch second moments.

3 Method

3.1 Problem Setup

Classifier-free guidance (CFG) (Ho & Salimans, 2022) combines the conditional and unconditional branch predictions as

$$p_{\text{cfg}} = sp_c + (1-s)p_u = g_s^\top \begin{bmatrix} p_c \\ p_u \end{bmatrix}, \qquad g_s = \begin{bmatrix} s \\ 1-s \end{bmatrix}, \tag{1}$$

where s is the guidance scale.

At a given denoising step and activation-quantization site, let $X_c, X_u \in \mathbb{R}^{N \times d}$ denote the matched conditional and unconditional branch activations, where N indexes token or spatial positions and d is the feature dimension. We stack them along the CFG branch axis as

$$X = \begin{bmatrix} X_c \\ X_u \end{bmatrix} \in \mathbb{R}^{2 \times N \times d}, \tag{2}$$

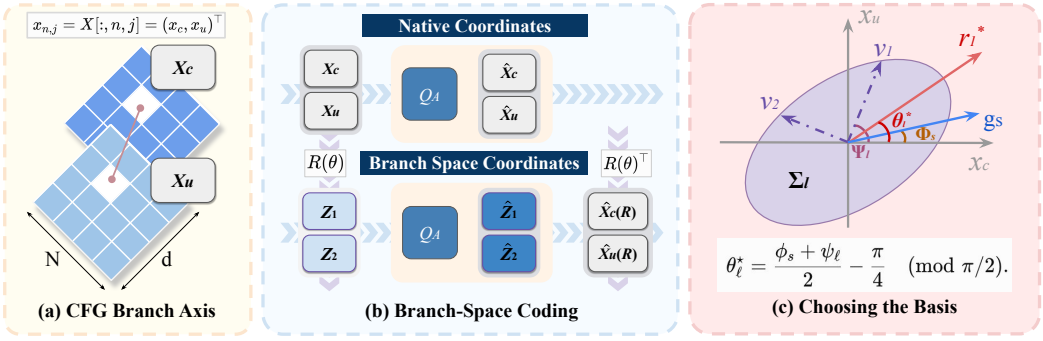


Figure 3: **Overview of branch-space coding. (a)** Matched CFG activations define a two-dimensional branch axis at each token–feature coordinate. **(b)** Native quantizer Q_A operates directly on (X_c, X_u), while branch-space coding rotates the pair, applies the same host quantizer in transformed coordinates, and rotates back. **(c)** The coding basis is determined jointly by the activation geometry Σ_ℓ and CFG branch direction g_s, yielding the optimal coding axis $r_1^\star$ and angle $\theta_\ell^\star$ where $\theta_\ell^\star = \angle(X_c, r_1^\star)$.

whose leading dimension defines the CFG branch space with native coordinates (c, u). At each token–feature coordinate (n, j), the corresponding matched branch-activation pair is $x_{n,j} = [x_{c,n,j}, x_{u,n,j}]^\top \in \mathbb{R}^2$.

Let $\mathcal{Q}_A$ denote the host A-bit activation quantizer. Standard PTQ quantizes the two branch activations in the native basis: $\widehat{X}_{\text{nat}} = \begin{bmatrix} \mathcal{Q}_A(X_c) \\ \mathcal{Q}_A(X_u) \end{bmatrix}$.

3.2 Branch-Space Structure and Transform Coding

Standard diffusion PTQ (Chen et al., 2025; Li et al., 2023; 2025; 2026; Lee et al., 2026; Zhao et al., 2025) quantizes CFG activations directly in the native (c, u) branch coordinates. To determine whether activation quantization can benefit from a different branch coding basis, we examine whether matched CFG activations exhibit cross-branch structure and whether changing only the branch coding basis affects quantization fidelity under a fixed activation-bit budget.

Observation I: ***CFG branches form a structured matched source.*** *On DiT-XL/2 with W4A4 SVDQuant (CFG $= 2$, 64 trajectories), the median correlation between matched branch activations across all linear sites in DiT-XL/2 is 0.859, compared with 0.328 for deliberately mismatched pairs (Figure 2, Table 6). The matched activations are also highly directional: their dominant axis lies at a median 42.71° from the conditional axis, close to the diagonal $c = u$. Figure 3(a) illustrates this structure schematically. These measurements motivate considering the matched activations jointly when choosing a quantization basis.*

Observation II: ***Activation-quantization fidelity depends on the branch basis.*** *Under the same activation-bit budget and host activation quantizer, changing only the branch coding basis produces measurable differences in generation fidelity. At CFG $= 4.5$, a guidance-aligned rotation improves PSNR and SSIM while reducing LPIPS under the same host quantizer and activation-bit budget; see Table 4.*

These observations motivate treating the branch coding basis as an explicit degree of freedom in activation quantization, rather than inheriting the native (c, u) execution coordinates. This leads to the following general formulation.

Definition 3.1 (Branch-space transform coding)**.** *Let $R \in \mathbb{R}^{2\times 2}$ be an orthogonal matrix, whose rows define an orthonormal coding basis. Given a host activation quantizer $\mathcal{Q}_A$, we quantize matched CFG activations in the transformed branch basis as*

$$\widehat{X}(R) := R^\top \mathcal{Q}_A(RX) = \begin{bmatrix} \widehat{X}_c(R) \\ \widehat{X}_u(R) \end{bmatrix}. \tag{3}$$

The transform acts only across matched conditional/unconditional pairs, leaving token and feature coordinates unchanged. Setting $R = I$ recovers native branchwise activation quantization.

3.3 GUIDANCE-CORRELATION BRANCH TRANSFORM

Branch-space coding leaves one question: *which basis should be used?* We choose the basis jointly from the geometry of matched CFG activations and the CFG branch direction.

At layer ℓ, for each matched token–feature coordinate (n, j), define

$$x_{n,j}^{(\ell)} = \begin{bmatrix} X_c^{(\ell)}[n,j] \\ X_u^{(\ell)}[n,j] \end{bmatrix} \in \mathbb{R}^2. \tag{4}$$

Suppressing (n, j) for brevity, we write a generic matched branch-activation pair as $x^{(\ell)} = [x_c^{(\ell)}, x_u^{(\ell)}]^\top$. The branch-space geometry at layer ℓ is summarized by

$$\Sigma_\ell = \mathbb{E}\left[x^{(\ell)} x^{(\ell)\top}\right], \tag{5}$$

where the expectation is taken over calibrated matched branch-activation pairs. Thus, Σ_ℓ is the uncentered second-moment matrix describing the two-dimensional branch geometry at layer ℓ.

For the remainder of this derivation, we fix layer ℓ and suppress the layer superscript on local branch-space quantities. Let $R = [r_1^\top; r_2^\top] \in \mathbb{R}^{2\times 2}$ be an orthogonal rotation matrix, where $r_1, r_2 \in \mathbb{R}^2$ define the coding basis. For a matched pair x, the transformed coordinates are $z = Rx$, with $z_i = r_i^\top x$.

DESIGN PRINCIPLE: ***Couple activation geometry with guidance sensitivity.*** Let $\hat{z}_i$ denote the quantized value of transformed coordinate z_i. We approximate the quantization distortion of each transformed coordinate as proportional to its projected second moment,

$$\mathbb{E}\left[(\hat{z}_i - z_i)^2\right] = \kappa_\ell \, r_i^\top \Sigma_\ell r_i, \qquad \kappa_\ell > 0, \tag{6}$$

where κ_ℓ is shared by the two coordinates, and cross-coordinate quantization-error terms are neglected.

Proposition 3.2 (Guidance-weighted distortion)**.** *Under this approximation, the expected squared branch-space distortion along the CFG direction is proportional to*

$$D_{\mathrm{sur}}(R) = \sum_{i=1}^{2} \underbrace{(r_i^\top g_s)^2}_{\substack{\text{guidance} \\ \text{sensitivity}}} \underbrace{(r_i^\top \Sigma_\ell r_i)}_{\substack{\text{projected} \\ \text{second moment}}}. \tag{7}$$

The objective therefore favors a branch basis that jointly accounts for how strongly each coding direction is represented in the activations and how strongly it contributes to the CFG.

Because the branch space is two-dimensional, the rotation can be parameterized by a single angle θ. We measure all angles relative to the native conditional axis and write

$$R(\theta) = \begin{bmatrix} r_1(\theta)^\top \\ r_2(\theta)^\top \end{bmatrix}, \qquad r_1(\theta) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \quad r_2(\theta) = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}. \tag{8}$$

Let ψ_ℓ denote the angle of the dominant eigenvector of Σ_ℓ, and let $\phi_s = \arctan2(1-s, s)$ denote the angle of the CFG direction g_s.

Theorem 3.3 (Closed-form optimal branch rotation)**.** *Let $\lambda_{1,\ell} > \lambda_{2,\ell}$ be the eigenvalues of Σ_ℓ. The global minimizers of Eq. 7 over two-dimensional orthogonal rotations are*

$$\theta_\ell^\star(s) = \frac{\phi_s + \psi_\ell}{2} - \frac{\pi}{4} \qquad \left(\mathrm{mod}\ \frac{\pi}{2}\right). \tag{9}$$

If $\lambda_{1,\ell} = \lambda_{2,\ell}$, the surrogate is rotationally invariant.

We call $R(\theta_\ell^\star)$ the ***Guidance-Correlation Branch Transform*** **(GCBT).** A guidance-aligned basis depends only on g_s, while the principal-axis basis of Σ_ℓ depends only on activation geometry. GCBT combines both quantities in a single coding basis. Calibration requires only the three distinct entries of the 2×2 matrix Σ_ℓ per layer; no gradient optimization or angle search is required. The derivation and proof are given in Appendix A.1 and A.2.

3.4 Compatibility with Rotation-Based and CFG-Aware PTQ Backbones

Rotation-based Quantizer Compatibility. Some diffusion PTQ backbones apply a shared feature-space rotation or linear transform before activation quantization. Let $U_\ell \in \mathbb{R}^{d\times d}$ denote such a transform at layer ℓ, shared by the conditional and unconditional branches, and let $R_\ell \in \mathbb{R}^{2\times 2}$ denote the GCBT branch transform. Because U_ℓ acts on the feature dimension whereas R_ℓ acts on the CFG branch dimension, the two operations commute exactly.

Proposition 3.4 (Exact composition with shared feature-space rotations)**.** *For branch activations $X \in \mathbb{R}^{2\times N\times d}$, a shared feature-space transform $U_\ell \in \mathbb{R}^{d\times d}$, and a branch-space transform $R_\ell \in \mathbb{R}^{2\times 2}$,*

$$R_\ell(XU_\ell) = (R_\ell X)U_\ell. \tag{10}$$

Consequently, a host activation quantizer that operates on XU_ℓ can be combined with GCBT without changing the transformed activation seen by the quantizer:

$$\mathcal{Q}_A(R_\ell(XU_\ell)) = \mathcal{Q}_A((R_\ell X)U_\ell)\,. \tag{11}$$

Thus, GCBT is algebraically compatible with host PTQ methods that already use shared feature-space rotations: branch-space and feature-space transforms are strictly composable rather than competing for the same degree of freedom.

CFG-aware calibration hosts. CFG-aware calibration and GCBT act on distinct quantization variables: a guidance-aware host optimizes quantization parameters against guided-output error while retaining the native (c, u) coordinates, whereas GCBT changes the branch basis in which quantization is performed. For a matched branch-activation pair x with $z = R_\ell x$,

$$g_s^\top x = g_s^\top R_\ell^\top z = (R_\ell g_s)^\top z. \tag{12}$$

Thus, the CFG branch information is preserved exactly under the basis change, so guided-output calibration does not determine the branch coding basis.

We test this complementarity by applying GCBT to a fixed CFG-aware W4A4 PixArt-Σ host on MJHQ and sDCI, holding the calibrated host parameters, quantization configuration, and generation seeds fixed. Table 1 shows consistent LPIPS and SSIM gains on both datasets, while the smaller yet positive PSNR changes are not statistically conclusive. The gains therefore persist after guided-output calibration, supporting branch-basis selection as a complementary quantization variable. Full calibration details and paired confidence intervals are reported in Appendix B.2.

Table 1: **GCBT after CFG-aware calibration.** PixArt-Σ, W4A4, we implement (Shafi & Suma, 2026) on SVDQuant given GAMP is closed-source. Bold: 95% CI excludes zero.

Data	Method	LPIPS ↓	SSIM ↑	PSNR ↑
MJHQ	CFG-Cal.	0.4775	0.5722	16.454
	+GCBT	**0.4634**	**0.5829**	16.547
sDCI	CFG-Cal.	0.5310	0.4768	15.279
	+GCBT	**0.5179**	**0.4872**	15.352

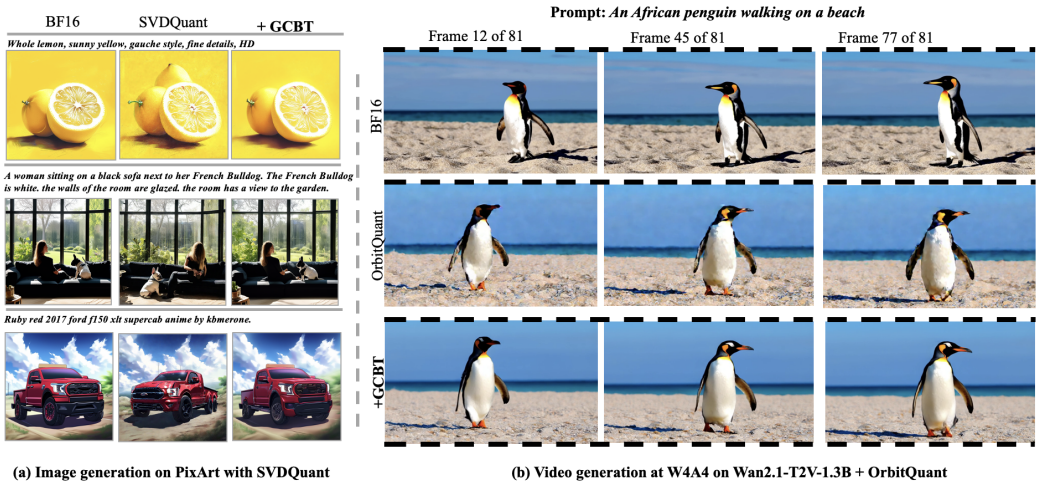
4 Experiments

4.1 Experimental Settings

Models and PTQ hosts. We evaluate GCBT across image and video diffusion models with multiple activation-quantization hosts. For image generation, the main evaluation includes PixArt-Σ (Chen et al., 2024) and SANA-1.6B (Xie et al., 2024) with SVDQuant (Li et al., 2025) (W4A4) and Q-DiT (Chen et al., 2025) (W4A8), with SD3.5-Large (Esser et al., 2024) included for large-scale validation. All host precision settings strictly follow the default evaluation protocols established in their respective original works. For video generation, we evaluate Wan2.1-T2V-1.3B and Wan2.1-T2V-14B (Wan et al., 2025), CogVideoX-2B (Yang et al., 2025), and OpenSora v1.2 (Zheng et al., 2024) with OrbitQuant (Lee et al., 2026) (W4A4) or ViDiT-Q (Zhao et al., 2025) (W4A8-MP). GCBT uses 64 calibration samples disjoint from evaluation. Full model-specific generation and calibration settings are provided in Appendix B.4.

Table 2: **Quantitative evaluation of image-generation fidelity on MJHQ and sDCI** ($n = 256$)**.** For each +GCBT row, the second line reports the pointwise paired 95% confidence interval of improvement (Native minus GCBT for LPIPS; GCBT minus Native for SSIM and PSNR). Bold denotes gains whose paired interval lies entirely above zero.

PTQ Hosts	Model	Method	MJHQ			sDCI		
			LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
SVDQuant W4A4	PixArt-Σ CFG 4.5	Native	0.327	0.651	17.472	0.355	0.574	16.450
		+ GCBT	**0.296** [0.024, 0.038]	**0.695** [0.039, 0.050]	**17.917** [0.287, 0.607]	**0.319** [0.029, 0.043]	**0.625** [0.047, 0.057]	**17.017** [0.418, 0.708]
	SANA-1.6B CFG 4.5	Native	0.230	0.702	17.283	0.234	0.651	16.565
		+ GCBT	**0.210** [0.015, 0.026]	**0.730** [0.023, 0.033]	**17.825** [0.383, 0.703]	**0.213** [0.016, 0.025]	**0.685** [0.029, 0.038]	**17.051** [0.364, 0.608]
Q-DiT W4A8	PixArt-Σ CFG 4.5	Native	0.269	0.711	17.939	0.281	0.648	17.135
		+ GCBT	**0.266** [0.001, 0.005]	**0.715** [0.003, 0.006]	**18.009** [0.035, 0.106]	**0.278** [0.001, 0.005]	**0.654** [0.004, 0.007]	**17.192** [0.019, 0.094]
	SANA-1.6B CFG 4.5	Native	0.175	0.764	18.907	0.178	0.722	17.941
		+ GCBT	0.174 [−0.002, 0.004]	0.765 [−0.002, 0.003]	18.877 [−0.120, 0.059]	0.179 [−0.002, 0.002]	0.722 [−0.002, 0.002]	17.928 [−0.089, 0.057]
Large-scale validation								
SVDQuant W4A4	SD3.5-Large CFG 4.5	Native	0.394	0.674	16.242	0.386	0.653	15.898
		+ GCBT	**0.378** [0.011, 0.021]	**0.690** [0.013, 0.019]	**16.385** [0.042, 0.243]	**0.371** [0.010, 0.020]	**0.667** [0.011, 0.018]	**16.018** [0.031, 0.211]



(a) Image generation on PixArt with SVDQuant

(b) Video generation at W4A4 on Wan2.1-T2V-1.3B + OrbitQuant

Figure 4: Qualitative comparison of GCBT on image and video generation.

Evaluation. We measure FP-reference fidelity using LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR (Huynh-Thu & Ghanbari, 2008). For video, we additionally report VBench subject consistency (S.C.), aesthetic quality (A.Q.), and imaging quality (I.Q.) (Huang et al., 2024). Evaluation uses fixed paired prompt sets ($n = 256$), and statistical significance is determined from paired 95% confidence intervals; full effect sizes and intervals are reported in the Appendix B.4.

4.2 Main Results

Image generation. Across the full image-generation matrix, GCBT yields statistically significant gains in most paired metric comparisons, while no comparison shows a statistically significant degradation. All statistically resolved effects therefore favor GCBT. The significant gains span both SVDQuant models, Q-DiT on PixArt-Σ, and the larger SD3.5-Large model, consistently across MJHQ and sDCI. The remaining comparisons are statistically inconclusive, with no evidence of

Table 3: **Quantitative evaluation of video generation on VBench ($n = 256$).** Bold indicates statistically significant improvements (paired 95% CIs lie entirely in the favorable direction). Full paired confidence intervals are provided in Appendix Table 9.

PTQ Host	Model	Method	Teacher Fidelity			VBench		
			LPIPS ↓	SSIM ↑	PSNR ↑	S.C. ↑	A.Q. ↑	I.Q. ↑
OrbitQuant W4A4	Wan2.1-T2V-1.3B CFG 5	Native	0.545	0.497	12.871	0.957	0.527	0.672
		+ GCBT	**0.504**	**0.518**	**12.965**	**0.975**	**0.584**	0.675
	CogVideoX-2B CFG 6	Native	0.427	0.639	17.474	0.927	0.494	0.593
		+ GCBT	**0.390**	**0.649**	17.554	**0.937**	**0.525**	**0.606**
ViDiT-Q W4A8-MP	OpenSora v1.2 CFG 7	Native	0.544	0.540	13.871	0.935	0.455	0.490
		+ GCBT	**0.541**	**0.542**	13.891	**0.938**	**0.459**	0.493
Large-scale validation								
OrbitQuant port W4A4	Wan2.1-T2V-14B CFG 5	Native	0.431	0.590	15.021	0.971	0.588	0.704
		+ GCBT	**0.413**	**0.601**	15.029	**0.976**	**0.607**	0.703

a systematic regression. Overall, these results show that the benefits of GCBT transfer across architectures, PTQ hosts, datasets, and model scales.

Video generation. The video experiments extend this pattern beyond image generation and, more importantly, beyond teacher-referenced reconstruction alone. Across Wan2.1, CogVideoX, and OpenSora, GCBT improves both FP-referenced fidelity and VBench generation-quality metrics. LPIPS and SSIM improve significantly for every evaluated video model, while subject consistency and aesthetic quality also improve significantly throughout. The effect persists on the larger Wan2.1-T2V-14B model, showing that the branch-space benefit transfers across video architectures, PTQ hosts, and model scales. These results indicate that the fidelity gains from GCBT are accompanied by improvements in generation-level quality, rather than being confined to closer reconstruction of the full-precision reference.

4.3 Ablations and System Analysis

We isolate the design choices underlying GCBT and examine its robustness and systems cost. Unless stated otherwise, ablations use PixArt-Σ with SVDQuant W4A4 on the fixed $n = 256$ MJHQ subset at CFG $= 4.5$, with 64 calibration samples disjoint from evaluation.

Branch-basis components. GCBT selects the coding basis jointly from the branch geometry Σ_ℓ and the CFG readout direction g_s. We first compare the native basis, a guidance-aligned basis using only g_s, and GCBT using both quantities under the same quantization and generation settings.

Table 4 shows a clear progression: aligning the basis with the CFG readout already improves over native branch coding, while incorporating the measured branch geometry yields a further gain across all three fidelity metrics. This indicates that guidance information alone does not account for the benefit of GCBT; the coding basis benefits from coupling CFG sensitivity with the geometry of the matched activations. Full statistics are reported in Appendix B.3.

Table 4: **Branch-basis ablation.** PixArt-Σ, SVDQuant W4A4 on MJHQ.

Basis	LPIPS ↓	SSIM ↑	PSNR ↑
Native	0.327	0.651	17.472
Guidance	0.315	0.678	17.707
GCBT	**0.296**	**0.695**	**17.917**

Adaptation across guidance scales. Because GCBT optimizes its basis using the CFG branch direction g_s, we evaluate its behavior across varying guidance scales while holding all other configurations fixed. As illustrated in Figure 5, GCBT exhibits a clear scale-dependent pattern: its fidelity gains over the host quantizer grow monotonically as guidance becomes stronger. At weak guidance ($s = 2$), the perceptual improvement is marginal; however, from $s = 3$ onward, GCBT consistently strengthens performance across LPIPS, SSIM, and PSNR. This trend directly reflects the role of g_s in

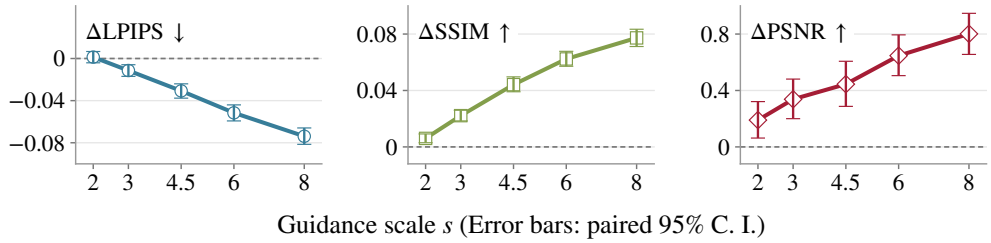


Figure 5: **Fidelity gains across CFG scales.** GCBT achieves monotonically larger gains over native PTQ across all metrics as guidance scale s increases.

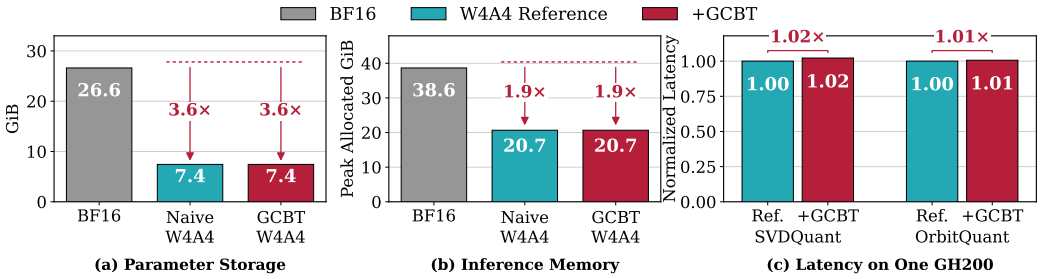


Figure 6: **System overhead comparison.** Host+GCBT matches native engines in end-to-end latency and peak GPU memory across image and video pipelines, imposing minimal practical overhead.

our objective function: as stronger CFG alters the guidance readout orientation, adapting the branch coding basis becomes increasingly consequential for preserving activation fidelity.

Runtime and memory overhead. To evaluate GCBT's practical system cost, we measure end-to-end generation latency and peak GPU memory footprint on representative SVDQuant and OrbitQuant deployments, with SVDQuant evaluated using the Nunchaku inference engine (offline calibration excluded). As shown in Figure 6, adding GCBT leaves the memory footprint unchanged and introduces only negligible runtime overhead (1–2%) relative to the corresponding host implementations. Thus, the fidelity gains of GCBT do not require a materially more expensive inference pipeline.

Calibration Sample Stability. GCBT depends only on the 2×2 branch second-moment matrix at each layer, so the required branch geometry can be estimated from a modest measurement pool. As shown in Table 5 (full results in Table 6), the measured geometry is already stable by $n_{\text{geom}} = 64$: increasing the pool to 128 changes the median principal-axis orientation by only $0.04°$, while the matched-branch correlation remains approximately 0.86 and the source remains strongly anisotropic. We therefore use 64 samples for GCBT calibration throughout. Full results are reported in Appendix B.1.

Table 5: **Calibration stability.** Layer-wise medians.

n_{geom}	Matched corr.	ψ_ℓ	$\lambda_{1,\ell}/\lambda_{2,\ell}$
16	0.864	42.89°	13.71
64	0.859	42.71°	13.27
128	0.857	42.67°	13.03

5 Conclusion

We identify CFG branch space as an overlooked coding dimension for diffusion activation quantization and introduce GCBT, a closed-form transform that jointly exploits matched-branch geometry and guidance sensitivity. Across many settings, GCBT consistently yields favorable fidelity and generation-quality improvements without modifying model parameters or the underlying quantization pipeline. These results establish branch-space coding as a lightweight and complementary design axis for diffusion PTQ, with negligible runtime and memory overhead.

AI USE STATEMENT

The core research idea, methodological formulation, and experimental design of this work were conceived and developed by the authors. Generative AI tools were subsequently used as secondary aids to critique and refine experimental plans, assist with research implementation and troubleshooting, and improve manuscript organization, wording, and figure presentation. All experiments were specified and executed by the authors, and all reported results, statistical analyses, mathematical derivations, and scientific claims were manually inspected and verified against the underlying experimental outputs and original sources. AI-assisted material was not accepted as final scientific content without substantive author review and revision. The manuscript itself was written and iteratively revised by the authors, with generative AI used as an editing and refinement tool rather than as a substitute for scientific authorship. The authors take full responsibility for the correctness, reproducibility, and final presentation of all content in this work.

ETHICS STATEMENT

This work does not involve human subjects, private user data, or the collection of sensitive information. We study post-training quantization of existing generative models using publicly released models and standard evaluation benchmarks. GCBT modifies only the coordinate basis used for activation quantization and does not introduce new generative capabilities. More efficient inference can nevertheless reduce the cost of deploying generative models and may therefore amplify both the benefits and existing risks of the underlying models, including inherited biases or potential misuse. Such risks remain governed by the capabilities, licenses, and usage policies of the corresponding host models.

REPRODUCIBILITY STATEMENT

We provide the mathematical formulation of GCBT and all assumptions underlying the distortion surrogate in the main text, with complete derivations and proofs in Appendix A. The appendix additionally reports calibration-size studies, paired statistical analyses for the main ablations and CFG-aware calibration, and the full model-specific generation and evaluation protocol. All primary comparisons use fixed paired prompt sets and generation seeds, with identical host configurations between the native and GCBT runs; calibration data are disjoint from evaluation data. We report the PTQ host, precision, CFG setting, calibration size, evaluation metrics, and paired confidence-interval procedure needed to reproduce the reported comparisons.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28306–28315. IEEE, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Zhihang Yuan, Jiangyong Yu, Chen Xu, Zhe Jiang, and Sifan Zhou. Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 37492–37517, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5cebc89b113920dbff7c79854ba765a3-Paper-Conference.pdf.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801, 2008. doi: 10.1049/el:20080522. URL https://digital-library.theiet.org/doi/abs/10.1049/el%3A20080522.

Donghyun Lee, Jitesh Chavan, Duy Nguyen, Sam Huang, Liming Jiang, Priyadarshini Panda, Timo Mertens, and Saurabh Shukla. Orbitquant: Data-agnostic quantization for image and video diffusion transformers, 2026. URL https://arxiv.org/abs/2607.02461.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank component for 4-bit diffusion models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 97831–97858, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f34f0630c33be15b8c89426bb8056798-Paper-Conference.pdf.

Xingyang Li, Samuel Tesfai, Zhekai Zhang, Haocheng Xi, Shuo Yang, Lvmin Zhang, Yufei Sun, Kelly Peng, Maneesh Agrawala, Ion Stoica, et al. Deltaquant: 4-bit video diffusion models with spatiotemporal delta smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43578–43588, 2026.

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.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Abdullah Al Shafi and Sumaiya Rahim Suma. Closing the null space: Guidance-aware quantization for classifier-free diffusion. *arXiv preprint arXiv:2607.08241*, 2026.

Yihua Shao, Deyang Lin, Minxi Yan, Siyu Chen, Fanhu Zeng, Minwen Liao, Ao Ma, Ziyang Yan, Haozhe Wang, Yan Wang, et al. Tr-dq: Time-rotation diffusion quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8869–8877, 2026.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57587–57613. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/sun25l.html.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *Proceedings of machine learning research*, 235:48630, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024. URL https://arxiv.org/abs/2410.10629.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Enshu Liu, Rui Wan, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E1N1oxd63b.

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

CONTENTS

1 Introduction **2**

2 Related Work **3**

3 Method **3**

3.1 Problem Setup 3

3.2 Branch-Space Structure and Transform Coding 4

3.3 Guidance-Correlation Branch Transform 5

3.4 Compatibility with Rotation-Based and CFG-Aware PTQ Backbones 6

4 Experiments **6**

4.1 Experimental Settings 6

4.2 Main Results 7

4.3 Ablations and System Analysis 8

5 Conclusion **9**

A Additional Details for GCBT **14**

A.1 Derivation of the Guidance-Weighted Distortion 14

A.2 Proof of Theorem 3.3 14

A.3 Commutation with Shared Linear Operators 15

B Additional Experiments **15**

B.1 Calibration-Size Stability 15

B.2 Paired Statistics for CFG-Aware Calibration 15

B.3 Additional Statistics for Branch-Basis Ablation 16

B.4 Full Experimental Protocol & Details 16

C Limitations **18**

Appendix

A Additional Details for GCBT

A.1 Derivation of the Guidance-Weighted Distortion

Fix a layer ℓ and suppress the layer superscript on local branch-space quantities. Let $R = \begin{bmatrix} r_1^\top \\ r_2^\top \end{bmatrix} \in \mathbb{R}^{2\times 2}$ be an orthogonal branch rotation, and let $z = Rx$ denote the transformed coordinates of a matched branch-activation pair x. Define the transformed-coordinate quantization error as $\delta = \widehat{z} - z = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$, with $\delta_i = \widehat{z}_i - z_i$.

Following the distortion surrogate in Eq. 6, we assume

$$\mathbb{E}[\delta_i^2] = \kappa_\ell \, r_i^\top \Sigma_\ell r_i, \qquad \mathbb{E}[\delta_1 \delta_2] = 0, \qquad \kappa_\ell > 0, \tag{13}$$

where the same κ_ℓ is used for both transformed coordinates. No zero-mean assumption on the quantization error is required.

After inversion of the branch rotation, the local branch-space error is $R^\top \delta$. Its component along the CFG readout direction g_s is

$$g_s^\top R^\top \delta = \sum_{i=1}^{2} (r_i^\top g_s)\delta_i. \tag{14}$$

Therefore,

$$\mathbb{E}\left[\left(g_s^\top R^\top \delta\right)^2\right] = \sum_{i=1}^{2} (r_i^\top g_s)^2 \mathbb{E}[\delta_i^2] + 2(r_1^\top g_s)(r_2^\top g_s)\mathbb{E}[\delta_1\delta_2] \tag{15}$$

$$= \kappa_\ell \sum_{i=1}^{2} (r_i^\top g_s)^2 (r_i^\top \Sigma_\ell r_i). \tag{16}$$

Because $\kappa_\ell > 0$ is independent of the branch basis, minimizing this expected squared distortion is equivalent to minimizing $D_{\mathrm{sur}}(R)$ in Eq. 7.

A.2 Proof of Theorem 3.3

Because the branch space is two-dimensional, any rotation can be parameterized by a single angle θ. Measuring angles relative to the native conditional axis, write

$$R(\theta) = \begin{bmatrix} r_1(\theta)^\top \\ r_2(\theta)^\top \end{bmatrix}, \qquad r_1(\theta) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \qquad r_2(\theta) = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}. \tag{17}$$

Let

$$\bar{\lambda}_\ell = \frac{\lambda_{1,\ell} + \lambda_{2,\ell}}{2}, \qquad \Delta_\ell = \frac{\lambda_{1,\ell} - \lambda_{2,\ell}}{2}, \tag{18}$$

and let ψ_ℓ denote the angle of the dominant eigenvector of Σ_ℓ and let ϕ_s denote the angle of g_s, both measured relative to the native conditional axis. Then

$$r_1^\top \Sigma_\ell r_1 = \bar{\lambda}_\ell + \Delta_\ell \cos 2(\theta - \psi_\ell), \tag{19}$$

$$r_2^\top \Sigma_\ell r_2 = \bar{\lambda}_\ell - \Delta_\ell \cos 2(\theta - \psi_\ell). \tag{20}$$

Similarly,

$$r_1^\top g_s = \|g_s\|_2 \cos(\phi_s - \theta), \tag{21}$$

$$r_2^\top g_s = \|g_s\|_2 \sin(\phi_s - \theta). \tag{22}$$

Substituting these expressions into Eq. 7 gives

$$\frac{D_{\mathrm{sur}}(\theta)}{\|g_s\|_2^2} \propto \bar{\lambda}_\ell + \Delta_\ell \cos 2(\phi_s - \theta) \cos 2(\theta - \psi_\ell) \tag{23}$$

$$= C_\ell + \frac{\Delta_\ell}{2} \cos(4\theta - 2\phi_s - 2\psi_\ell), \tag{24}$$

where

$$C_\ell = \bar{\lambda}_\ell + \frac{\Delta_\ell}{2} \cos 2(\phi_s - \psi_\ell)$$

is independent of θ.

For $\Delta_\ell > 0$, the minimum is attained when

$$4\theta - 2\phi_s - 2\psi_\ell = -\pi \qquad (\mathrm{mod}\, 2\pi), \tag{25}$$

which gives

$$\theta_\ell^\star(s) = \frac{\phi_s + \psi_\ell}{2} - \frac{\pi}{4} \qquad \left(\mathrm{mod}\, \frac{\pi}{2}\right). \tag{26}$$

This is Eq. 9. For $\Delta_\ell = 0$, equivalently $\lambda_{1,\ell} = \lambda_{2,\ell}$, the objective is independent of θ, and the surrogate is rotationally invariant.

A.3 Commutation with Shared Linear Operators

Let $\mathcal{L}(X) = XW$ be a linear operator shared by the two CFG branches. Since R acts along the branch dimension while W acts along the feature dimension,

$$\mathcal{L}(R^\top Z) = R^\top \mathcal{L}(Z). \tag{27}$$

Let $\widehat{Z} = \mathcal{Q}_A(RX)$ denote the quantized activation in transformed branch coordinates. Equation 27 allows the shared linear operator to act directly on $\widehat{Z}$,

$$\widehat{Z} \xrightarrow{W} Y' \xrightarrow{R^\top} Y, \tag{28}$$

rather than first reconstructing the quantized activation in the native branch basis.

For an affine operator with a bias shared by the two CFG branches, the bias is added after the inverse branch transform. Thus, branch-space coding does not require reconstruction of the full native-basis activation before the shared quantized linear operation.

B Additional Experiments

B.1 Calibration-Size Stability

GCBT estimates the three distinct entries of the 2×2 second-moment matrix Σ_ℓ at each layer. To assess how much calibration data is required for these statistics to stabilize, we vary the calibration-set size on DiT-XL/2 with SVDQuant W4A4 and report layerwise median branch-space statistics. The mismatched correlation is computed by deliberately pairing conditional and unconditional activations from different samples, providing a reference for the matched-branch structure.

The estimated branch geometry is already stable at modest calibration sizes. In particular, beyond $n_{\mathrm{cal}} = 64$, the principal-axis orientation changes negligibly, while matched-branch correlation remains close to 0.855 and consistently exceeds the mismatched reference. The source also remains strongly anisotropic throughout the sweep. We therefore use 64 calibration samples in the main experiments.

B.2 Paired Statistics for CFG-Aware Calibration

Setup. GAMP (Shafi & Suma, 2026) has no public implementation, so we implement a GAMP-inspired guided-output calibration on the official SVDQuant W4A4 PixArt-Σ host. This is not an official GAMP reproduction. The W4 GPTQ residual weights, smoothing, and FP rank-32 branch are

Table 6: **Measurement-pool-size stability of CFG branch geometry on DiT-XL/2.** Values are medians across 112 linear sites under repository-native SVDQuant W4A4 with an FP rank-32 branch. All sizes use nested class/noise inputs and frozen weights.

n_{geom}	Matched corr.	Mismatched corr.	ψ_ℓ	$\lambda_{1,\ell}/\lambda_{2,\ell}$
16	0.8635	0.3166	42.89°	13.71
32	0.8604	0.3243	42.71°	13.46
64	**0.8587**	**0.3276**	$\mathbf{42.71^\circ}$	**13.27**
128	0.8569	0.3231	42.67°	13.03

Table 7: **Paired effects of GCBT after CFG-aware calibration.** Positive values favor GCBT. LPIPS reports relative reduction, while SSIM and PSNR report absolute paired improvements. Brackets denote paired 95% confidence intervals.

Dataset	LPIPS reduction ↑	ΔSSIM ↑	ΔPSNR (dB) ↑
MJHQ	2.97% [2.17, 3.78]%	+0.0107 [0.0079, 0.0135]	+0.093 [−0.005, 0.190]
sDCI	2.47% [1.71, 3.23]%	+0.0104 [0.0077, 0.0132]	+0.073 [−0.017, 0.169]

frozen. For each of the 168 activation quantizers (dynamic group-64 A4), we fit one positive clipping multiplier in $[0.1, 2.0]$ by minimizing the MSE between the quantized and full-precision guided noise predictions, $\epsilon_u + s(\epsilon_c - \epsilon_u)$ with $s = 4.5$. We use 1,024 paired states from 64 training prompts and 256 states from 16 validation prompts, 800 Adam updates (learning rate 0.01, batch size one), and keep the checkpoint with the lowest validation error. The calibrated host is then frozen, and GCBT angles are computed from 64 complete Native trajectories on this host. Host-training, validation, GCBT-calibration, and evaluation prompts are disjoint.

This study uses a separate execution setup from Table 2: BF16 execution and a BF16 teacher, with initial noise drawn from a CPU generator. Evaluation uses the same 256 MJHQ and 256 sDCI prompts and seeds (1024×1024, 20 DPM-Solver++ steps), but the different dtype and noise generator yield different teacher images. Absolute values are therefore not comparable across the two tables; the evidence is the paired Host versus +GCBT comparison under an identical host, prompts, seeds, and teacher.

As shown in Table 7, the paired intervals are entirely favorable for LPIPS and SSIM on both datasets, whereas the PSNR intervals overlap zero. These statistics provide the paired uncertainty estimates underlying the significance annotations in Table 1.

B.3 Additional Statistics for Branch-Basis Ablation

Table 8 provides the statistical details underlying the branch-basis ablation in Table 4. All variants are evaluated on the same 256 MJHQ prompts with identical generation seeds, enabling paired prompt-level comparisons. We report bootstrap 95% confidence intervals from 10,000 prompt-level resamples for both the absolute metric means and the paired improvements between basis choices.

The paired analysis supports the stepwise pattern observed in the main text. Guidance alignment significantly improves over Native in LPIPS, SSIM, and PSNR, and GCBT provides a further statistically supported improvement over the guidance-aligned basis on all three metrics. Consequently, the full GCBT–Native comparison is also favorable throughout. We base these comparisons on the paired intervals in Table 8(b), rather than overlap between the marginal confidence intervals of the absolute means in panel (a).

B.4 Full Experimental Protocol & Details

Image generation. PixArt-Σ and SANA-1.6B are evaluated at 1024×1024 resolution with 20 sampling steps and CFG $= 4.5$. Both use the second-order DPM-Solver++ scheduler packaged with their respective checkpoints; PixArt-Σ uses epsilon prediction, whereas SANA uses flow prediction. SD3.5-Large is evaluated at 1024×1024 resolution with 28 steps and CFG $= 4.5$, using

Table 8: **Additional statistics for branch-basis ablation.** Evaluated on PixArt-Σ (SVDQuant W4A4, CFG $s = 4.5$) on MJHQ ($n = 256$). Panel (a) reports absolute metric means with 95% CIs. Panel (b) reports paired improvements over the reference (positive values indicate improvement).

(a) Absolute metrics: mean [95% CI]

Basis	LPIPS ↓	SSIM ↑	PSNR ↑
Native	0.327 [0.315, 0.338]	0.651 [0.638, 0.664]	17.472 [17.150, 17.805]
Guidance	0.315 [0.303, 0.327]	0.678 [0.664, 0.691]	17.707 [17.378, 18.044]
GCBT	**0.296** [0.284, 0.308]	**0.695** [0.682, 0.708]	**17.917** [17.584, 18.256]

(b) Paired improvements: mean [95% CI]

Candidate vs. reference	LPIPS reduction ↑	SSIM gain ↑	PSNR gain ↑
Guidance vs. Native	+0.012 [0.004, 0.020]	+0.027 [0.021, 0.032]	+0.236 [0.069, 0.406]
GCBT vs. Native	+0.031 [0.024, 0.038]	+0.044 [0.039, 0.050]	+0.445 [0.287, 0.607]
GCBT vs. Guidance	+0.019 [0.014, 0.024]	+0.017 [0.014, 0.020]	+0.209 [0.104, 0.307]

its checkpoint-provided FlowMatch Euler scheduler with shift = 3. We use the official non-distilled checkpoints and fixed, empty negative prompts. Within each model, the full-precision reference, native quantized host, and GCBT use identical prompts, generation seeds, samplers, and generation settings.

Video generation. Wan2.1-T2V-1.3B and Wan2.1-T2V-14B generate 832×480 videos with 81 frames at 16 FPS, using 50 FlowUniPC steps and CFG = 5; their flow shifts are 8 and 5, respectively. CogVideoX-2B generates 720×480 videos with 49 frames at 8 FPS, using 50 steps of its checkpoint-provided CogVideoX DDIM scheduler and CFG = 6, with dynamic CFG disabled. OpenSora v1.2 generates 512×512 videos with 64 frames at 24 FPS, using 30 rectified-flow steps, timestep transformation, and CFG = 7; its text preprocessing uses an AES score of 6.5. We use non-distilled checkpoints with active two-branch CFG throughout. For each model, the full-precision reference and both quantized variants share the prompt, generation seed, resolution, frame count, frame rate, and sampling configuration.

Table 9: **Paired confidence intervals for video-generation metrics on the fixed $n = 256$ VBench subset.** Each +GCBT cell reports the absolute mean, followed by the pointwise paired 95% confidence interval of improvement: Native minus GCBT for LPIPS, and GCBT minus Native for all other metrics. Positive values favor GCBT. Bold marks favorable intervals excluding zero. The intervals describe paired differences, not uncertainty intervals around the absolute means.

PTQ Host	Model	Method	Teacher Fidelity			VBench		
			LPIPS ↓	SSIM ↑	PSNR ↑	S.C. ↑	A.Q. ↑	I.Q. ↑
OrbitQuant W4A4	Wan2.1-T2V-1.3B CFG 5	Native	0.545	0.497	12.871	0.957	0.527	0.672
		+ GCBT	**0.504** [0.036070, 0.046001]	**0.518** [0.017282, 0.025529]	**12.965** [0.017197, 0.179532]	**0.975** [0.017177, 0.020618]	**0.584** [0.051873, 0.063167]	0.675 [−0.002505, 0.007421]
	CogVideoX-2B CFG 6	Native	0.427	0.639	17.474	0.927	0.494	0.593
		+ GCBT	**0.390** [0.030245, 0.044470]	**0.649** [0.004370, 0.015454]	17.554 [−0.093723, 0.251918]	**0.937** [0.007323, 0.011380]	**0.525** [0.026006, 0.036021]	**0.606** [0.005985, 0.018786]
ViDiT-Q W4A8-MP	OpenSora v1.2 CFG 7	Native	0.544	0.540	13.871	0.935	0.455	0.490
		+ GCBT	**0.541** [0.001084, 0.003664]	**0.542** [0.001102, 0.003845]	13.891 [−0.014214, 0.059996]	**0.938** [0.000561, 0.004034]	**0.459** [0.000861, 0.007971]	0.493 [−0.000222, 0.006053]
Large-scale validation								
OrbitQuant port W4A4	Wan2.1-T2V-14B CFG 5	Native	0.431	0.590	15.021	0.971	0.588	0.704
		+ GCBT	**0.413** [0.013673, 0.022822]	**0.601** [0.007352, 0.015371]	15.029 [−0.088516, 0.103788]	**0.976** [0.003256, 0.006483]	**0.607** [0.014723, 0.023360]	0.703 [−0.003987, 0.002373]

Paired evaluation protocol. For video, we use a fixed $n = 256$ subset of VBench (Huang et al., 2024), stratified by category. Prompt subsets and generation seeds are fixed once and reused for the full-precision reference, Host PTQ, and Host+GCBT. Within each paired comparison, the checkpoint, host configuration, bitwidth, CFG scale, sampling steps, resolution, and generation seed are identical;

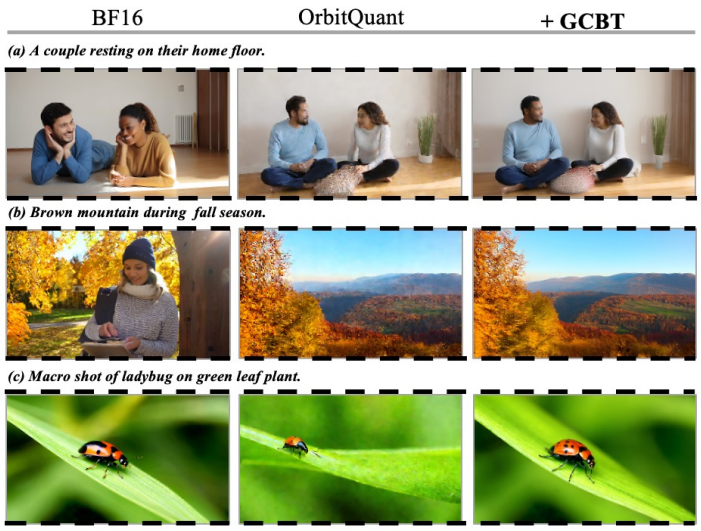


Figure 7: **Additional qualitative examples for the appendix.** We show full-precision, native quantized, and GCBT outputs under the same prompts and generation settings.

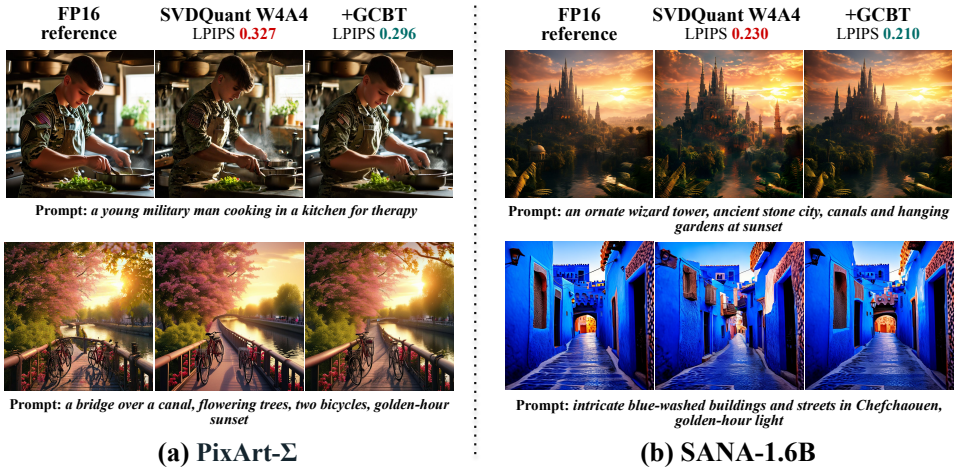


Figure 8: **More Examples of GCBT Generation**

only GCBT is changed. Framewise fidelity metrics are averaged within each clip, with prompts treated as the paired statistical units.

On the absolute LPIPS values for Wan. The relatively high BF16-referenced LPIPS on Wan is not specific to GCBT: it is already present in the OrbitQuant baseline. In video diffusion, quantization can shift the sampling trajectory enough to produce a different yet still prompt-consistent realization, making frame-aligned distance to a single BF16 sample large even when visual quality remains good. Figure 7 illustrates this behavior. Accordingly, we interpret LPIPS here as BF16-reference fidelity rather than standalone generation quality, and report VBench metrics alongside it.

C LIMITATIONS

GCBT has several scope limitations. First, its closed-form optimality is derived under the equal-rate quantization-noise surrogate in Sec. 3.3, which assumes that transformed-coordinate quantization

error scales with projected second moments and neglects cross-coordinate error terms. The resulting rotation is therefore optimal for this surrogate, rather than being guaranteed to minimize the exact distortion of every host quantizer or end-to-end generation error.

Second, GCBT requires an explicit matched conditional–unconditional branch pair and is therefore directly applicable to inference pipelines with active two-branch CFG. We do not study distilled or single-pass guidance schemes, nor timestep-varying or dynamic CFG schedules.

Finally, our CFG-aware comparison uses a GAMP-inspired calibration that we implemented ourselves because no public implementation of GAMP is available. This experiment should therefore be interpreted as evidence that GCBT remains complementary to guided-output calibration, rather than as an official reproduction of GAMP; results with the original implementation may differ.