STRIDE: TRAINING-FREE DIVERSITY GUIDANCE VIA PCA-DIRECTED FEATURE PERTURBATION IN SINGLE-STEP DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distilled one-step (T=1) or few-step (T$\leq$4) diffusion models enable real-time image generation but often exhibit reduced sample diversity compared to their multi-step counterparts. In multi-step diffusion, diversity can be introduced through schedules, trajectories, or iterative optimization; however, these mechanisms are unavailable in the few-step or single-step setting, limiting the effectiveness of existing diversity-enhancing methods. A natural alternative is to perturb intermediate features, but naive feature perturbation is often ineffective, either yielding limited diversity gains or degrading generation quality. We argue that effective diversity injection in few-step models requires perturbations that respect the model's learned feature geometry. Based on this insight, we propose STRIDE, a training-free method that requires no per-image optimization or backward passes and operates in a single forward pass. STRIDE injects spatially coherent (pink) noise into intermediate transformer features, projected onto the principal components of the model's own activations, ensuring that perturbations lie on the learned feature manifold. This design enables controlled variation along meaningful directions in the representation space. Extensive experiments on FLUX.1-schnell and SD3.5 Turbo across COCO, DrawBench, PartiPrompts, and GenEval show that STRIDE consistently improves diversity while maintaining strong text alignment. In particular, STRIDE reduces intra-batch similarity with minimal impact on CLIP score, and Pareto-dominates existing training-free baselines on the diversity-fidelity frontier. These results highlight that, in the absence of iterative refinement, improving diversity in few-step and one-step diffusion depends not on increasing perturbation strength, but on aligning perturbations with the model's internal representation structure.



(a) (b)

Figure 1: **STRIDE motivation.** **(a)** STRIDE vs. random-seed baseline on FLUX.1-schnell: each block shows four samples from one DrawBench prompt; STRIDE produces diverse poses and styles. **(b)** Unstructured noise pushes features off-manifold and is treated as corruption, while PCA-directed perturbation keeps noise on the learned manifold $\mathcal{M}$.

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

Distilled one-step image generators such as FLUX.1-schnell (Black Forest Labs, 2024) and few-step generators like SD3.5-Turbo (Stability AI, 2024) have turned text-to-image generation into a real-time operation, closing the gap with their multi-step teachers on single-sample quality while running an order of magnitude faster. A trade-off of collapsing generation into one or a few forward passes, however, is reduced sample diversity: given a fixed prompt and different random seeds, distilled students often produce visually similar outputs. Gandikota and Bau (Gandikota & Bau, 2026) measure a twenty-two percent drop in average pairwise DreamSim relative to an undistilled teacher, and attribute this to early structural commitment in the generation process. For applications that rely on stochastic exploration at a fixed prompt (creative iteration, dataset synthesis, design variation), few-step models are efficient but offer limited variability.

Existing approaches to improving diversity largely adapt techniques developed for multi-step diffusion, such as conditioning annealing across timesteps (Sadat et al., 2023), interval-restricted guidance (Kynkäänniemi et al., 2024), trajectory-level optimization (Jalali et al., 2025), or latent optimization under diversity objectives (e.g., DPP-based methods (Harrington et al., 2025)). These methods fundamentally rely on temporal degrees of freedom (schedules, trajectories, or iterative refinement) that are absent in the $T{=}1$ or $T{\leq}4$ setting. As a result, when applied to few-step or especially single-step models, they either become ineffective or degenerate to their first iteration, leaving the diversity gap largely unaddressed.

A natural alternative is to introduce stochasticity directly into intermediate feature representations, treating transformer blocks as the locus of variation rather than the noise schedule. Feature perturbation is compatible with single-pass inference and architecture-agnostic. However, naive feature perturbation is often ineffective: unstructured noise injected into intermediate features tends to yield limited diversity gains or degrade generation quality. This suggests that, in few-step models, not all perturbations are equally interpretable as valid variations by the network.

We argue that effective diversity injection requires perturbations that respect the model's learned feature structure. In particular, meaningful variation should lie along directions that are consistent with the model's internal representation geometry. To this end, we approximate the local feature manifold using the principal components of the model's own activations, computed per image. This provides a data-dependent basis capturing dominant modes of variation in the current representation. Constraining perturbations to this subspace lets noise be interpreted as plausible variation rather than arbitrary corruption.

Based on this insight, we introduce **STRIDE**, a training-free method that injects structured noise into intermediate transformer features during a single forward pass. STRIDE combines three key design choices: (i) spatially coherent ($1/f$-spectrum) noise to preserve structured variation across tokens, (ii) projection onto the top K-principal components of feature activations to approximate the learned feature manifold, and (iii) injection at early, structure-forming layers where global composition is determined. These components work jointly to enable controlled diversity while maintaining fidelity.

Our method is implemented via a single forward hook and requires no training, backpropagation, or iterative optimization, with a single scalar controlling perturbation strength. To our knowledge, STRIDE is among the first methods that simultaneously operate in a training-free, inference-only, and single-forward-pass regime for diversity enhancement in distilled diffusion models.

On FLUX.1-schnell and SD3.5-Turbo across DrawBench, PartiPrompts, COCO, and GenEval, STRIDE consistently improves diversity while preserving text alignment, Pareto-dominating single-pass training-free baselines on the diversity-fidelity frontier (Figure 2), and transfers to **FLUX.2-klein-4B** and **HunyuanImage-2.1** with only α recalibrated (App. A.6).[1] Full code will be released upon acceptance.

2 RELATED WORK

Mode collapse in distilled one-step generators. Adversarial, consistency, and score-distillation schedulers (Sauer et al., 2024b; Song et al., 2023; Yin et al., 2024b) collapse multi-step denoising

[1] Tables and figures numbered with an S prefix are in the appendix.

into one forward pass but reduce per-prompt sample diversity. Gandikota and Bau (Gandikota & Bau, 2026) report a twenty-two percent drop in pairwise DreamSim (0.337 to 0.264) on SDXL-DMD2 and localize the cause to a first-step structural commitment in the distilled student; their teacher-for-step-one remedy is unavailable at $T{=}1$. Training-time counterparts such as DP-DMD (Wu et al., 2026) replace the distillation objective itself. To our knowledge, STRIDE is the first diversity method that is simultaneously training-free, inference-only, and single-forward-pass at $T{=}1$: every prior method drops at least one of these three axes.

Inference-time remedies that assume $T > 1$. CADS (Sadat et al., 2023) anneals conditioning across timesteps; its schedule is undefined at $T{=}1$. Kynkäänniemi et al. (Kynkäänniemi et al., 2024) restrict guidance to a middle noise-level interval and show early denoising steps set diversity, a finding STRIDE operationalizes when the interval collapses to the only step available. SPARKE (Jalali et al., 2025) backpropagates a kernel-entropy repulsion reward through the denoising trajectory, which is structurally inapplicable at $T{=}1$. TPSO (Meng et al., 2025) iteratively searches text-embedding offsets under a dissimilarity objective. Harrington et al. (Harrington et al., 2025) replace the initial latent with pink noise and optimize the batch via a Determinantal Point Process against an HPSv2 quality kernel, requiring roughly fifteen forward passes per batch. STRIDE retains Harrington's spectral prior (pink beats white) but forgoes the DPP loop and its quality kernel for a single-pass PCA projection onto the learned feature manifold. Particle Guidance (Corso et al., 2023) and SGI (Parmar et al., 2025) enforce batch-level diversity via joint potentials or post-hoc selection and are composable with ours.

Feature-level perturbation at inference. Self- and perturbed-attention guidance (Hong et al., 2023; Ahn et al., 2024; Hong, 2024) target fidelity and reduce diversity as a side effect. Token Perturbation Guidance (TPG Rajabi et al. (2025)) is the closest neighbor: it shuffles DiT tokens at selected layers but requires two forward passes per timestep and applies an unstructured permutation agnostic to the model's feature geometry.

STRIDE uses a single pass and projects the perturbation onto the current block's principal directions, so the injected energy lies on the learned feature manifold rather than its orthogonal complement.

Principal components of diffusion features. Concept Sliders (Gandikota et al., 2024), SliderSpace (Gandikota et al., 2025), and h-space analyses (Kwon et al., 2023) identify semantically meaningful directions in diffusion feature spaces for targeted attribute editing; each principal component there serves as a controllable axis to move along. STRIDE uses the same geometric objective (principal components of a block's activations) for a complementary purpose: rather than move along a single learned direction, it injects spatially coherent pink noise inside the entire principal subspace, so each generation takes an independent structural step that preserves on-manifold plausibility but leaves its content unconstrained. Recent low-rank analyses of diffusion-transformer activations (Gan et al., 2025) are consistent with this subspace view.

3 METHOD

We present **STRIDE** (**S**patial **T**ransformer **R**epresentation-space **I**njection for **D**iversity **E**nhancement), a training-free method for increasing output diversity in single or few-step diffusion transformers without modifying weights or requiring iterative optimization. STRIDE injects structured noise into intermediate transformer representations during a single forward pass. The key idea is to align perturbations with the model's internal representation structure (Figure 1b). To this end, STRIDE projects noise onto the principal components of the model's feature activations, computed per image. This projection provides a data-dependent approximation of the local feature manifold, capturing dominant directions of variation in the representation space. By constraining perturbations to this subspace, STRIDE enables controlled diversity while preserving semantic consistency.

Motivation: A natural approach to recovering diversity in distilled models is to perturb the initial latent Harrington et al. (2025). To evaluate this strategy, we conduct controlled experiments on FLUX.1-schnell by injecting pink noise[2] at the input latent and measuring its effect on output diversity (Table 1). We find this to be insufficient at $T{=}1$, since input-level pink noise yields only modest diversity gains on FLUX.1-schnell and, at matched diversity, degrades alignment (Table 2). This suggests that, in single-step models, much of the stochastic variation introduced at the input is

[2] As shown in Harrington et al. (2025) and confirmed in our study, injecting pink noise improves diversity.

Table 1: **Main results on COCO** (2000 prompts $\times$ 10 images), **DrawBench** (199 $\times$ 4), and **PartiPrompts** (1632 $\times$ 4). Methods use their respective Pareto-optimal settings. (App. A.2.2);*comparable on the joint diversity–alignment–quality frontier (Figure 2 and Table 2), not on individual columns across rows*. KID ($\times 10^{-3}$) is reported only on COCO. Per-prompt standard deviations and significance analyses appear in Tables S8, S9 and S7. †CADS collapses on SD3.5 Turbo (CLIP < 23 across all tested settings). DDC and SPELL require multiple passes and are inapplicable to 1-step FLUX.

	COCO					DrawBench				PartiPrompts			
Method	InBSim↓	VD/p↑	CLIP↑	HPS↑	KID↓	InBSim↓	VD/p↑	CLIP↑	HPS↑	InBSim↓	VD/p↑	CLIP↑	HPS↑
FLUX.1-schnell (1-step)													
Baseline	0.738	2.794	26.69	0.293	6.27	0.662	2.236	28.47	0.293	0.722	2.025	28.13	0.290
CADS	0.659	3.433	26.44	0.291	6.75	0.556	2.578	27.09	0.290	0.630	2.328	27.37	0.288
Input Noise	0.661	3.443	26.11	0.288	9.03	0.566	2.570	27.43	0.284	0.632	2.345	27.31	0.283
STRIDE No-PCA	0.762	2.603	28.05	0.276	8.09	0.673	2.205	29.04	0.280	0.744	1.952	28.77	0.273
STRIDE (Ours)	0.683	3.268	27.66	0.283	5.74	0.590	2.496	28.61	0.284	0.653	2.275	28.56	0.281
SD3.5 Large Turbo (4-step)													
Baseline	0.771	2.527	26.77	0.290	9.62	0.707	2.067	28.18	0.291	0.769	1.855	28.09	0.289
CADS†	0.391	5.888	22.56	0.263	17.54	0.318	3.309	22.67	0.262	0.356	3.179	22.88	0.257
Input Noise	0.749	2.703	26.72	0.291	9.14	0.668	2.203	28.34	0.292	0.740	1.959	28.02	0.288
DDC	0.679	3.301	26.70	0.276	7.12	0.600	2.436	27.84	0.274	0.675	2.187	27.70	0.273
SPELL	0.603	3.971	24.44	0.247	39.91	0.638	2.334	26.32	0.274	0.690	2.157	26.69	0.269
STRIDE No-PCA	0.766	2.573	26.82	0.289	9.41	0.704	2.082	28.07	0.291	0.766	1.870	28.00	0.287
STRIDE (Ours)	0.725	2.905	26.69	0.289	8.68	0.639	2.300	28.07	0.288	0.714	2.051	27.91	0.286

Table 2: CLIP / HPS at matched InBSim on DrawBench. STRIDE: absolute scores; others: differences from STRIDE. †CADS collapses on SD3.5 Turbo. Full results: Table S29 and S30.

Backbone @ InBSim	CLIP / HPS	Pink latent (Δ)	CADS (Δ)	No-PCA (Δ)
FLUX.1-schnell @ 0.62	29.20 / 0.288	-1.10 / $+0.003$	-1.38 / $+0.003$	-0.57 / -0.012
SD3.5 Turbo @ 0.65	28.08 / 0.289	-0.70 / $+0.002$	n/r†	-0.84 / -0.002

suppressed by the deterministic mapping of the network. An alternative is to perturb intermediate features, which may offer a more direct handle on the learned representation. However, naive feature perturbation proves counterproductive: unstructured noise disrupts the learned representation and increases InBSim above baseline.

The key insight of this paper is that not all directions in feature space correspond to meaningful variation. The model's representations concentrate along a low-dimensional manifold shaped by dominant modes of variation, and perturbations that deviate from this manifold are interpreted as noise rather than diversity. Therefore, useful variation must be introduced along these intrinsic directions, as illustrated in Figure 1b. This motivates the design of structured, feature-aware perturbations that explicitly respect the geometry of the learned representation space.

Feature-Level PCA-Directed Noise Injection: Building on the insight that meaningful variation must align with the intrinsic geometry of the feature space, we construct perturbations that are explicitly constrained to lie along the dominant modes of the model's representations. Concretely, we achieve this via a PCA-directed noise injection mechanism applied to intermediate transformer features.

Let $\mathbf{h} \in \mathbb{R}^{B\times N\times D}$ denote the hidden states at a target transformer block, where B is the batch size, $N = H \times W$ is the number of spatial tokens, and D is the feature dimension. STRIDE modifies $\mathbf{h}$ via a forward hook applied after each target block:

$$\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{d} \tag{1}$$

where α is a scalar perturbation strength and $\mathbf{d}$ is a manifold-aligned structured perturbation.

To construct geometry-aligned perturbations, we must first identify the dominant modes of variation in the feature space. However, directly applying PCA to token-wise features treats each spatial location independently and fails to capture coherent spatial structures. We therefore group spatially adjacent tokens into patches, enabling the subsequent analysis to reflect structured, spatial patterns.

Table 3: **FLUX.2-klein-4B and HunyuanImage-2.1 on DrawBench**. HPS is preserved within 0.003 on both backbones and is statistically unchanged on klein (paired Wilcoxon, Table S7).

Backbone (STRIDE config)	InBSim↓	VD/p↑	CLIP↑	HPS↑	△HPS
FLUX.2-klein-4B, 4-step (K=512, α=1.5)	0.736 → 0.686	1.98 → 2.15	28.08 → 28.20	0.295 → 0.295	−0.0004 (n.s.)
HunyuanImage-2.1, 8-step (K=1024, α=14)	0.694 → 0.630	2.12 → 2.32	26.43 → 26.00	0.292 → 0.289	−0.003

Step 1: Spatial Patchification. We reshape the spatial grid of tokens into $P \times P$ patches with stride S, concatenating the features within each patch:

$$\mathbf{h}_{\text{2D}} \in \mathbb{R}^{H \times W \times D} \longrightarrow \mathbf{H}_{\text{patches}} \in \mathbb{R}^{M \times (P^2 D)} \tag{2}$$

where $M = \lfloor (H-P)/S + 1 \rfloor \times \lfloor (W-P)/S + 1 \rfloor$ is the number of patches and $P^2 D$ is the concatenated feature dimension.

This couples adjacent tokens so that PCA captures spatially coherent modes of variation (e.g., layout and composition).

Step 2: PCA. We compute truncated PCA on the centered patch features of the *current* image:

$$\bar{\mathbf{H}} = \mathbf{H}_{\text{patches}} - \boldsymbol{\mu}, \qquad \bar{\mathbf{H}} \approx \mathbf{U}\mathbf{S}\mathbf{V}^\top \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{P^2 D \times K}$ contains the top-K right singular vectors (principal directions) and $\mathbf{S} = \text{diag}(s_1, \ldots, s_K)$ contains the corresponding singular values. We compute this via randomized SVD Halko et al. (2011) (`torch.pca_lowrank`, q=K, niter=2).

The PCA is computed **online** from each image's own activations to ensure that the principal directions are *prompt-specific*: the spatial PCA basis for "a red car" differs from that of "an astronaut on the moon," capturing the content-specific modes of variation available to the model (Supp. Table S2). Crucially, these directions define the local tangent space of the model's representation manifold, i.e., the directions along which variation is semantically meaningful.

Step 3: Noise Generation and Projection. We generate spatially correlated noise via a 2D FFT-based filter $\mathcal{F}$ that biases the perturbation toward low spatial frequencies as discussed in (Harrington et al., 2025):

$$\boldsymbol{\epsilon}_{\text{pink}} = \mathcal{F}^{-1}\left[\frac{1}{(1+|\mathbf{f}|)^{f_\alpha}} \cdot \mathcal{F}[\boldsymbol{\epsilon}_{\text{white}}]\right] \tag{4}$$

where $\boldsymbol{\epsilon}_{\text{white}} \sim \mathcal{N}(0,1)$, $|\mathbf{f}|$ denotes the radial frequency in 2D Fourier space, and $f_\alpha \geq 0$ controls the spectral slope Harrington et al. (2025). The filter $1/(1+|\mathbf{f}|)^{f_\alpha}$ attenuates high-frequency components: f_α=0 recovers white noise (uniform power spectrum), while larger f_α progressively suppresses fine-scale variation in favour of spatially coherent low-frequency structure. This shaping matters because high-frequency noise injected into transformer features tends to manifest as texture-level artefacts that the model treats as corruption Harrington et al. (2025), whereas low-frequency noise induces coherent spatial perturbations (e.g., layout shifts, compositional rearrangements) that align with the structural decisions distilled few-step models make in their early forward passes.

The noise is patchified identically to the features (Eq. 2), yielding $\boldsymbol{\epsilon}_{\text{patches}}$, and projected onto the PCA basis with singular-value weighting:

$$\mathbf{c} = \boldsymbol{\epsilon}_{\text{patches}}\,\mathbf{V}, \qquad \mathbf{c}_{\text{scaled}} = \mathbf{c} \odot \frac{\mathbf{s}}{\bar{s}}, \qquad \mathbf{d}_{\text{patches}} = \mathbf{c}_{\text{scaled}}\,\mathbf{V}^\top \tag{5}$$

where $\mathbf{V} \in \mathbb{R}^{P^2 D \times K}$ is the PCA basis from Eq. 3, $\mathbf{s} = (s_1, \ldots, s_K)$ are the corresponding singular values, and $\bar{s}$ is their mean. The projection $\mathbf{c} = \boldsymbol{\epsilon}_{\text{patches}}\mathbf{V}$ expresses the patchified pink noise in the basis of the model's dominant variance directions, discarding the orthogonal complement that lies off the local feature manifold. The singular-value weighting allocates perturbation energy according to the model's intrinsic variance structure, ensuring that dominant modes (e.g., global structure and layout) are emphasized while suppressing directions corresponding to noise or unstable features. The resulting directed perturbation $\mathbf{d}_{\text{patches}}$ is then unpatchified back to the full spatial grid and injected via Eq. 1, ensuring that all perturbations remain aligned with the model's learned feature geometry.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate on two primary models spanning single-step and few-step inference (Sauer et al., 2024b;a; Luo et al., 2023; Song et al., 2023; Esser et al., 2024; Liu et al., 2023; Chen et al., 2025b): **FLUX.1-schnell** Black Forest Labs (2024), a 12B-parameter MMDiT with 1-step distilled inference (19 double-stream blocks, 38 single-stream blocks, 4096 tokens at 1024×1024), and **SD3.5 Large Turbo** Stability AI (2024), an 8.1B-parameter MMDiT distilled to 4 denoising steps(38 joint transformer blocks, 4096 tokens at 1024×1024). These models represent the two dominant acceleration paradigms: single-step (no iterative refinement) and few-step distillation (compressed denoising schedule). Transfer to two further backbones, FLUX.2-klein-4B (4-step) and HunyuanImage-2.1 (8-step), is reported on DrawBench in Table 3 and App. A.6.

Datasets. We evaluate on four standard text-to-image benchmarks: **MS-COCO** Lin et al. (2014) for large-scale distributional diversity, **DrawBench** Saharia et al. (2022) for compositionally challenging prompts, **PartiPrompts** Yu et al. (2022) for broad category coverage, and **GenEval** Ghosh et al. (2023) for compositional accuracy. For MS-COCO, we use a subset of 2,000 captions from the 2014 validation split. See Supp. Sec A.2.1.

Metrics. We report four primary metrics: **InBatchSim (InBSim)** $\downarrow$, the average pairwise CLIP similarity among images generated from the same prompt (lower is more diverse); **Vendi Score per prompt (VD/$\mathbf{p_{DINO}}$)** (Friedman & Dieng, 2022) $\uparrow$, the effective number of distinct outputs per prompt using DINO (Oquab et al., 2023) features; **CLIP Score** (Hessel et al., 2021) $\uparrow$ for text-image alignment; and **HPSv2** (Wu et al., 2023) $\uparrow$ (hereafter **HPS** for brevity) for perceptual quality. On COCO we additionally report **KID** (Bińkowski et al., 2018) $\downarrow$ (distributional distance to real images), and on GenEval **DreamSim (DrSim)** (Fu et al., 2023) $\uparrow$, the average pairwise distance among generations from the same prompt, to match the diversity protocol of Harrington et al. (2025).

Baselines. We compare STRIDE against five training-free diversity methods. **CADS** Sadat et al. (2023) adds noise to the text conditioning embeddings; for single-step models (FLUX) the annealing schedule collapses to a single noise level, while for SD3.5 Turbo (4 steps) the full per-timestep schedule operates as designed. **Input Noise** blends pink latent noise into the initial Gaussian noise before denoising, following the frequency-shaped approach of Harrington et al. Harrington et al. (2025): $\mathbf{z}' = (1-\alpha)\,\mathbf{z} + \alpha\,\mathbf{z}_{\text{pink}}$. **STRIDE No-PCA** (ablation) adds pink noise directly to transformer features *without* spatial PCA projection, serving as the "without on-manifold projection" ablation of our method. **DDC** Gandikota & Bau (2026) uses the Hybrid Inference variant (Algorithm 1, k=1): SD3.5 Large (teacher) for the first denoising step, SD3.5 Turbo (student) for steps 2–4. **SPELL** Kirchhof et al. (2024) is applied to SD3.5 Turbo only (structurally inapplicable to single-step FLUX); both hyperparameters (r, λ) are jointly swept following the paper's calibration protocol (Appendix A.2.2). All baseline hyperparameters are selected via grid search to identify their respective Pareto-optimal operating points; see Supp. Sec. A.2.2 for grids and final values. Statistical robustness is verified across both prompts (per-prompt standard errors, Table S8) and seeds (three-seed reproducibility, Table S9), and by paired per-prompt Wilcoxon tests (Table S7).

STRIDE. STRIDE is applied to the *early* transformer blocks at the *first* denoising step only, consistent with our finding (Supp. Sec. A.3) and the analysis of Gandikota & Bau (2026) that compositional structure is established at step 0 in the early MM-DiT layers. Two design choices adapt to each model: (i) the layer set spans roughly the first third of the dual-stream blocks (L0–7 of 19 for FLUX, L0–11 of 38 for SD3.5 Turbo); (ii) we scale the perturbation magnitude α inversely with the number of denoising steps to prevent compounding across the schedule, with diversity primarily modulated through the frequency-shaping exponent f_α. We use K=1024 principal components for both backbones. Per-model hyper-parameters are reported in Supp. Sec. A.2.3.

4.2 MAIN RESULTS

Tables 1 and 4 report each method at its selected Pareto-optimal operating point, where STRIDE achieves quality comparable to existing baselines and closely related methods. At matched diversity, Table 2 and the Pareto curves in Figures 2 and 3 show that STRIDE achieves stronger text–image alignment, with a minimal HPS cost discussed in Section 5.

Table 4: **Results on GenEval** (553 prompts × 4 images; FLUX at 512^2, SD3.5-Turbo at 1024^2). *Comparable on the joint diversity–alignment–quality frontier (Figure 2 and Table 2), not on individual columns across rows*. On SD3.5 Turbo, CADS collapses†. Rows marked ⋆ are taken directly from Harrington et al. (2025) for a comparable FLUX-schnell setting; their GI row reproduces Parmar et al. (2025). Per-prompt standard deviations in Table S8; further setup details in Sec. A.2.2. Each method is reported at its own Pareto-optimal operating point (selected on DrawBench, App. A.2.2). STRIDE-induced diversity is comparable to optimisation-based approaches.

	FLUX.1-schnell 1-Step					**SD3.5 Turbo** 4-Step				
Method	**InBSim↓**	**DrSim↑**	**VD/p↑**	**CLIP↑**	**HPS↑**	**InBSim↓**	**DrSim↑**	**VD/p↑**	**CLIP↑**	**HPS↑**
Baseline	0.669	0.363	2.204	28.34	0.296	0.741	0.264	1.948	29.20	0.290
CADS	0.559	0.431	2.550	27.34	0.291	0.299†	0.634†	3.348†	22.82†	0.257†
Input Noise	0.592	0.445	2.476	27.32	0.287	0.701	0.314	2.087	28.92	0.291
DDC	–	–	–	–	–	0.627	0.368	2.343	28.52	0.274
SPELL	–	–	–	–	–	0.651	0.369	2.284	27.88	0.272
i.i.d.⋆	–	0.307	2.013	–	0.304	–	–	–	–	–
GI⋆	–	0.413	2.473	–	0.296	–	–	–	–	–
NoiseOpt⋆	–	0.446	2.753	–	0.293	–	–	–	–	–
STRIDE No-PCA	0.696	0.322	2.123	29.26	0.279	0.730	0.274	1.986	29.13	0.290
STRIDE (Ours)	0.607	0.423	2.428	29.04	0.282	0.679	0.337	2.175	28.86	0.287

FLUX.1-schnell. STRIDE is the only evaluated method to improve both diversity and text alignment across all four datasets (Tables 1 and 4). On COCO, DrawBench, and PartiPrompts, it reduces InBSim by 7.5%, 10.9%, and 9.6%, while increasing CLIP by 0.96, 0.14, and 0.43, respectively. CADS and input noise achieve comparable or larger diversity gains but reduce CLIP by 0.25–1.38(Table 2). Pareto curves show STRIDE's stronger alignment at matched diversity over the evaluated range (Figures 2a and 3a); qualitative comparisons appear in Figure 1a.

Paired per-prompt Wilcoxon tests confirm diversity gains across backbones ($p < 0.001$, Holm-corrected; $d_z = 0.64$ on FLUX.1-schnell), with no significant CLIP change on FLUX.1-schnell or SD3.5 Turbo (Table S7). One-step FLUX.1-schnell+STRIDE also exceeds 50-step FLUX.1-dev in diversity and CLIP at approximately $1/12$ the runtime, with an HPS cost of 0.015 (App. A.7, Table S32). On GenEval, STRIDE achieves DreamSim and Vendi/p scores of 0.423 and 2.428, comparable to concurrent noise optimisation (0.446, 2.753; Harrington et al., 2025). Its computational cost remains modest by comparison even at $K = 1024$, requiring a single forward pass without an optimisation loop rather than approximately 80 backward iterations per prompt.

SD3.5 Large Turbo. CADS collapses on this backbone (CLIP < 23 across all tested settings; Table 1). STRIDE reduces InBSim by 6.0% on COCO and 9.6% on DrawBench, with small CLIP decreases (0.08 and 0.11, respectively), and achieves the second-lowest KID (8.68 vs. baseline 9.62). Input noise yields smaller InBSim reductions ($2.9\%/5.5\%$), while No-PCA achieves only 0.7% on COCO. STRIDE therefore improves diversity without quality collapse (Table 2, Figures 2b and 3b).

DDC-Hybrid (Gandikota & Bau, 2026) reaches a comparable Pareto point (InBSim 0.60, CLIP 27.84) but loads both teacher and student, requiring 56.16 GB versus STRIDE's 28.11 GB. This $2\times$ memory requirement necessitates CPU offload on an A100-40GB, increasing runtime to 10.5 s/image ($+805\%$ over baseline). STRIDE uses only the student in a single forward pass, taking 1.50 s/image ($+76.3\%$ at $K = 512$). SPELL (Kirchhof et al., 2024) achieves comparable DrawBench diversity (9.8% lower InBSim), but incurs approximately $15\times$ the CLIP loss (1.86 vs. 0.12), illustrating the quality cost of repulsion in this few-step setting.

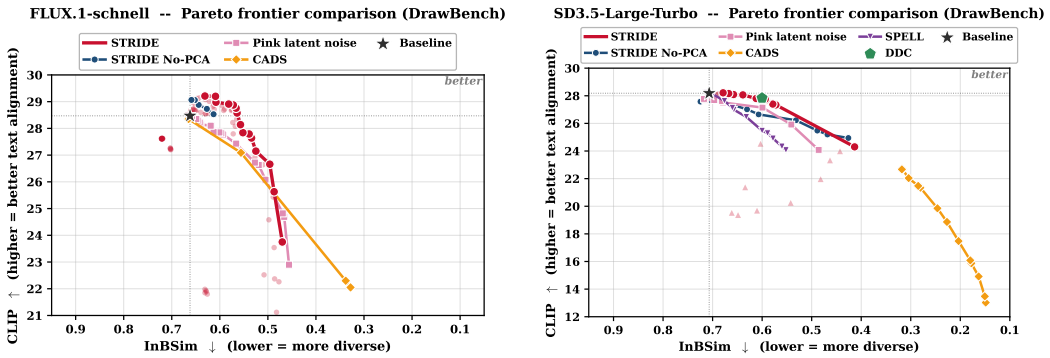
Quality trade-off. Diversity gains incur small but statistically significant HPS reductions (ΔHPS $= -0.0091$ on FLUX.1-schnell and -0.0029 on SD3.5 Turbo; both $p < 0.001$), with no significant CLIP change (Table S7). On FLUX-Klein, the HPS decrease is smaller and not statistically significant (Table 3), suggesting a reduced HPS trade-off in multi-step ($T > 1$) setting. Reducing the rank to K=512 mitigates this trade-off (ΔHPS $= -0.0034$, ΔCLIP $= +0.61$). At matched diversity, noise baselines achieve 0.001–0.005 higher HPS but 0.6–1.4 lower CLIP than STRIDE (Table 2).

Compute cost. Jointly adjusting rank and perturbation strength provides a practical runtime–diversity trade-off (Tables S17 and S24). At $K = 1024$, online SVD adds $195\%/167\%$ runtime on FLUX.1-schnell/SD3.5 Turbo, without additional forward passes or increased peak VRAM. On FLUX.1-schnell, $K = 512$, $\alpha = 0.8$ retains 76% of the diversity gain (InBSim 0.608), improves

CLIP/HPS (28.98/0.287 vs. 28.61/0.284), and more than halves the overhead. On SD3.5 Turbo, $K = 512$, $\alpha = 0.2$ achieves comparable diversity (InBSim 0.633 vs. 0.639) and HPS, with CLIP lower by 0.18, while reducing overhead from 167% to 76% (Table S11). For comparison, the noise-optimisation method of Harrington et al. (2025) requires approximately 80 backward iterations per prompt, whereas STRIDE uses a single forward pass without backpropagation.

Hyperparameter ablation. We fix $P = 2$: larger patches ($P \in \{4, 8, 16\}$) return InBSim toward baseline on both backbones (Table 5), consistent with quadratic growth in the joint patch dimension P^2D reducing the projected subspace's relative size. Full ablations appear in Supp. Sec. A.3 and Tables S13 and S20; per-prompt standard deviations and significance analyses in Supp. Sec. A.2.5; and qualitative comparisons in Supp. Sec. A.8.

Human Evaluation. In a two-alternative forced-choice study on 24 DrawBench prompts, 119 participants chose which of two 4-image sets was more diverse. Across 1,370 responses STRIDE was preferred **67.9%** of the time (95% Wilson CI $[65.4\%, 70.3\%]$; binomial $p < 10^{-39}$), winning 22 of 24 prompts (Table S6; protocol, per-prompt results and the single baseline-preferred prompt in Supp. Sec. A.2.4). For detailed analysis and ablation, see Supp. Sec. A.3.



(a) FLUX.1-schnell (b) SD3.5 Turbo

Figure 2: **Pareto frontiers for diversity-quality trade-off.** We compare STRIDE against baselines using InBSim↓ and CLIP↑ on FLUX.1-schnell and SD3.5 Turbo. STRIDE achieves the strongest Pareto frontier on both architectures, improving diversity while preserving image-text alignment.

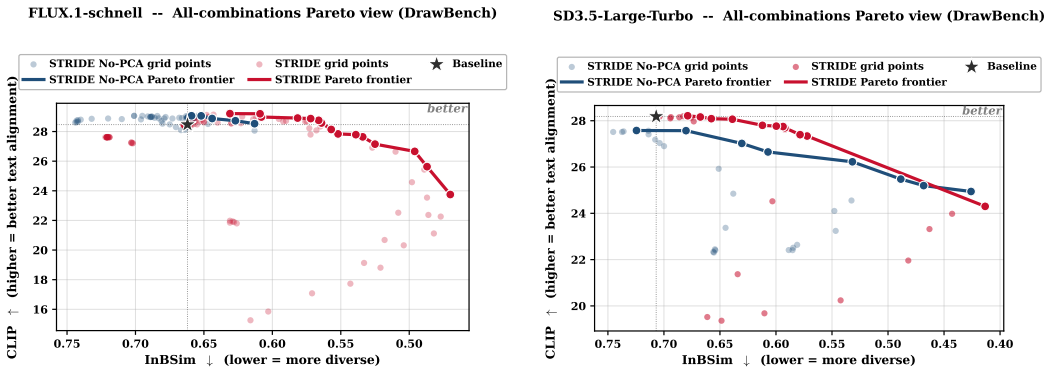
Table 5: **Patch size P ablation on DrawBench** (199 prompts × 4 images), at the per-backbone perturbation strength. Cells are InBSim↓ / CLIP↑. The locked configuration P=2 is bolded; larger P dilutes per-patch PCA coverage and reverts InBSim toward baseline.

Backbone	Base	P=2	P=4	P=8	P=16
FLUX.1-schnell (α=0.5)	0.664 / 28.40	**0.590 / 28.61**	0.663 / 28.50	0.663 / 28.39	0.664 / 28.41
SD3.5 Turbo (α=0.1)	0.707 / 28.18	**0.639 / 28.07**	0.693 / 28.18	0.701 / 28.18	0.701 / 28.16

5 OBSERVATIONS

Why PCA is essential. STRIDE No-PCA (unstructured pink noise injected directly at the feature level, without PCA projection) shows that on-manifold projection is *necessary*, not merely beneficial Table 2. On FLUX, unstructured noise makes outputs *more* similar: COCO InBSim *rises* +3.3% (0.762 vs. baseline 0.738) and PartiPrompts +3.0% (0.744 vs. 0.722). However, PCA projection inverts this at identical α, yielding −7.5% InBSim on COCO. No-PCA's higher CLIP in Tables 1 and 4 is a symptom of this reduced diversity, not an alignment benefit: at matched diversity it is worse than STRIDE on both CLIP and HPS (Table S29). We hypothesise that off-subspace energy is displaced from the learned activation distribution and re-grounded against the text condition by the corrective dynamics of subsequent layers, raising CLIP while tightening the output distribution; see App. A.4 for a detailed discussion and the supporting measurements.

Layer selection and timestep gating. On both MM-DiT models, diversity originates from perturbation at **early** transformer blocks (L0–7 for FLUX, L0–11 for SD3.5 Turbo); late blocks produce negligible diversity, consistent with the interpretation that early blocks set spatial composition while later blocks refine local detail Ahn et al. (2025a); Jiang et al. (2025). For SD3.5 Turbo's 4-step schedule, perturbation at **step 0** is most effective; early steps commit to the global composition, so late-step perturbations can only alter local details (see Supp. Table S19).



(a) FLUX.1-schnell

(b) SD3.5 Turbo

Figure 3: **Diversity-quality Pareto frontier on DrawBench.** We compare STRIDE No-PCA (undirected pink noise) and STRIDE (PCA-directed pink noise) across frequency exponent f_α and perturbation strength α on FLUX.1-schnell and SD3.5 Turbo using 199 prompts with 4 images per prompt. STRIDE offers a better diversity-quality trade-off, achieving lower InBSim at comparable CLIP across both backbones.

We present **STRIDE**, a training-free method for improving output diversity in distilled few-step diffusion models through PCA-directed spatial feature perturbation.

Our central finding is that perturbation geometry is critical: at identical α, PCA-projected feature noise reduces InBSim by 7.5%, whereas unstructured noise increases it by 3.3% (Tables 1 and 2). STRIDE uses a single forward hook without iterative optimisation and improves the diversity–alignment trade-off over evaluated single-pass, training-free baselines on FLUX.1-schnell (1-step) and SD3.5 Large Turbo (4-step). On FLUX, it uniquely improves both diversity and text alignment; on SD3.5 Turbo, it delivers larger diversity gains than input noise while largely preserving CLIP, whereas CADS collapses. Transfer to FLUX.2-klein-4B and HunyuanImage-2.1 requires recalibrating only α, with structural hyperparameters fixed (App. A.6). Noise optimisation (Harrington et al., 2025), post-hoc selection (Parmar et al., 2025), and hybrid base-distilled inference (Gandikota & Bau, 2026) also improve diversity but require iterative search, overgeneration, or access to the undistilled model, respectively; STRIDE remains complementary to these approaches.

Limitations. STRIDE has three practical costs. **(i) Residual artefacts.** Because perturbation occurs inside the forward pass, single-step models have no subsequent denoising steps to attenuate residual noise; generated images may retain subtle high-frequency artefacts at large α or f_α, unlike multi-step models where later refinement steps smooth the perturbation. HPS decreases by 0.009 on FLUX.1-schnell and 0.003 on SD3.5 Turbo, consistently across prompts; this cost diminishes at lower rank and is absent on FLUX.2-klein-4B (Table S7). **(ii) Compute overhead.** Online PCA adds wall-clock overhead that scales with K (Table S11), and prompt-dependence of the spatial principal directions (Supp. Table S2) precludes a simple precomputed basis. More discussion in Supp. Sec A.3 **(iii) Architecture-specific tuning.** The optimal layer set and timestep gating differ across model families (L0–7 for FLUX, L0–11 with step-0-only timestep gating for SD3.5 Turbo); transferring STRIDE to a new architecture requires re-identifying the early structure-forming blocks. In practice all structural hyperparameters transferred unchanged to FLUX.2-klein-4B and HunyuanImage-2.1 with only α recalibrated (App. A.6.1)

Future work. Three directions follow naturally: (1) Amortised PCA to reduce per-image overhead; (2) Composing STRIDE with conditioning-level and latent-level diversity methods to test whether gains are super-additive; and (3) Automating noise injection-site selection for new architectures.

AI USE STATEMENT

In this work, we used generative AI tools to draft portions of the manuscript and assist in producing result visualisations. We have not used generative AI tools to develop research ideas, design experiments, implement the proposed method, analyse experimental results, or formulate conclusions. Additionally, we used generative AI tools to improve readability through grammar, spelling, and word-choice edits. We have reviewed all AI-assisted work, and manuscript drafts produced with AI assistance were subsequently revised by the authors. All research ideas, experimental design, implementations, experiments, analyses, and conclusions are the authors' own. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Minjae Kim, Jaewon Min, Wooseok Jang, Sangwu Lee, Sayak Paul, Susung Hong, and Seungryong Kim. Where and how to perturb: On the design of perturbation guidance in diffusion and flow models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. URL https://arxiv.org/abs/2506.10978.

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Minjae Kim, Jaewon Min, Wooseok Jang, Saungwu Lee, Sayak Paul, Susung Hong, and Seungryong Kim. Fine-grained perturbation guidance via attention head selection. *arXiv preprint arXiv:2506.10978*, 2025b.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Black Forest Labs. Announcing Black Forest Labs and the FLUX.1 suite of models. https://bfl.ai/blog/24-08-01-bfl, August 2024. Accessed: 2026-04-27.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. *arXiv preprint arXiv:2505.21179*, 2025a. URL https://arxiv.org/abs/2505.21179.

Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16185–16195, 2025b.

Geoffrey Cideron, Andrea Agostinelli, Johan Ferret, Sertan Girgin, Romuald Elie, Olivier Bachem, Sarah Perrin, and Alexandre Ramé. Diversity-rewarded cfg distillation. *arXiv preprint arXiv:2410.06084*, 2024. URL https://arxiv.org/abs/2410.06084.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023.

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.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.09344.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. *arXiv preprint arXiv:2505.18584*, 2025. URL https://arxiv.org/abs/2505.18584.

Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1304–1313, 2026.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. *arXiv preprint arXiv:2502.01639*, 2025. URL https://arxiv.org/abs/2502.01639.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2310.11513.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. *arXiv preprint arXiv:2601.00090*, 2025.

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

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2408.00760.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. In *NeurIPS 2025-Thirty-ninth Conference on Neural Information Processing Systems*, 2025.

Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. Sparke: Scalable prompt-aware diversity guidance in diffusion models via rke score. *arXiv preprint arXiv:2506.10173*, 2025.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025.

Kwanyoung Kim and Byeongsu Sim. Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2503.07677.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. *arXiv preprint arXiv:2410.06025*, 2024.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.10960.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

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 components for 4-bit diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2411.05007.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. URL https://arxiv.org/abs/1405.0312.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Debin Meng, Chen Jin, Zheng Gao, Yanran Li, Ioannis Patras, and Georgios Tzimiropoulos. Training-free generation of diverse and high-fidelity images via prompt semantic space optimization. *arXiv preprint arXiv:2511.19811*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. *arXiv preprint arXiv:2508.15773*, 2025.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. *arXiv preprint arXiv:2506.10036*, 2025.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Stability AI. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, October 2024. Accessed: 2026-04-27.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026. URL https://arxiv.org/abs/2602.03139.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. URL https://arxiv.org/abs/2306.09341.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://arxiv.org/abs/2405.14867.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. URL https://arxiv.org/abs/2311.18828.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. URL https://arxiv.org/abs/2206.10789.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.14919.

A Appendix

A.1 Extended Related Work

This appendix expands Section 2. We cover training-time diversity remedies, additional attention-guidance variants, timestep-aware recovery, input-level noise shaping, prompt-embedding diversification, classical feature-space analysis, and diversity metrics and benchmarks.

Training-time diversity recovery. DMD2 (Yin et al., 2024a) combines distribution matching distillation with a GAN objective and improves quality but retains the mode-seeking bias of reverse-KL distillation. DP-DMD (Wu et al., 2026) assigns v-prediction to the first step to preserve diversity and DMD to subsequent steps. Adversarial Score-Identity Distillation (SiDA; Zhou et al. (2025)) augments SiD with an adversarial term to offset mode seeking. Diversity-rewarded CFG distillation (Cideron et al., 2024) adds a per-prompt Vendi reward to the student objective. SliderSpace (Gandikota et al., 2025) trains a per-concept LoRA to expose semantically independent directions. All four operate at training time and require either distillation access or a LoRA fine-tune per concept; STRIDE requires neither.

Additional inference-time guidance. Perturbed-attention, self-attention, and smooth-energy guidance (Ahn et al., 2024; Hong et al., 2023; Hong, 2024) target fidelity. Entropy Rectifying Guidance (Ifriqi et al., 2025) temperature-scales softmax to form a CFG-like difference. Normalized Attention Guidance (Chen et al., 2025a) renormalizes scores after perturbation. Frequency-Decoupled Guidance splits guidance strength across frequency bands. SoftPAG (Ahn et al., 2025a) modulates the PAG strength continuously and reports diversity as a secondary target. PLADIS (Kim & Sim, 2025) aggressively extrapolates attention. None of these operate under a single forward-pass constraint with diversity as the primary objective, and all but two require either a guidance difference across timesteps or a second forward pass.

Diversity-via-repulsion and post-hoc selection. Particle Guidance (Corso et al., 2023) enforces repulsion across a batch during multi-step sampling via a joint potential and does not apply to single-particle, single-step generation. Scaling Group Inference (SGI; Parmar et al. (2025)) is complementary to STRIDE: it generates $M \gg K$ candidates and prunes via iterative quadratic integer programming.

Timestep-aware diversity recovery. Kynkäänniemi et al. (Kynkäänniemi et al., 2024) observe that mode collapse in distilled models originates at the first denoising timestep. Their fix uses the undistilled base model for one step and switches to the distilled model for the remainder, requiring access to both models and $T>1$. Their Tweedie-formula diagnostic shows that distilled models sometimes generate desirable features at intermediate stages and subsequently suppress them, a phenomenon consistent with our observations for unstructured feature perturbation.

Input-level noise shaping. Harrington et al. (Harrington et al., 2025) replace the initial Gaussian latent with frequency-shaped (pink) noise via the filter $1/(1+f)^{\alpha}$ and introduce a Determinantal Point Process optimization loop over initial latents that iteratively refines a batch for diversity subject to an HPSv2 quality constraint. The DPP loop requires roughly fifteen forward passes per batch for convergence on FLUX.1-schnell. STRIDE inherits the spectral insight (pink is strictly better than white) but applies it at intermediate features rather than inputs, and replaces the DPP optimization with a single-pass PCA projection.

Prompt-embedding diversification. TPSO (Meng et al., 2025) optimizes text-embedding offsets per prompt to place generations in underrepresented embedding regions, preserving fidelity via a trust-region constraint on the original embedding. STRIDE is orthogonal: TPSO modifies *what* the model is asked to generate, while STRIDE modifies *how* the model realizes a fixed prompt. The two are composable.

Principal components and massive activations. Kwon et al. (Kwon et al., 2023) analyze an h-space of UNet bottleneck features and extract attribute directions via PCA for editing. Concept Sliders (Gandikota et al., 2024) similarly learn direction LoRAs. Recent work on diffusion transformers

documents massive activations, where a small number of channels carry disproportionately large magnitudes and govern the effective subspace of the feature map (Gan et al., 2025). Channel-aware low-rank work for quantization (Li et al., 2025) independently documents this low-rank structure from a different motivation. This geometry is what makes a per-image truncated SVD with a small K a cheap and accurate surrogate for the manifold STRIDE must project onto.

Concept Sliders (Gandikota et al., 2024) train low-rank adapters (LoRAs) on attention-output projections of a frozen diffusion model to expose individual semantic axes, e.g., a single LoRA per concept (age, weather, style), each defining one learned direction along which a generation can be moved at inference time. The principal directions are obtained by contrastive optimization between paired prompts, not from the model's own activation statistics. STRIDE uses a related geometric objective (principal directions of a block's activations) but in three different ways: (i) it computes the basis online from a single image's features rather than learning it from a paired-prompt corpus, (ii) it injects noise across the entire principal subspace rather than moving along one direction, and (iii) it requires no training. Concept Sliders modify what the model expresses; STRIDE modifies how the model resolves a single prompt.

HeadHunter (Ahn et al., 2025b) (Ahn et al., NeurIPS 2025) builds a perturbation-guidance (Ahn et al., 2024) design space for DiTs and flow models, identifying which attention heads are responsible for which aspects of the generation. Their SoftPAG variant continuously modulates the PAG strength applied to selected heads. Both require an additional forward pass per timestep (PAG-style guidance difference) and target fidelity with diversity as a secondary observation. STRIDE differs in operating on block-level token features (not per-head attention maps), in projecting noise onto data-driven principal directions of those features (not perturbing attention scores directly), and in working in a single forward pass without a guidance difference. The mechanisms are orthogonal: HeadHunter modifies attention; STRIDE modifies the residual stream feeding into attention.

A.2 Implementation Details

A.2.1 Benchmark Configuration.

We evaluate on four standard text-to-image benchmarks spanning complementary evaluation axes. Table S1 summarizes the per-model generation settings.

MS-COCO Lin et al. (2014) provides large-scale diversity and distributional evaluation, with captions covering 80 everyday object categories across natural photographic scenes. We sample 2,000 captions from the 2014 validation split, generating 10 images per prompt (20,000 images total per method) to enable reliable per-prompt Vendi Score estimation and KID computation against the real image distribution.

DrawBench Saharia et al. (2022) (199 prompts) tests compositionally challenging descriptions across 11 categories including colours, counting, conflicting interactions, descriptions, misspellings, positional relations, rare words, Reddit-sourced creative prompts, and text rendering. We generate 4 images per prompt (796 images per method).

PartiPrompts Yu et al. (2022) (1,632 prompts) provides broad category coverage spanning 12 semantic categories (animals, vehicles, food, indoor/outdoor scenes, people, world knowledge, illustrations, etc.) and 11 challenge aspects from basic descriptions to complex compositionality and writing/symbols. We generate 4 images per prompt (6,528 images per method), offering the largest prompt-level statistical power among our benchmarks.

GenEval Ghosh et al. (2023) (553 prompts) evaluates compositional accuracy through structured prompts across six tasks: single object, two objects, counting (2 to 4), colours (10 types), 2D positional relations (left/right/above/below), and colour and object attribute binding, with objects sourced from COCO categories. We generate 4 images per prompt (2,212 images per method) at 512×512 resolution, following the benchmark's standard evaluation protocol.

All models generate at native resolution (1024×1024 for both FLUX.1-schnell and SD3.5 Large Turbo), with the exception of FLUX GenEval, which we evaluate at 512×512 following common practice to match Harrington et al. (2025). SD3.5 GenEval remains at 1024×1024. We use a single denoising step for FLUX.1-schnell (1-step model) and 4 steps for SD3.5 Large Turbo following its distillation schedule, with guidance scale 0.0 for both models as recommended in their official

Table S1: **Per-benchmark and per-model generation configuration.** The top block reports the prompt count, images-per-prompt, and total images generated for each of the four benchmarks. The lower blocks list the per-model native resolution, denoising-step count, and maximum tokenizer sequence length. All experiments use classifier-free guidance scale 0 and a fixed base seed s=8 with per-variant offsets (s+1, s+2, ...) for reproducibility. The only resolution exception is FLUX GenEval, evaluated at 512^2 to match the protocol of (Harrington et al., 2025).

Setting	COCO	DrawBench	PartiPrompts	GenEval
Benchmark size				
Prompts	2,000	199	1,632	553
Images / prompt	10	4	4	4
Total images	20,000	796	6,528	2,212
FLUX.1-schnell (12B parameters, MM-DiT, 1 step)				
Resolution	1024^2	1024^2	1024^2	512^2
Denoising steps	1	1	1	1
Max seq. length	256	256	256	256
SD3.5 Large Turbo (8.1B parameters, MM-DiT, 4 steps)				
Resolution	1024^2	1024^2	1024^2	1024^2
Denoising steps	4	4	4	4
Max seq. length	256	256	256	256

configurations. All experiments use a fixed base seed (s=8) with per-variant offsets (s+1, s+2, ...) for reproducibility while ensuring distinct samples per prompt.

A.2.2 Baseline Implementation Details

All baselines are training-free and operate at inference time only. For each baseline, we conduct a grid search on DrawBench (199 prompts $\times$ 4 images) to identify Pareto-optimal hyperparameters balancing diversity (InBSim↓) and quality (CLIP↑), then evaluate the selected configuration across all four benchmarks. Table S3 summarises the final hyperparameters for every method-model combination.

CADS Sadat et al. (2023). We reimplement Condition-Annealed Diffusion Sampling following the original formulation. For the single-step model (FLUX.1-schnell), the per-timestep annealing schedule $\gamma(t) = (\tau_2 - t)/(\tau_2 - \tau_1)$ collapses to a single scalar noise level: the schedule endpoints τ_1, τ_2 have no effect when only one denoising step is taken. We therefore treat γ and s as direct hyperparameters and ablate them manually to identify Pareto-optimal configurations (see supplementary Table S16). The noisy embedding is computed as $\mathbf{y}' = \sqrt{\gamma}\,\mathbf{y} + s\sqrt{1-\gamma}\,\boldsymbol{\epsilon}$, where γ controls signal retention, s controls noise magnitude, and rescaling (ψ) preserves the original embedding mean and variance. For FLUX.1-schnell, noise is applied to the T5 sequence embeddings only (not pooled embeddings). For SD3.5 Large Turbo (4-step), CADS operates as originally designed: the linear annealing schedule between τ_1 and τ_2 modulates conditioning noise across denoising steps, with γ varying smoothly from full noise at $t \leq \tau_1$ to no noise at $t \geq \tau_2$. For SD3.5 Large Turbo we inject noise into the concatenated CLIP+T5 sequence embeddings at each step. Notably, CADS **collapses on SD3.5 Large Turbo** at every tested configuration combinations (see Table S23).

Input Noise (Harrington et al. Harrington et al. (2025)). We reimplement the pink latent noise approach from Harrington et al. Harrington et al. (2025) as the original code is not publicly available. Following their pink-noise formulation, we blend frequency-shaped noise into the initial Gaussian latent: $\mathbf{z}' = (1-\alpha)\,\mathbf{z} + \alpha\,\mathbf{z}_{\text{pink}}$, renormalized to unit variance. The pink noise is generated via 2D FFT with a radial frequency filter $1/(1+f)^{f_\alpha}$, where f_α controls the frequency roll-off exponent (their "α" parameter; we denote it f_α to avoid confusion with our perturbation strength α). Noise is generated independently per sample.

For FLUX.1-schnell, the latent is in packed format ($B \times N \times D_{\text{latent}}$) and is unpacked to spatial dimensions before pink noise generation, then repacked. For SD3.5 Large Turbo, latents are in standard (B, C, H, W) format and noise is applied directly. We verified our implementation by confirming that f_α=0 (white noise) produces no diversity change at any α, matching the expected behaviour that isotropic noise is absorbed by the model's normalisation layers.

The optimal α and f_α differ substantially across models due to sensitivity to the number of denoising steps: single-step models tolerate higher α, while SD3.5 Turbo's 4-step schedule amplifies latent perturbations and requires lower α with higher f_α. See Tables S15 and S22

Concurrent Noise Optimization (Harrington et al. Harrington et al. (2025)). For our GenEval-FLUX comparison rows in Table 4 (*-marked), we report numbers directly from Table 5 of Harrington et al. (2025) as their code is not publicly available. Their evaluation matches ours on the dimensions that matter most for fair comparison: FLUX.1-schnell at 512×512 resolution, the same 553 GenEval prompts, 4 images per prompt, and DINOv2-based set diversity metrics (DreamSim, Vendi) together with HPSv2 for image quality. The three *-marked rows correspond to: *i.i.d.* (white-noise sampling with no intervention), **GI** (*Group Inference*; their reproduction of Parmar et al. Parmar et al. (2025), which reduces a candidate pool of 64 to 4 via group inference), and **NoiseOpt** (*Noise Optimization*; their proposed end-to-end optimization of *white-noise* latents with the DPP (*Determinantal Point Process*) diversity objective and a CLIP-HPSv2 (*Human Preference Score v2*) quality reward). A key methodological distinction: NoiseOpt optimizes the *initial white-noise latent* via backpropagation, whereas STRIDE injects *pink noise* (with frequency-shaped power spectrum) into intermediate transformer features along PCA-aligned directions in a single forward pass without any optimization loop or backpropagation through the diffusion model. NoiseOpt requires **80 backpropagation iterations** per prompt (Harrington et al., 2025, Sec. 4.2), whereas STRIDE operates in the single forward pass for generation. The comparison is necessarily approximate; our purpose is not to claim a precise comparison, but to demonstrate that STRIDE attains diversity gains *in the same regime* (DreamSim 0.423 vs. 0.446; Vendi 2.428 vs. 2.753) under a fundamentally different mechanism.

Diversity Distillation (Gandikota & Bau Gandikota & Bau (2026) We compare against the inference-time diversity recovery method proposed for few-step distilled models, applied here to SD3.5 Large Turbo with the SD3.5 Large checkpoint as the non-distilled base. This baseline (Hybrid Inference, Algorithm 1 in (Gandikota & Bau, 2026), k=1) replaces the distilled model's first denoising step with a single step from the non-distilled base model, while keeping the turbo's 4-step noise schedule unchanged. Concretely, the base transformer is invoked with the turbo scheduler's first timestep value, and the remaining three steps use the distilled turbo transformer. Both transformers run at guidance scale 0 as discussed in (Gandikota & Bau, 2026), finding that natural diversity is preserved without additional steering. Text encoders, tokenisers, VAE, and scheduler are shared between the two checkpoints (both inherit the SD3 family backbone), so only the transformer module is swapped per timestep. We evaluate under our standard SD3.5 Turbo protocol (1024×1024, 4 turbo timesteps, base seed s=8 with per-variant offsets, native CFG=0). Hyperparameters use the paper-recommended optimum.

SPELL (Kerrigan et al. Kirchhof et al. (2024)). We reimplement Sparse Repellency following Algorithm 1 and the parallelised pseudocode in Appendix D of (Kirchhof et al., 2024). SPELL is applied to SD3.5 Large Turbo only; we exclude it from the FLUX.1-schnell evaluation as the single-step schedule reduces SPELL to a one-shot post-prediction adjustment with no opportunity for the iterative repellency dynamics the method is designed for.

We use SPELL in its dynamic batch-coupled mode, with no static reference set. At each denoising step t and for each particle i, The shielded set $\mathcal{S}_t^{(i)} = \{\hat{\mathbf{x}}_0^{(j)} : j \neq i\}$ is rebuilt from the predicted clean endpoints of the other particles in the batch. If any $\hat{\mathbf{x}}_0^{(j)} \in \mathcal{S}_t^{(i)}$ falls within Euclidean distance r of $\hat{\mathbf{x}}_0^{(i)}$, the latter is projected radially outward to the boundary of the r-ball, and the corresponding correction is propagated back to $\mathbf{x}_t^{(i)}$ through the scheduler's denoising update. Outside this threshold, the repellency is exactly zero (the "sparse" regime), producing no intervention on already-diverse batches and tapering off naturally as the predicted endpoints sharpen towards the end of the trajectory.

Table S2: **Token PCA vs. Spatial PCA on FLUX.1-schnell:** cross-prompt alignment of principal directions (1.0 = identical subspaces). For each layer, we compute the principal subspace similarity between PCA bases extracted from different prompts and average across prompt pairs. Token directions remain $\sim$0.81-aligned across prompts, while spatial directions drop to $\sim$0.32, a $2.5\times$ gap. Channel space is largely prompt-invariant while spatial-channel correlations are prompt-dependent.

Method	**Top-k**	L0	L1	L2	L3	L4	L5	L6	L7	**Mean $\pm$ Std**
Token (D=3072)	5	0.987	0.971	0.954	0.970	0.988	0.979	0.988	0.981	0.977 ± 0.011
Token (D=3072)	16	0.941	0.895	0.878	0.904	0.927	0.944	0.932	0.960	0.923 ± 0.026
Token (D=3072)	64	0.863	0.818	0.810	0.790	0.822	0.765	0.854	0.825	0.818 ± 0.030
Spatial (D'=12288)	5	0.990	0.930	0.996	0.987	0.975	0.994	0.988	0.969	0.979 ± 0.020
Spatial (D'=12288)	16	0.996	0.510	0.771	0.818	0.832	0.860	0.851	0.816	0.807 ± 0.128
Spatial (D'=12288)	64	0.446	0.246	0.296	0.328	0.335	0.350	0.349	0.337	0.336 ± 0.053

Distances are computed in latent space on the standard (B, C, H, W) SD3.5 Turbo latent. The original paper reports $r = 200$, $\lambda = 1.6$ for SD3-Medium at 256×256 on CC12M with 28 denoising steps and CFG=5.5 (Kirchhof et al., 2024, App. D); these conditions differ from ours along four axes (Turbo distillation, 4-step schedule, CFG=0, 1024×1024 resolution), so both hyper-parameters require recalibration. We follow the calibration protocol of (Kirchhof et al., 2024, App. D): we generate one batch without repellency, measure the median pairwise $\|\hat{\mathbf{x}}_0^{(i)} - \hat{\mathbf{x}}_0^{(j)}\|_2$ between the predicted clean endpoints at the final denoising step (grand median M=391.9; Table S25), and sweep $r \in \{0.5M, 1.0M, 1.5M, 2.0M\}$ paired with the λ values explicitly studied in the original paper Kirchhof et al. (2024). The Pareto-optimal configuration (r=588, λ=1.0; Table S26) is the strongest diversity gain that retains CLIP > 26. The projection is applied at every denoising step (4 of 4 for SD3.5 Turbo).

STRIDE No-PCA (Uniform Pink Feature Noise). STRIDE No-PCA adds unstructured pink noise directly to transformer hidden states *without* spatial PCA projection: $\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{n}_{\text{pink}}$. This serves as the critical "without on-manifold projection" ablation of our method. Pink noise is generated per-layer using the same 2D FFT procedure as the input noise baseline, reshaped to match the token grid dimensions. STRIDE No-PCA uses the same layer selection and timestep gating as STRIDE for each model (Table S3), enabling a controlled comparison where the *only* difference is the presence or absence of PCA projection.

STRIDE (PCA-Directed Spatial Feature Perturbation). Our proposed method. Full details are in Section 3. Hyperparameters are model-specific (Table S3) and were selected via staged ablation: first layer selection and timestep gating, then α and f_α jointly. For all models, we use online PCA (no precomputed components), spatial patch size P=2, and singular-value-weighted projection ($S/\bar{S}$ scaling).

A.2.3 STRIDE Configuration Details

Grid sweeps. For each model, α and f_α are swept over the grids in Tables S13 and S20, with $K = 1024$. The layer set chosen by the ablations in Sec. A.3. The hyperparameter settings are discussed in Table S4

A.2.4 Human Survey Details

Setup. We conducted a two-alternative forced-choice (2AFC) study to assess whether the diversity gains reported by quantitative metrics (InBSim,Vendi/p, DreamSim) align with human perception. The study consists of 24 prompts randomly sampled from DrawBench Saharia et al. (2022), chosen to cover the benchmark's range of difficulty: simple object descriptions ("A black colored dog"), compositional prompts ("A banana on the left of an apple"), text rendering, rare object combinations, and creative scenes. For each prompt, we generated 4 images using the FLUX.1-schnell baseline (CFG = 0, distinct seeds) and 4 images using STRIDE (same seeds, our configuration: α=0.5, f_α=5, K=1024, layers L0–7).

Table S3: **Hyperparameter configurations** for all methods across both models. Each configuration was selected via grid search on DrawBench to identify the Pareto-optimal diversity-quality operating point, except where noted. †CADS collapses on SD3.5 Turbo at all tested configurations (CLIP<23); reported config is the least degraded. ‡Hybrid Inference uses paper-recommended optima (k=1, base CFG=0). §SPELL is applied to SD3.5 Turbo only.

Method	Parameter	FLUX.1-schnell	SD3.5 Large Turbo
CADS	γ / τ_1	γ=0.9	τ_1=0.8†
	s / τ_2	s=0.5	s=0.15 / τ_2=0.9†
	ψ	0.9	1.0†
	Target embeddings	T5 sequence	CLIP+T5 sequence†
Input Noise	α (blend)	0.2	0.1
	f_α (freq. exp.)	5.0	10.0
Hybrid Inference‡	Base-model steps (k)	—	1 (step 0 only)
	Base CFG scale	—	0.0
	Turbo CFG scale	—	0.0
	Schedule	—	turbo 4-step (n=4)
SPELL§	Repellency radius r	—	588
	Overcompensation λ	—	1.0
	Reference set	—	intra-batch (dynamic)
	Distance space	—	latent (B, C, H, W)
	Steps with projection	—	all (4 of 4)
STRIDE No-PCA	α	0.5	0.1
	f_α	5.0	20.0
	Layers	L0–7	L0–11
	Steps	all (1)	S0 only
STRIDE	α	0.5	0.1
	f_α	5.0	20.0
	K (PCA rank)	1024	1024
	P (patch size)	2	2
	Layers	L0–7	L0–11
	Steps	all (1)	S0 only
	Singular Value scaling	$S/\bar{S}$	$S/\bar{S}$
	Stride(S)	S=1	S=1

Table S4: STRIDE hyperparameters per model. All configurations use spatial patch size $P = 2$, full-rank K, and singular value scaling ($S/\bar{S}$).

Model	Layer set	K	α	f_α	Steps
FLUX.1-schnell	L0–7	1024	0.5	5	1
SD3.5 Large Turbo	L0–11	1024	0.1	20	4 (S0 only)

Interface and protocol. Participants accessed the survey through a custom web application built on Flask with a SQLite backend. The landing page (Figure S1) explains the task, defines "diverse" (*"different poses, viewpoints, compositions, lighting, styles, or interpretations of the prompt"*), and obtains voluntary consent before participation. Each question (Figure S2) presents two image grids labelled *Set A* and *Set B*, each containing 4 images generated for the same prompt. The assignment of baseline/STRIDE to A/B is **randomized independently per question**, so participants cannot infer which method is which from positional cues. Participants are asked: *"Which set of images is more diverse?"*, and select either Set A or Set B. A 5-point confidence scale is shown alongside each question (*1: Unsure / Guessing* to *5: Absolutely certain*).

Recruitment and response accounting. The survey was distributed through academic mailing lists and social media. **All 119 participants** who answered at least one question contribute to the analysis: **50** completed all 24 questions, and **69** provided partial responses before closing the survey. Each recorded answer counts as one response regardless of whether the participant later finished the full study. The total response count is 1,370, with per-prompt counts varying from 52 to 61 due to

participant drop-off and individual question skips. Responses are anonymous, and no personal data was collected.

Per-prompt results. Table S5 reports the response distribution and binomial significance for each of the 24 prompts. STRIDE was preferred on **22 of 24** prompts (overall preference rate $930/1{,}370 = 67.9\%$, 95% Wilson CI $[65.4\%, 70.3\%]$, binomial $p < 10^{-39}$ against $H_0 : p{=}0.5$), with one tie. Of the 22 STRIDE wins, **18** reach significance at $p < 0.05$. The remaining 4 STRIDE wins that fall short of significance involve compositional prompts with strict spatial constraints (e.g., *"A banana on the left of an apple"*, *"A stack of 3 cubes. . . "*) where both methods produce visually similar layouts because the prompt admits limited variation.

Table S5: Per-prompt human preference results. STRIDE % denotes the proportion of participants choosing STRIDE over baseline. *indicates $p < 0.05$ under a two-sided binomial test against $H_0 : p{=}0.5$. STRIDE wins on 22/24 prompts (18 at $p < 0.05$), with one tie (#13); the single significant baseline win (#20, *Mario*) is discussed above.

#	Prompt	n	Base	STRIDE	STRIDE %
0	A black colored dog.	59	18	41	69.5%*
1	A sphere made of kitchen tile. A sphere with the texture of kitchen tile.	59	14	45	76.3%*
2	A large keyboard musical instrument with a wooden case enclosing a soundboard and metal strings, which are struck by hammers when the keys are depressed. The strings' vibration is stopped by dampers when the keys are released and can be regulated for length and volume by two or three pedals.	58	21	37	63.8%*
3	A yellow colored giraffe.	60	13	47	78.3%*
4	A storefront with 'Hello World' written on it.	61	21	40	65.6%*
5	Painting of Mona Lisa but the view is from behind of Mona Lisa.	55	10	45	81.8%*
6	A cube made of denim. A cube with the texture of denim.	60	20	40	66.7%*
7	An ancient Egyptian painting depicting an argument over whose turn it is to take out the trash.	56	13	43	76.8%*
8	A separate seat for one person, typically with a back and four legs.	52	15	37	71.2%*
9	A banana on the left of an apple.	52	21	31	59.6%
10	A stack of 3 cubes. A red cube is on the top, sitting on a red cube. The red cube is in the middle, sitting on a green cube. The green cube is on the bottom.	56	22	34	60.7%
11	A spider with a moustache bidding an equally gentlemanly grasshopper a good day during his walk to work.	56	11	45	80.4%*
12	New York Skyline with 'Hello World' written with fireworks on the sky.	57	21	36	63.2%
13	A car playing soccer, digital art.	56	28	28	50.0%‡
14	An illustration of a large red elephant sitting on a small blue mouse.	60	17	43	71.7%*
15	Photo of an athlete cat explaining its latest scandal at a press conference to journalists.	60	21	39	65.0%*
16	New York Skyline with 'Deep Learning' written with fireworks on the sky.	55	12	43	78.2%*
17	A pink colored giraffe.	55	15	40	72.7%*
18	A realistic photo of a Pomeranian dressed up like a 1980s professional wrestler with neon green and neon orange face paint and bright green wrestling tights with bright orange boots.	58	17	41	70.7%*
19	Three dogs on the street.	57	12	45	78.9%*
20	A real life photography of super mario, 8k Ultra HD.	57	38	19	33.3%*†
21	A long curved fruit which grows in clusters and has soft pulpy flesh and yellow skin when ripe.	57	18	39	68.4%*
22	A stop sign on the right of a refrigerator.	60	19	41	68.3%*
23	A connection point by which firefighters can tap into a water supply.	54	23	31	57.4%
Total		**1,370**	**440**	**930**	**67.9%***

†Baseline-preferred (canonical fictional character; see text). ‡Tie (28–28).

A.2.5 Statistical Significance

Rather than repeating experiments across multiple global seeds (prohibitively expensive for 8–12B parameter models generating 20,000+ images per configuration), we report per-prompt standard deviation across the prompt population (Table S8). With 2,000 prompts on COCO and 1,632 on PartiPrompts, the standard error of mean InBSim is below ± 0.005 for all methods, making the reported differences (ΔInBSim ≥ 0.04) statistically significant at $p \ll 0.001$. On the smallest benchmark (DrawBench, 199 prompts), the standard error is ± 0.014, and all STRIDE improvements ($\Delta \geq 0.068$) exceed $5\times$ the SE. The human evaluation (Section A.2.4) provides independent confirmation ($p < 10^{-39}$).

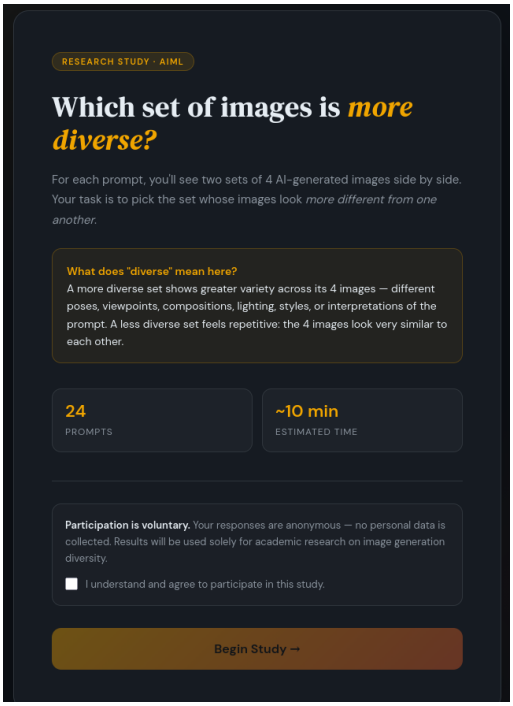


Figure S1: Survey landing page. Participants read task instructions, the diversity definition, and consent before beginning.

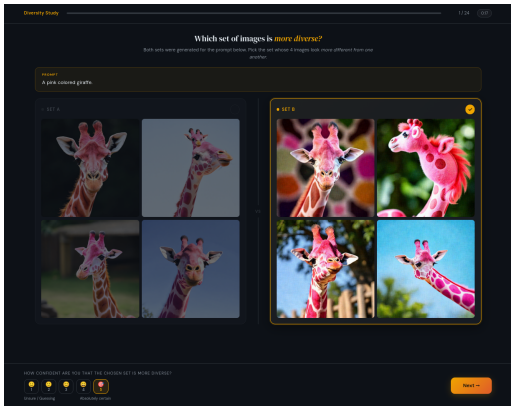


Figure S2: A single survey question. Set A and Set B contain 4 images each from baseline and STRIDE in randomized order; participants pick which set is more diverse and rate confidence on a 5-point scale.

Table S6: Human evaluation (2AFC diversity preference). All 119 participants who answered at least one question contributed; the 50 who completed all 24 questions and 69 partial sessions together produced 1,370 valid responses.

	Preference (%)
STRIDE	**67.9** (930/1,370)
Baseline (i.i.d)	32.1 (440/1,370)
$p < 10^{-39}$ (binomial); 95% Wilson CI $[65.4, 70.3]$	
119 participants (50 complete, 69 partial), 24 prompts, 1,370 responses	
STRIDE preferred on 22/24 prompts (18 at $p < 0.05$); one tie	

Algorithm S1 STRIDE: spatial-grouped PCA-directed feature perturbation

Require: Hidden state $\mathbf{H} \in \mathbb{R}^{1\times N\times D}$ at chosen transformer block; PCA rank K; patch size P; perturbation strength α; pink-noise exponent f_α

$\mathbf{n} \leftarrow \textsc{PinkNoise}(N, D;\, f_\alpha)$ ▷ Eq. 4
$\mathbf{H}_p,\, \mathbf{n}_p \leftarrow \textsc{Patchify}(\mathbf{H}, P),\ \textsc{Patchify}(\mathbf{n}, P)$ ▷ Eq. 2
$\widetilde{\mathbf{H}}_p \leftarrow \mathbf{H}_p - \overline{\mathbf{H}}_p$ ▷ centre patches
$(\mathbf{U}, \mathbf{S}, \mathbf{V}) \leftarrow \textsc{TruncatedSVD}(\widetilde{\mathbf{H}}_p,\, q{=}K)$ ▷ Eq. 3
$\mathbf{s} \leftarrow \mathrm{diag}(\mathbf{S})$ ▷ singular values as vector
$\mathbf{c} \leftarrow \mathbf{n}_p \cdot \mathbf{V}$ ▷ project noise (Eq. 5)
$\mathbf{c} \leftarrow \mathbf{c} \odot (\mathbf{s} \,/\, \bar{s})$ ▷ singular-value scaling (Eq. 5)
$\mathbf{\Delta}_p \leftarrow \mathbf{c} \cdot \mathbf{V}^\top$ ▷ back to patch space (Eq. 5)
$\mathbf{\Delta} \leftarrow \textsc{Unpatchify}(\mathbf{\Delta}_p)$ ▷ $\mathbb{R}^{1\times N\times D}$
return $\mathbf{H} + \alpha\, \mathbf{\Delta}$ ▷ Eq. 1

Multi-seed robustness. To confirm that the reported STRIDE results are not artifacts of a single random seed, we run STRIDE experiment on DrawBench (199 prompts $\times$ 4 images) at three independent seeds (8, 25, 39) for each backbone (Table S9). Cross-seed variation is small across all four metrics: InBSim varies by ≤ 0.006 on FLUX and ≤ 0.002 on SD3.5 Turbo, and CLIP by ≤ 0.20 and ≤ 0.09 respectively. These cross-seed deviations are one to two orders of magnitude smaller than the per-prompt standard errors of ± 0.014 (InBSim) and ± 0.36 (CLIP) on DrawBench reported in Table S8, indicating that the dominant source of measurement uncertainty is the prompt distribution itself rather than seed selection. The mean values across the three seeds match the single-seed numbers in Tables 1 within rounding (e.g., FLUX InBSim 0.592 multi-seed vs. 0.590 single-seed), confirming the reported configuration is robustly reproducible.

Paired per-prompt significance tests. To test the main comparisons directly rather than through standard errors, we run paired per-prompt Wilcoxon signed-rank tests on DrawBench (n=199; Δ = STRIDE $-$ baseline of the same backbone; Holm-corrected within each metric; Table S7). At the locked FLUX.1-schnell configuration, ΔInBSim $= -0.073$ with paired effect size d_z=0.64 and $p{=}6\times10^{-16}$. Diversity improves significantly on every backbone; text alignment is not significantly reduced in any comparison and improves significantly at K=512 on FLUX.1-schnell ($+0.61$); the HPS reductions are small in absolute magnitude, decrease substantially at K=512, and are statistically indistinguishable from zero on FLUX.2-klein-4B.

Table S7: **Paired per-prompt significance tests on DrawBench** (199 prompts $\times$ 4 images; Wilcoxon signed-rank; Δ = STRIDE $-$ baseline of the same backbone; Holm-corrected within each metric). n.s.: not significant after correction. Rows marked K=1024 use the locked per-backbone configuration of Table S4; the FLUX.1-schnell K=512 row uses α=0.5; FLUX.2-klein-4B uses K=512, α=1.5 (App. A.6).

Comparison (n=199, paired by prompt)	**ΔInBSim↓**	**ΔCLIP↑**	**ΔHPS↑**
FLUX.1-schnell +STRIDE (K=1024)	-0.073, $p < 0.001$	$+0.14$, n.s.	-0.0091, $p < 0.001$
FLUX.1-schnell +STRIDE (K=512)	-0.027, $p < 0.001$	$+0.61$, $p < 0.001$	-0.0034, $p < 0.001$
SD3.5 Large Turbo +STRIDE (K=1024)	-0.068, $p < 0.001$	-0.12, n.s.	-0.0029, $p < 0.001$
FLUX.2-klein-4B +STRIDE (K=512)	-0.049, $p < 0.001$	$+0.12$, n.s.	-0.0004, n.s.
HunyuanImage-2.1 +STRIDE (K=1024, α=14)	-0.064, $p < 0.001$	-0.43, $p < 0.01$	-0.0026, $p < 0.001$

Table S8: **Per-prompt standard deviation** (σ_p) and standard error (SE $= \sigma_p/\sqrt{n}$) for all main experiments. All STRIDE vs. Baseline InBSim differences exceed $5\times$ SE, confirming statistical significance without multi-seed repetition.

(a) FLUX.1-schnell

				InBSim		VD/p_{DINO}		CLIP		HPS	
Dataset	**Method**	n_p	ipp	$\mu \pm \sigma_p$	SE	$\mu \pm \sigma_p$	SE	$\mu \pm \sigma_p$	SE	$\mu \pm \sigma_p$	SE
COCO	Baseline	2000	10	$.738 \pm .138$	.003	2.794 ± 1.116	.025	26.69 ± 3.02	.068	$.293 \pm .016$	.000
	CADS	2000	10	$.659 \pm .159$	.004	3.433 ± 1.375	.031	26.44 ± 2.83	.063	$.291 \pm .016$	.000
	Input Noise	2000	10	$.661 \pm .146$	.003	3.443 ± 1.291	.029	26.11 ± 2.87	.064	$.288 \pm .016$	.000
	STRIDE No-PCA	2000	10	$.762 \pm .118$	.003	2.603 ± 0.923	.021	28.05 ± 3.02	.068	$.276 \pm .014$	.000
	STRIDE	2000	10	$.683 \pm .145$	.003	3.268 ± 1.257	.028	27.66 ± 2.95	.066	$.283 \pm .015$	.000
DrawBench	Baseline	199	4	$.662 \pm .188$	.013	2.236 ± 0.636	.045	28.47 ± 5.04	.357	$.293 \pm .021$	.001
	CADS	199	4	$.556 \pm .202$	.014	2.578 ± 0.654	.046	27.09 ± 4.72	.335	$.290 \pm .020$	.001
	Input Noise	199	4	$.566 \pm .186$	.013	2.570 ± 0.624	.044	27.43 ± 4.98	.353	$.284 \pm .019$	.001
	STRIDE No-PCA	199	4	$.673 \pm .181$	.013	2.205 ± 0.615	.044	29.04 ± 4.90	.347	$.280 \pm .018$	.001
	STRIDE	199	4	$.590 \pm .191$	.014	2.496 ± 0.641	.045	28.61 ± 4.83	.342	$.284 \pm .018$	.001
PartiPrompts	Baseline	1632	4	$.722 \pm .171$	.004	2.025 ± 0.585	.014	28.13 ± 4.74	.117	$.290 \pm .018$	.000
	CADS	1632	4	$.630 \pm .197$	.005	2.328 ± 0.652	.016	27.37 ± 4.53	.112	$.288 \pm .018$	.000
	Input Noise	1632	4	$.632 \pm .181$	.004	2.345 ± 0.616	.015	27.31 ± 4.63	.115	$.283 \pm .020$	.000
	STRIDE No-PCA	1632	4	$.744 \pm .150$	.004	1.952 ± 0.515	.013	28.77 ± 4.42	.109	$.273 \pm .018$	.000
	STRIDE	1632	4	$.653 \pm .181$	.004	2.275 ± 0.616	.015	28.56 ± 4.55	.113	$.281 \pm .019$	.000
GenEval	Baseline	553	4	$.669 \pm .164$	.007	2.204 ± 0.560	.024	28.34 ± 3.52	.150	$.296 \pm .012$	.000
	CADS	553	4	$.559 \pm .191$	.008	2.550 ± 0.627	.027	27.34 ± 3.49	.148	$.290 \pm .013$	.001
	Input Noise	553	4	$.592 \pm .169$	.007	2.476 ± 0.572	.024	27.32 ± 3.40	.145	$.287 \pm .012$	.000
	STRIDE No-PCA	553	4	$.696 \pm .149$	.006	2.123 ± 0.517	.022	29.26 ± 3.37	.143	$.278 \pm .011$	.000
	STRIDE	553	4	$.607 \pm .165$	.007	2.428 ± 0.557	.024	29.04 ± 3.42	.145	$.282 \pm .011$	.000

(b) SD3.5 Large Turbo

				InBSim		VD/p_{DINO}		CLIP		HPS	
Dataset	**Method**	n_p	ipp	$\mu \pm \sigma_p$	SE	$\mu \pm \sigma_p$	SE	$\mu \pm \sigma_p$	SE	$\mu \pm \sigma_p$	SE
COCO	Baseline	2000	10	$.771 \pm .130$	.003	2.527 ± 1.016	.023	26.77 ± 3.13	.070	$.290 \pm .015$	.000
	CADS†	2000	10	$.391 \pm .147$	.003	3.169 ± 0.627	.014	22.56 ± 1.33	.030	$.263 \pm .020$	.000
	Input Noise	2000	10	$.749 \pm .135$	.003	2.703 ± 1.073	.024	26.72 ± 3.13	.070	$.291 \pm .014$	.000
	DDC	2000	10	$.679 \pm .152$	.003	3.301 ± 1.316	.029	26.70 ± 2.81	.063	$.275 \pm .013$	.000
	SPELL	2000	10	$.603 \pm .111$	.002	3.971 ± 1.037	.023	24.44 ± 3.02	.067	$.248 \pm .016$	.000
	STRIDE No-PCA	2000	10	$.766 \pm .131$	.003	2.573 ± 1.021	.023	26.82 ± 3.11	.070	$.289 \pm .015$	.000
	STRIDE	2000	10	$.725 \pm .145$	.003	2.905 ± 1.193	.027	26.69 ± 3.05	.068	$.289 \pm .015$	.000
DrawBench	Baseline	199	4	$.707 \pm .188$	.013	2.067 ± 0.627	.044	28.18 ± 5.09	.361	$.291 \pm .021$	.001
	CADS†	199	4	$.318 \pm .188$	.013	2.323 ± 0.562	.040	22.67 ± 1.05	.074	$.262 \pm .022$	.002
	Input Noise	199	4	$.668 \pm .193$	.014	2.203 ± 0.647	.046	28.34 ± 5.11	.362	$.292 \pm .020$	.001
	DDC	199	4	$.600 \pm .200$	.014	2.436 ± 0.654	.046	27.84 ± 4.97	.352	$.274 \pm .017$	.001
	SPELL	199	4	$.638 \pm .160$	.011	2.334 ± 0.546	.039	26.32 ± 4.73	.335	$.273 \pm .017$	.001
	STRIDE No-PCA	199	4	$.704 \pm .182$	.013	2.082 ± 0.617	.044	28.07 ± 5.19	.368	$.291 \pm .020$	.001
	STRIDE	199	4	$.639 \pm .201$	.014	2.300 ± 0.665	.047	28.07 ± 5.18	.367	$.288 \pm .019$	.001
PartiPrompts	Baseline	1632	4	$.769 \pm .163$	.004	1.855 ± 0.554	.014	28.09 ± 4.54	.112	$.289 \pm .018$	.000
	CADS†	1632	4	$.356 \pm .183$	.005	2.268 ± 0.537	.013	22.88 ± 0.99	.025	$.257 \pm .022$	.001
	Input Noise	1632	4	$.740 \pm .171$	.004	1.959 ± 0.582	.014	28.02 ± 4.53	.112	$.288 \pm .018$	.000
	DDC	1632	4	$.675 \pm .181$	.004	2.187 ± 0.509	.013	27.70 ± 4.38	.108	$.273 \pm .016$	.000
	SPELL	1632	4	$.690 \pm .149$	.004	2.157 ± 0.522	.013	26.69 ± 4.29	.106	$.269 \pm .019$	.000
	STRIDE No-PCA	1632	4	$.766 \pm .157$	.004	1.870 ± 0.541	.013	28.00 ± 4.59	.114	$.287 \pm .018$	.000
	STRIDE	1632	4	$.714 \pm .174$	.004	2.051 ± 0.593	.015	27.91 ± 4.55	.113	$.286 \pm .018$	.000
GenEval	Baseline	553	4	$.741 \pm .160$	.007	1.948 ± 0.543	.023	29.20 ± 3.36	.143	$.291 \pm .012$	.001
	CADS†	553	4	$.299 \pm .161$	.007	3.348 ± 0.493	.021	22.82 ± 3.63	.154	$.258 \pm .013$	.001
	Input Noise	553	4	$.701 \pm .161$	.007	2.087 ± 0.539	.023	28.92 ± 3.37	.143	$.291 \pm .012$	.001
	DDC	553	4	$.627 \pm .180$	.008	2.343 ± 0.584	.025	28.52 ± 3.10	.132	$.273 \pm .011$	.000
	SPELL	553	4	$.651 \pm .163$	.007	2.284 ± 0.558	.024	27.88 ± 3.18	.135	$.272 \pm .013$	.001
	STRIDE No-PCA	553	4	$.730 \pm .159$	.007	1.986 ± 0.537	.023	29.13 ± 3.40	.144	$.290 \pm .012$	.001
	STRIDE	553	4	$.679 \pm .169$	.007	2.175 ± 0.578	.025	28.86 ± 3.37	.143	$.288 \pm .012$	.001

Notes: n_p = number of prompts; ipp = images per prompt; σ_p = standard deviation across prompts; SE $= \sigma_p/\sqrt{n_p}$. †CADS collapses CLIP by $\sim$5–6 points on SD3.5 across all benchmarks.

Table S9: **Multi-seed robustness of STRIDE on DrawBench** (199 prompts × 4 images). Three independent seeds per backbone with the locked main-paper configuration as per Table S3). Cross-seed variation is one to two orders of magnitude smaller than the per-prompt standard errors reported in Table S8.

Backbone	Seed	InBSim↓	VD/p↑	CLIP↑	HPS↑
FLUX.1-schnell	8	0.5900	2.496	28.61	0.2841
	25	0.5961	2.476	28.73	0.2841
	39	0.5910	2.491	28.81	0.2849
	mean±*std*	0.592 ± 0.003	2.488 ± 0.008	28.72 ± 0.08	0.2844 ± 0.0004
SD3.5 Turbo	8	0.6389	2.300	28.07	0.2881
	25	0.6405	2.307	28.10	0.2882
	39	0.6398	2.301	28.16	0.2892
	mean±*std*	0.640 ± 0.001	2.303 ± 0.003	28.11 ± 0.04	0.2885 ± 0.0005

A.3 Extended Ablation and Analysis

This section reports the per-architecture ablation grids that select the locked configurations used in the main paper. We sweep PCA solver (Table S10), layer set, timestep gating, PCA rank K, perturbation strength α, and frequency exponent f_α on FLUX.1-schnell and SD3.5 Large Turbo, and contrast STRIDE against STRIDE No-PCA, CADS, input-noise, and SPELL on the same protocol. We further report inference latency and peak GPU memory across methods in Table S11, and the K–α grids that motivate the recommended lower-cost operating points (Tables S17 and S24).

Table S10: PCA solver ablation. STRIDE (Spatial PCA, P=2), α=0.5, FLUX.1-schnell, DrawBench 199 prompts × 4 images. All solvers produce comparable results, with the default randomized SVD (n_{iter}=2) offering the best quality-diversity tradeoff at lowest compute cost.

PCA Solver	InBSim↓	CLIP↑	HPS↑	V_D/p↑
Randomized SVD (n_{iter}=2, default)	0.529	27.66	0.275	2.697
Randomized SVD (n_{iter}=5)	0.524	27.35	0.273	2.722
Randomized SVD + oversample (q=K+32)	0.529	27.52	0.273	2.706
Exact SVD (`torch.linalg.svd`)	**0.510**	27.17	0.273	2.764

Inference cost. We measure wall-clock time and peak GPU memory for all methods on a single NVIDIA A100-SXM4-40GB at the model's native resolution and step count, batch size 1 at 1024×1024 with one image per prompt, synchronising CUDA before each measurement and discarding the first three generations as warm-up; Table S11 reports mean and standard deviation over ≥ 50 prompts. Memory is unaffected: peak VRAM stays within 0.4 GB of the baseline at every K. Wall-clock overhead is the one-time cost of the truncated SVD per hooked block and scales with K and the number of hooked blocks: K=256 adds +35%/+29% (FLUX.1-schnell/SD3.5 Turbo), K=512 adds +90%/+76%, and K=1024, used in the main results to maximise diversity, adds +195%/+167%.

PCA rank K vs perturbation strength α STRIDE's diversity gains depend jointly on PCA rank K, perturbation strength α, and spectral exponent f_α. The K–α ablations (Tables S17 and S24) show that increasing α can compensate for reduced rank. On FLUX.1-schnell, K=512, α=0.8 achieves InBSim 0.608, retaining 76% of the reduction obtained with the fixed K=1024, α=0.5 configuration (0.590), with higher CLIP (28.98 vs. 28.61) and less than half the runtime overhead (+89.5% vs. +195.1%). On SD3.5 Turbo, K=512, α=0.2 yields comparable diversity to the fixed K=1024, α=0.1 configuration (InBSim 0.633 vs. 0.639; V_D/p 2.326 vs. 2.300), with CLIP differing by less than 0.2 and overhead reduced from +166.8% to +76.3%. These results support lower ranks with stronger perturbations under limited compute budgets, although the appropriate strength depends on both backbone and rank: on SD3.5 Turbo, α=0.2 at K=1024 reduces CLIP to 21.96.

Table S11: **Inference cost** on NVIDIA A100-40GB, batch 1, 1024×1024. Wall-clock is mean±std per image over ≥50 prompts (3 warm-up runs discarded); overhead relative to the baseline. *Resident* memory is the max VRAM held during inference; *co-resident* is the projected total if all weights were on the GPU at once (relevant only for DDC†, which requires CPU offload to fit on 40 GB).

Method	**s/image**	**std**	**Overhead**	$\text{VRAM}_{\text{resident}}$	$\text{VRAM}_{\text{co-resident}}$
(a) FLUX.1-schnell — 1-step					
Baseline	0.706	0.001	—	33.79 GB	33.79 GB
CADS	0.742	0.084	+5.1%	33.79 GB	33.79 GB
STRIDE (K=256)	0.956	0.002	+35.4%	33.82 GB	33.82 GB
STRIDE (K=512)	1.338	0.010	+89.5%	33.83 GB	33.83 GB
STRIDE (K=1024)	2.083	0.005	+195.1%	33.82 GB	33.82 GB
(b) SD3.5 Large Turbo — 4-step					
Baseline	1.160	0.001	—	28.11 GB	28.11 GB
CADS	1.162	0.001	+0.2%	28.12 GB	28.12 GB
DDC†	10.499	0.239	+805.0%	30.76 GB	**56.16 GB**
STRIDE (K=256)	1.498	0.001	+29.1%	28.11 GB	28.11 GB
STRIDE (K=512)	2.045	0.011	+76.3%	28.49 GB	28.49 GB
STRIDE (K=1024)	3.095	0.003	+166.8%	28.11 GB	28.11 GB

Multi-step sensitivity: α vs. f_α. SD3.5 Turbo reveals a qualitatively different tuning landscape than single-step models. While FLUX tolerates α up to 0.5 (Table S13) without quality loss, SD3.5 collapses at $\alpha \geq 0.2$ with moderate f_α(Table S20), because perturbations accumulate across the 4 denoising steps. The solution is to fix α=0.1 (small per-step perturbation) and control diversity entirely through f_α (frequency shaping strength), which scales smoothly from 0.2 to 30 without collapse (Table S20). This decomposition cleanly separates the two control axes: α governs perturbation *magnitude*, while f_α governs perturbation *structure*. Together, they yield a principled adaptation strategy for multi-step models.

A.3.1 FLUX.1-SCHNELL (MMDIT, 1-STEP)

Layer ablation We report the layer ablation for FLUX.1-schnell in Table S12. The single-layer sweep (a) reveals a stark concentration of diversity in the earliest block: injecting at L0 alone reduces InBSim from 0.662 to **0.557**, while every other individual layer (L1–L18) leaves InBSim within 0.003 of baseline. The grouped sweep (b) confirms this asymmetry: the early-block range L0–7 achieves the best diversity-quality trade-off (InBSim = 0.569, CLIP = 28.63, +0.16 over baseline), late blocks L8–18 alone barely move InBSim (0.658), and extending to all 19 layers (L0–18) yields only marginal additional diversity (0.553) at a small CLIP cost (28.29). We therefore adopt **L0–7** as the default layer set for FLUX.1-schnell.

PCA rank ablation Table S17 evaluates PCA rank K and perturbation strength α on FLUX.1-schnell with f_α=5 and layers L0–7 fixed. Two trends emerge. *(i) Diversity gains require sufficient rank.* For $K \leq 128$, all metrics remain within measurement noise of baseline across the evaluated α values: InBSim ranges from 0.650 to 0.665, while CLIP and V_D/p deviate by at most 0.25 and 0.05, respectively. Gains emerge at K=256–512, suggesting that effective perturbation requires broader coverage of the principal activation subspace. *(ii) Rank and perturbation strength jointly control the diversity–quality trade-off.* At K=512, increasing α from 0.4 to 0.8 improves diversity (InBSim 0.647 → 0.608; V_D/p 2.283 → 2.429) with nearly unchanged CLIP (28.99 → 28.98) and a small HPS decline (0.290 → 0.287). At K=1024, the same sweep yields larger diversity gains (V_D/p up to 2.820) at a greater CLIP and HPS cost (HPS 0.270 at α=0.8, vs. baseline 0.294). We select K=1024, α=0.5 for the main results (InBSim 0.590, CLIP 28.61, HPS 0.284), retaining CLIP within 0.4 points of baseline at an HPS cost of 0.010. For lower computational cost, we recommend K=512, α=0.8, which retains 76% of the InBSim reduction at less than half the overhead (Table S11) and at a smaller HPS cost (0.287 vs. 0.284).

STRIDE ablation We sweep α (perturbation strength) and f_α (frequency roll-off exponent) on FLUX.1-schnell at the locked L0–7 configuration in Table S13. Three patterns emerge. *(i) PCA projection enables a wide operating range:* even at α=1.0 with low f_α, the method achieves InBSim =0.591 at CLIP =28.82, a noticeable diversity gain at quality *above* baseline. *(ii) Moderate spectral shaping improves CLIP:* for $\alpha \in \{0.2, 0.5\}$, CLIP rises above baseline (peak CLIP =29.54 at α=0.5, f_α=0.1), indicating that PCA-directed perturbation can break the model's deterministic mode without harming text alignment. *(iii) Aggressive spectral shaping causes degenerate collapse:* at $\alpha \geq 0.5$ with $f_\alpha \geq 8$, InBSim plateaus or *reverses* (e.g., α=1.0: 0.498 $\rightarrow$ 0.616 as f_α rises from 3 to 12) while CLIP drops below 20 . We adopt α=0.5, f_α=5 as the default operating point for FLUX.1-schnell, balancing strong diversity (InBSim =0.510) with maintained quality (CLIP =26.64).

STRIDE No-PCA ablation We perform the same (α, f_α) sweep with STRIDE No-PCA (uniform feature-level noise, no PCA projection) on FLUX.1-schnell in Table S14. The contrast with STRIDE is striking. *(i) STRIDE No-PCA **reduces** diversity at moderate α:* for $\alpha \in [0.5, 1.0]$ in the designed Harrington range ($f_\alpha \leq 1$), InBSim *rises* above baseline (e.g., α=1.0, f_α=0: InBSim =0.744 vs. baseline 0.662). *(ii) CLIP rises above baseline at most settings* (peak =29.06), consistent with the interpretation in App. A.4 that off-subspace noise is displaced from the learned activation distribution and re-grounded against the text condition by subsequent layers, raising CLIP while suppressing sample variation; this is a hypothesis supported by the displacement and attention measurements, not an established mechanism. *(iii) Only at extreme spectral shaping* ($f_\alpha \geq 8$ with high α) does STRIDE No-PCA yield diversity gains, and only by paying substantial CLIP cost (f_α=10, α=1.0: InBSim =0.613 at CLIP =28.06). At matched moderate settings (α=0.5, f_α=5), STRIDE No-PCA yields InBSim =0.666 versus STRIDE's 0.510. STRIDE achieves roughly 4$\times$ the diversity gain at comparable CLIP, isolating PCA-projected, on-manifold perturbation as the mechanism that converts feature-level noise into actual diversity

Noise Baseline ablation We sweep α and f_α for the input-latent noise blend on FLUX.1-schnell in Table S15. *(i) Within the designed Harrington range* ($f_\alpha \leq 1$), the best operating point is α=1.0, f_α=1.0 (InBSim =0.505, CLIP =26.08): meaningful diversity at the cost of -2.3 CLIP. *(ii) Sharp spectral shaping fails at the input:* at $f_\alpha \geq 3$ with $\alpha \geq 0.5$, CLIP collapses below 22 while InBSim simultaneously *reverses* (e.g., α=1.0, f_α=5: InBSim =0.735, well above baseline). *(iii) The viable sharp-regime window is narrow:* only α=0.2, f_α=5 maintains usable quality (InBSim =0.566, CLIP =27.43), and this is the configuration we adopt as the input-noise default. Comparison with STRIDE at matched settings (Table S13, α=0.5, f_α=5: InBSim =0.510, CLIP =26.64) shows that feature-level PCA-projected injection produces stronger diversity at higher CLIP than any input-level configuration, as input-level noise is fundamentally limited because the perturbation must propagate through the full denoising trajectory, where high-magnitude noise destroys fidelity before yielding diversity.

CADS Baseline We ablate CADS hyperparameters on FLUX.1-schnell in Table S16. At T=1, γ and s act as scalar noise-level and noise-magnitude controls, mixing Gaussian noise into the T5 sequence embedding before passing it to the transformer. Diversity gains scale monotonically with noise strength but are accompanied by severe CLIP degradation: even in moderate settings (γ=0.7, s=0.5), collapse CLIP to 22.30, and aggressive settings (γ=0.3, s=2.0) reach InBSim =0.090 at CLIP =11.61, as the model produces low-similarity but visually incoherent outputs. The only configuration retaining usable text alignment is γ=0.9, s=0.5 (InBSim =0.563, CLIP =27.21), which we report in the main results.

Table S12: FLUX.1-schnell layer ablation. STRIDE Spatial PCA, P=2, for efficiency, pink noise. The layer-importance ordering reported here is consistent across both regimes. DrawBench 199 prompts × 4 images. Baseline: InBSim=0.662, CLIP=28.47. **(a)** Individual layers. **(b)** Layer groups.

(a) Single-layer injection

Layer	**InBSim↓**	**CLIP↑**	**HPS↑**
Baseline	0.662	28.47	0.294
L0	**0.557**	28.34	0.280
L1	0.665	28.55	0.293
L2	0.669	28.46	0.294
L3	0.670	28.44	0.293
L4	0.669	28.45	0.293
L5	0.668	28.44	0.293
L6	0.669	28.45	0.293
L7	0.668	28.46	0.292
L8	0.667	28.45	0.293
L9	0.667	28.47	0.294
L10	0.667	28.48	0.293
L11	0.667	28.46	0.293
L12	0.669	28.45	0.293
L13	0.668	28.46	0.293
L14	0.668	28.47	0.294
L15	0.668	28.46	0.293
L16	0.668	28.45	0.293
L17	0.668	28.45	0.293
L18	0.669	28.45	0.293

(b) Multi-layer injection

Layers	**N**	**InBSim↓**	**CLIP↑**	**HPS↑**
Baseline	0	0.662	28.47	0.294
L0–7	**8**	**0.569**	**28.63**	**0.281**
L8–18	11	0.658	28.64	0.292
L0–18	19	0.553	28.29	0.280

Table S13: **FLUX.1-schnell STRIDE ablation over α and f_α.** Spatial PCA, P=2, layers L0–7, DrawBench 199 prompts × 4 images. Each cell shows InBSim↓ / CLIP↑. Baseline: InBSim=0.664, CLIP=28.40. The sweep spans the full spectral range from white noise (f_α=0) through pink ($f_\alpha \approx 5$) to highly low-frequency-biased noise (f_α=12).

	Perturbation strength α					
f_α	0.0	0.1	0.2	0.5	0.7	1.0
Low-frequency regime ($f_\alpha \leq 1$)						
0.0 (white)	.664/28.40	.660/28.67	.654/29.21	.646/29.52	—	.647/29.15
0.1	.664/28.40	.660/28.68	.655/29.21	.644/29.54	—	.644/29.19
0.2	.664/28.40	.660/28.68	.654/29.22	.643/29.53	—	.639/29.21
0.5	.664/28.40	.660/28.70	.655/29.15	.638/29.48	—	.625/29.08
1.0	.664/28.40	.660/28.68	.653/29.11	.636/29.49	—	.591/28.82
High-frequency regime ($f_\alpha \geq 2$)						
2.0	.664/28.40	.660/28.69	.648/29.12	.609/29.20	.576/28.67	.527/26.91
3.0	.664/28.40	.657/28.69	.643/29.14	.575/28.70	.525/27.15	.498/24.58
4.0	.664/28.40	.656/28.72	.630/29.08	.539/27.79	.489/25.43	.508/22.52
5.0	.664/28.40	.653/28.69	.622/29.01	**.510/26.64**	.487/23.54	.518/20.68
6.0	.664/28.40	.650/28.67	.612/28.89	.487/25.63	.486/22.37	.533/19.13
8.0	.664/28.40	.640/28.60	.592/28.67	.470/23.75	.504/20.32	.571/17.08
10.0	.664/28.40	.630/28.55	.573/28.22	.477/22.26	.521/18.81	.603/15.85
12.0	.664/28.40	.623/28.56	.552/27.84	.482/21.12	.543/17.73	.616/15.26

Table S14: **FLUX.1-schnell STRIDE No-PCA ablation over α and f_α.** Uniform feature-level noise without PCA projection, layer L0-7, DrawBench 199 prompts $\times$ 4 images. Each cell shows InBSim↓ / CLIP↑. Baseline: InBSim=0.662, CLIP=28.37. In contrast to STRIDE (Table S13), feature-level perturbation *without* PCA **increases** InBSim.

	Perturbation strength α					
f_α	0.0	0.1	0.2	0.5	0.7	1.0
Low-frequency regime ($f_\alpha \leq 1$, designed Harrington range)						
0.0 (W)	.662/28.37	.659/28.73	.662/28.92	.689/28.99	.701/29.06	.744/28.61
0.1	.662/28.37	.659/28.75	.662/28.94	.688/29.01	—	.743/28.63
0.2	.662/28.37	.660/28.75	.662/28.95	.688/29.00	—	.742/28.66
0.5	.662/28.37	.660/28.78	.661/28.93	.689/28.99	—	.743/28.74
0.7	.662/28.37	.660/28.78	.662/28.92	.689/28.97	.701/29.06	.742/28.77
1.0	.662/28.37	.660/28.79	.662/28.89	.686/28.93	—	.740/28.80
High-frequency regime ($f_\alpha \geq 2$)						
2.0	.662/28.37	.659/28.78	.663/28.89	.683/28.89	.695/29.06	.732/28.86
3.0	.662/28.37	.659/28.79	.662/28.94	.679/28.88	.691/29.02	.720/28.88
4.0	.662/28.37	.659/28.84	.662/29.00	.673/28.88	.684/29.05	.710/28.84
5.0	.662/28.37	.658/28.85	.663/29.04	.666/28.91	.678/28.98	.696/28.82
6.0	.662/28.37	.657/28.88	.660/29.03	.660/28.86	.668/28.93	.681/28.71
8.0	.662/28.37	.657/28.92	.655/29.06	.644/28.88	.644/28.75	.647/28.51
10.0	.662/28.37	.657/28.97	.652/29.06	.627/28.73	.613/28.53	.613/28.06

Table S15: **FLUX.1-schnell input-latent noise ablation over α and f_α.** Harrington-style blend $\mathbf{z}' = (1-\alpha)\,\mathbf{z} + \alpha\,\mathbf{z}_{\text{pink}}$ with frequency filter $1/(1+f)^{f_\alpha}$ generated via 2D FFT, applied to the initial latent. DrawBench 199 prompts $\times$ 4 images, 1 denoising step. Each cell shows InBSim↓ / CLIP↑. Baseline: InBSim=0.668, CLIP=28.41. Input-level noise produces meaningful diversity only in a narrow operating window: at $f_\alpha \geq 3$ with $\alpha \geq 0.5$, sharp noise catastrophically destroys CLIP while diversity reverses, indicating a degenerate-mode failure where the model produces uniformly degraded outputs.

	Perturbation strength α					
f_α	0.0	0.1	0.2	0.5	0.7	1.0
Low-frequency regime ($f_\alpha \leq 1$, designed Harrington range)						
0.0 (W)	.668/28.41	.665/28.32	.666/28.35	.669/28.42	.666/28.48	.681/28.43
0.1	.668/28.41	.665/28.33	.666/28.33	.666/28.42	.658/28.46	.669/28.40
0.2	.668/28.41	.665/28.31	.664/28.34	.660/28.36	.649/28.35	.659/28.33
0.5	.668/28.41	.665/28.30	.662/28.33	.640/28.15	.621/28.06	.613/27.86
1.0	.668/28.41	.664/28.32	.660/28.31	.593/27.77	.527/26.72	.505/26.08
High-frequency regime ($f_\alpha \geq 2$)						
2.0	.668/28.41	.662/28.28	.648/28.23	.469/24.82	.492/21.68	.521/20.92
3.0	.668/28.41	.659/28.27	.636/28.07	.500/21.53	.566/18.92	.596/18.34
4.0	.668/28.41	.655/28.34	.601/27.85	.560/19.29	.645/17.16	.672/16.53
5.0	.668/28.41	.646/28.29	**.566/27.43**	.615/17.82	.707/15.93	.735/15.42
6.0	.668/28.41	.642/28.22	.519/26.63	.663/16.73	.753/15.16	.780/14.86
8.0	.668/28.41	.619/28.10	.466/24.69	.731/15.43	.797/14.75	.803/14.59
10.0	.668/28.41	.597/27.82	.456/22.90	.763/15.02	.804/14.68	.802/14.59

Table S16: **FLUX.1-schnell CADS ablation over γ and s.** At T=1 the per-step annealing schedule collapses; we therefore treat γ and s as direct hyperparameters. Configuration: ψ=0.9 rescaling, T5-sequence-only noise injection, DrawBench 199 prompts × 8 images. Each cell shows InBSim↓ / CLIP↑. Baseline (γ=1.0, no noise): InBSim=0.665, CLIP=28.35. CADS reduces InBSim only at substantial CLIP cost.

Configuration	InBSim↓	CLIP↑
γ=0.9, s=0.5*	**0.563**	**27.21**
γ=0.9, s=1.0	0.328	22.05
γ=0.9, s=2.0	0.152	15.22
γ=0.7, s=0.5	0.338	22.30
γ=0.7, s=1.0	0.157	15.50
γ=0.7, s=2.0	0.103	12.41
γ=0.5, s=0.5	0.204	18.00
γ=0.5, s=1.0	0.122	13.61
γ=0.5, s=2.0	0.093	11.84
γ=0.3, s=0.5	0.157	15.15
γ=0.3, s=1.0	0.105	12.73
γ=0.3, s=2.0	0.090	11.61

*Operating point reported in the main results.

Table S17: **FLUX.1-schnell STRIDE ablation over K (PCA rank) and α (perturbation strength).** f_α=5, singular-value scaling ($S/\bar{S}$), P=2, layers L0–7, 1 denoising step. DrawBench 199 prompts × 4 images. Baseline: InBSim=0.664, CLIP=28.40, V_D/p_{DINO}=2.228, HPS=0.294.

(a) InBSim↓ / CLIP↑

	Perturbation strength α					
K	0.1	0.2	0.4	**0.5***	0.6	0.8
4	.664 / 28.40	.663 / 28.39	.664 / 28.37	.663 / 28.41	.661 / 28.36	.661 / 28.37
8	.663 / 28.38	.664 / 28.36	.664 / 28.40	.663 / 28.42	.664 / 28.42	.663 / 28.45
16	.663 / 28.38	.664 / 28.38	.665 / 28.45	.665 / 28.50	.663 / 28.49	.660 / 28.45
32	.663 / 28.36	.663 / 28.38	.662 / 28.44	.662 / 28.40	.661 / 28.45	.659 / 28.43
64	.663 / 28.40	.665 / 28.44	.664 / 28.45	.662 / 28.48	.660 / 28.45	.655 / 28.48
128	.663 / 28.38	.664 / 28.42	.663 / 28.48	.660 / 28.50	.657 / 28.51	.650 / 28.62
256	.663 / 28.42	.664 / 28.47	.661 / 28.61	.655 / 28.69	.651 / 28.81	.646 / 28.96
512	.665 / 28.48	.661 / 28.59	.647 / 28.99	.636 / 29.08	.631 / 29.21	.608 / 28.98
1024	.661 / 28.63	.655 / 28.96	.614 / 28.83	.590 / 28.61	.557 / 28.14	.496 / 26.66

(b) V_D/p_{DINO}↑ / HPS↑

	Perturbation strength α					
K	0.1	0.2	0.4	**0.5***	0.6	0.8
4	2.231 / .293	2.232 / .294	2.231 / .294	2.232 / .294	2.239 / .294	2.239 / .294
8	2.235 / .293	2.232 / .294	2.229 / .293	2.233 / .293	2.232 / .294	2.236 / .293
16	2.233 / .294	2.233 / .293	2.228 / .293	2.226 / .294	2.233 / .294	2.240 / .293
32	2.233 / .294	2.232 / .294	2.237 / .293	2.235 / .293	2.238 / .294	2.245 / .293
64	2.232 / .294	2.228 / .294	2.229 / .294	2.236 / .293	2.240 / .293	2.255 / .292
128	2.230 / .294	2.228 / .294	2.233 / .295	2.242 / .293	2.251 / .293	2.272 / .291
256	2.232 / .294	2.226 / .293	2.239 / .293	2.260 / .292	2.270 / .291	2.284 / .290
512	2.224 / .293	2.237 / .293	2.283 / .290	2.322 / .290	2.343 / .289	2.429 / .287
1024	2.245 / .293	2.269 / .290	2.417 / .287	2.496 / .284	2.606 / .282	2.820 / .270

A.3.2 SD3.5 LARGE TURBO (MMDIT, 4-STEP)

Layer and Timestep ablation We report the layer-range probe for SD3.5 Large Turbo in Table S18. To identify a suitable layer range cheaply, we run this probe at a milder f_α=5 (main experiments use f_α=20 for stronger diversity). Early-block ranges (L0–6 through L0–37) form a competitive cluster, all within $\sim$1% InBSim and $\sim$2% V_D/p of one another, while late-only injection (L20–37) collapses to baseline (InBSim 0.703, V_D/p 2.079 vs. baseline 0.707 / 2.067), indicating that early transformer blocks carry the structural decisions STRIDE is designed to perturb. We adopt **L0–11** as a representative early-block range for all main experiments.

We further ablate *which* denoising step receives the STRIDE perturbation in Table S19. Among single-step configurations, the first step (S0) is the most effective injection point: diversity decreases monotonically as the perturbed step moves later, with InBSim rising from 0.639 at S0 to 0.708 at S3 (essentially baseline) and V_D/p falling correspondingly from 2.300 to 2.063. Perturbing all four steps (S0–3) yields marginally higher diversity than S0 alone (InBSim 0.621, V_D/p 2.386) but degrades both CLIP (-0.25) and HPS (-0.015), placing it off the diversity vs quality Pareto frontier. This pattern is consistent with the view that compositional commitments are made at the first denoising step in distilled few-step models Kynkäänniemi et al. (2024); Gandikota & Bau (2026): once the global structure has been committed at S0, further perturbation mainly disrupts local refinement without contributing meaningful additional diversity. We therefore restrict STRIDE injection to S0 for SD3.5 Large Turbo.

PCA rank ablation Table S24 evaluates PCA rank K on SD3.5 Large Turbo at f_α=20. Low ranks ($K < 256$) at the fixed α=0.1 yield InBSim within 0.02 of baseline and CLIP changes within measurement noise, suggesting insufficient subspace coverage for effective perturbation. Diversity gains emerge at K=512 and peak at K=1024, where the full-rank projection reduces InBSim to 0.639 (9.6% below the DrawBench baseline of 0.707), with a 0.11 decrease in CLIP. At α=0.2, diversity improves monotonically through K=512 (InBSim 0.633, CLIP 27.89), whereas K=1024 sharply degrades CLIP to 21.96. These results support selecting K=1024 with α=0.1, while K=512 with α=0.2 provides a lower-cost alternative.

STRIDE ablation We sweep α and f_α for STRIDE on SD3.5 Large Turbo at the locked L0–11 / K=1024 / S0-only configuration in Table S20. The behaviour differs qualitatively from FLUX. *(i) Diversity is monotonic in f_α at fixed α=0.1:* InBSim decreases steadily from 0.706 (f_α=0.2) to 0.572 (f_α=100), with CLIP remaining within 0.86 of baseline across the entire range. The model is far more tolerant of strongly low-frequency-biased noise than FLUX is, suggesting that high-frequency content in the perturbation is what destabilises the distilled trajectory. *(ii) α=0.2 collapses CLIP at moderate f_α:* the same f_α values that work cleanly at α=0.1 collapse CLIP to $\sim$20 at α=0.2 ($f_\alpha \in [5, 15]$); recovery only begins above f_α=20 and remains 5 CLIP points below baseline even at f_α=100. *(iii) α=0.3 collapses entirely:* CLIP drops below 21 at every tested f_α and InBSim *rises* above baseline at $f_\alpha \geq 3$ (e.g., f_α=5: InBSim =0.925, CLIP =15.04), the model produces uniformly degraded outputs. We adopt α=0.1, f_α=20 as the SD3.5 Turbo default.

STRIDE No-PCA ablation We perform the same (α, f_α) sweep with STRIDE No-PCA (uniform feature-level noise, no PCA projection) on SD3.5 Large Turbo at the locked L0–11 / S0-only configuration in Table S21. The contrast with STRIDE is again striking. *(i) STRIDE No-PCA fails to produce diversity at safe α:* at α=0.1, InBSim *rises* above baseline at every f_α, peaking at 0.745 at f_α=5. The model interprets the unstructured perturbation as corruption and compensates by producing *more similar* outputs. *(ii) Higher α trades quality for diversity but cannot escape the trade-off:* α=0.2 yields InBSim =0.532 at the cost of CLIP collapse to 26.23 (f_α=50); α=0.3 reaches InBSim =0.426 at CLIP =24.94. Comparison at α=0.1, f_α=20, STRIDE yields InBSim =0.639 at CLIP =28.07, while STRIDE No-PCA yields InBSim =0.724 at CLIP =27.58. STRIDE produces a -9.6% diversity gain over baseline ; STRIDE No-PCA yields essentially zero gain. Removing PCA projection eliminates the method's ability to convert feature-level noise into actual diversity, exactly as observed on FLUX.

Noise Baseline Ablation We sweep α and f_α for the input-latent pink-noise blend on SD3.5 Large Turbo in Table S22. Three patterns emerge. *(i) Diversity emerges only at high f_α:* for both α=0.1 and α=0.2, InBSim is essentially flat ($\approx$0.72–0.73) for $f_\alpha \leq 5$ and only begins to decrease at

$f_\alpha \geq 8$. *(ii) At α=0.2, the 4-step schedule amplifies latent perturbation and collapses CLIP:* moving from f_α=10 to f_α=50 at α=0.2 trades modest diversity gains (InBSim $0.570 \rightarrow 0.528$) for catastrophic CLIP loss ($24.76 \rightarrow 15.92$); InBSim itself stops decreasing and turns back upward ($0.502 \rightarrow 0.528$) at f_α=50 as the model enters a degenerate failure mode. *(iii) α=0.1 remains usable across a wider range:* the operating point α=0.1, f_α=10 achieves InBSim =0.681 at CLIP =27.62, and we adopt this configuration as the input-noise default on SD3.5 Turbo.

CADS Baseline We sweep CADS hyperparameters on SD3.5 Large Turbo in Table S23, holding τ_2=0.9 and ψ=1.0 and varying the schedule start τ_1 and noise magnitude s. Three patterns emerge. *(i) CADS collapses CLIP at every tested configuration:* the highest CLIP value across the entire grid is 22.67 (at τ_1=0.8, s=0.15), still 5.5 points below baseline (28.18), and CLIP drops below 14 at aggressive settings (τ_1=0.4, $s \geq 0.8$). *(ii) InBSim drops far below baseline only because outputs are visually degraded:* the lowest InBSim values (e.g., 0.149 at τ_1=0.4, s=1.0) co-occur with CLIP =13.03, indicating that the model produces uniformly incoherent outputs that are trivially dissimilar from each other rather than coherently diverse samples covering the prompt's mode space. *(iii) τ_1=0.8 saturates:* when the perturbation window $[\tau_1, \tau_2]$=[0.8, 0.9] covers only the last 0.1 of the timestep range, the schedule barely activates within the 4-step turbo trajectory and InBSim/CLIP become nearly invariant in s (all s values yield InBSim $\approx$ 0.317, CLIP $\approx$ 22.64). We adopt τ_1=0.8, s=0.15 as the least-degraded operating point and report it in the main results with a † marker indicating CADS collapses on this model. We include CADS as a baseline despite the collapse to document a methodologically important fact: the dominant pre-existing inference-time diversity method does not transfer to few-step distilled MM-DiT generation.

SPELL (Sparse Repellency) Kirchhof et al. (2024) Baseline We reimplement Sparse Repellency following Algorithm 1 and the parallelised pseudocode in Appendix D of the original paper, as the code is not publicly available. SPELL is applied to SD3.5 Large Turbo only; we exclude it from the FLUX.1-schnell evaluation as the single-step schedule reduces SPELL to a one-shot post-prediction adjustment with no opportunity for the iterative repellency dynamics the method is designed for.

We use SPELL in its dynamic batch-coupled mode (Algorithm 1, lines 5–9 only), with no static reference set. At each denoising step t and for each particle i, the shielded set $\mathcal{S}_t^{(i)} = \{\hat{\mathbf{x}}_0^{(j)} : j \neq i\}$ is rebuilt from the predicted clean endpoints of the other particles in the batch ($B = 4$). If any $\hat{\mathbf{x}}_0^{(j)} \in \mathcal{S}_t^{(i)}$ falls within Euclidean distance r of $\hat{\mathbf{x}}_0^{(i)}$, the latter is projected radially outward to the boundary of the r-ball, scaled by an overcompensation factor λ, and the corresponding correction is propagated back to $\mathbf{x}_t^{(i)}$ through the scheduler's denoising update. Outside this threshold the repellency is exactly zero (the "sparse" regime), tapering off naturally as the predicted endpoints sharpen towards the end of the trajectory. The predicted clean endpoint is recovered from the velocity prediction $v_\theta(\mathbf{x}_t, t)$ via the rectified-flow identity $\hat{\mathbf{x}}_0 = \mathbf{x}_t - \sigma_t \cdot v_\theta$, where σ_t is the FlowMatchEulerDiscreteScheduler's time variable along the linear-interpolation path $\mathbf{x}_t = (1 - \sigma_t)\,\mathbf{x}_0 + \sigma_t\,\boldsymbol{\epsilon}$. Distances are computed in fp32 on the standard (B, C, H, W) SD3.5 Turbo latent. The projection is applied at every denoising step (4 of 4 for SD3.5 Turbo).

Hyperparameter calibration. SPELL exposes two hyperparameters: the repellency radius r (size of the forbidden ball around each shielded endpoint) and the overcompensation factor λ (multiplier on the correction vector that controls whether the trajectory escapes the ball gradually across steps or aggressively early). The original paper uses $\lambda = 1.6$ for SD3-Medium and other latent-space text-to-image models on CC12M (Kirchhof et al., 2024, §3, App. D).

For the radius r, the (Kirchhof et al., 2024, App. D) notes that the optimal value depends on the latent-space dimensionality and resolution, and prescribes a measurement-based calibration protocol: generate one batch per prompt without repellency, measure the median pairwise $\|\hat{\mathbf{x}}_0^{(i)} - \hat{\mathbf{x}}_0^{(j)}\|_2$ at the final denoising step, and sweep $r \in [0,\ 2 \times \text{median}]$. We follow this protocol exactly for our setting (Turbo distillation, 4-step schedule, CFG=0, 1024×1024 resolution) and sweep both jointly as both r and λ control trajectory dynamics and may interact non-trivially in the compressed 4-step regime.

Calibration probe (Table S25). We run the calibration probe on a sampled subset of 20 DrawBench prompts. For each prompt, we generate $K = 4$ images at 1024×1024 with CFG=0 and the

standard 4-step Turbo schedule, with SPELL disabled ($r = 0$, which sets the ReLU clamp to zero everywhere). At each denoising step, we compute the predicted clean endpoint $\hat{\mathbf{x}}_0$ via the rectified-flow identity above, and record the median pairwise Euclidean distance across the six pairs in the $B = 4$ batch. Table S25 reports per-prompt medians at all four denoising steps. As expected, distances grow monotonically with the trajectory: predicted endpoints are diffuse at σ=1.0 and sharpen towards $\sigma \approx 0$, where they correspond to the final outputs. The final-step grand median across all 20 prompts is $M = 391.9$, with per-prompt final medians ranging from 303 to 635.

Following the paper's $r \in [0, 2M]$ prescription, we sweep $r \in \{0.5M, 1.0M, 1.5M, 2.0M\} = \{196, 392, 588, 784\}$ crossed with the three λ values explicitly studied in the original paper, $\lambda \in \{1.0, 1.6, 2.0\}$. This 4×3 grid of 12 configurations is evaluated on full DrawBench (199 prompts $\times$ 4 images), with SPELL applied at all four denoising steps. Table S26 reports diversity (InBSim↓, V_D/p_{DINO} ↑) and quality (CLIP↑, HPS↑) for every grid cell. We select the configuration with the best diversity vs quality trade-off on DrawBench and apply ($r^\star = 588$, $\lambda^\star = 1.0$) uniformly across all four benchmarks (DrawBench, PartiPrompts, GenEval, MS-COCO). The selected values are reported in Table S3

Table S18: SD3.5 Large Turbo STRIDE layer-range probe on DrawBench (199 prompts $\times$ 4 images). K=1024, P=2, α=0.1, f_α=5, step 0 only, 4 denoising steps, guidance scale 0.0. We use f_α=20 in the main experiments for stronger diversity; here a milder f_α is used purely to identify a suitable layer range. We adopt L0–11 as a representative early-block range for all production runs.

Layers	**InBSim↓**	**CLIP↑**	**V_D/p_{DINO}↑**	**HPS↑**
Baseline (α=0)	0.707	28.18	2.067	0.291
L0 (1 block)	0.701	28.21	2.088	0.289
L0–2 (3 blocks)	0.700	28.23	2.086	0.289
L0–6 (7 blocks)	0.688	28.06	2.136	0.290
L0–11 (12 blocks)	0.694	28.16	2.116	0.290
L0–19 (20 blocks)	0.686	28.10	2.143	0.288
L0–37 (all 38)	0.692	28.13	2.126	0.289
L20–37 (late 18)	0.703	28.12	2.079	0.291

Table S19: **SD3.5 Large Turbo step-index ablation.** Effect of which denoising step receives STRIDE perturbation, evaluated on DrawBench (199 prompts $\times$ 4 images). K=1024, P=2, L0–11, α=0.1, f_α=20, 4 denoising steps total, guidance scale 0.0. Single-step diversity decreases monotonically as the perturbed step moves later. Perturbing all steps (S0–3) yields marginally higher diversity than S0 alone but degrades CLIP and HPS; we adopt S0 as the Pareto-optimal operating point for production runs.

Step index	**InBSim↓**	**CLIP↑**	**V_D/p_{DINO}↑**	**HPS↑**
S0 (step 0 only)	0.639	28.07	2.300	0.288
S1 (step 1 only)	0.687	**28.35**	2.148	0.288
S2 (step 2 only)	0.697	28.25	2.111	0.280
S3 (step 3 only)	0.708	28.24	2.063	**0.291**
S0–3 (all steps)	**0.621**	27.82	**2.386**	0.273

Table S20: **SD3.5 Large Turbo STRIDE ablation over α and f_α.** Spatial PCA, K=1024, P=2, layers L0–11, step 0 only, 4 denoising steps, guidance scale 0.0. DrawBench: α=0.1 and α=0.2 runs use 199 prompts × 4 images; α=0.3 uses an early 30-prompt × 4 image screening set, abandoned once collapse was observed. Each cell shows InBSim↓ / CLIP↑. Baseline (199 prompts): InBSim≈0.707, CLIP≈28.18.

	Perturbation strength α		
f_α	0.1	0.2	0.3‡
0.2	.706/28.16	.673/27.97	.631/20.19
0.5	.704/28.17	.663/27.80	.693/19.68
3.0	.694/28.10	.603/24.52	.890/15.83
5.0	.694/28.16	.634/21.37	.925/15.04
7.0	.686/28.15	.661/19.52	.910/15.32
8.0	.683/28.19	.670/19.24	—
9.0	.680/28.19	.664/19.24	—
10.0	.679/28.21	.649/19.36	—
11.0	.676/28.14	.633/19.41	—
12.0	.668/28.16	.610/19.68	—
15.0	.658/28.09	.542/20.24	—
20.0⋆	**.639/28.07**	.482/21.96	—
30.0	.612/27.80	.463/23.32	—
35.0	.600/27.77	.456/23.56	—
40.0	.594/27.75	.451/23.68	—
45.0	.592/27.69	.451/23.93	—
50.0	.592/27.61	.443/23.98	—
80.0	.579/27.40	.419/24.61	—
100.0	.572/27.34	.413/24.30	—

Table S21: **SD3.5 Large Turbo STRIDE No-PCA ablation over α and f_α.** Uniform feature-level noise without PCA projection, layers L0–11, step 0 only, 4 denoising steps, guidance scale 0.0. DrawBench 100 prompts × 4 images. Each cell shows InBSim↓ / CLIP↑. Baseline (199 prompts): InBSim≈0.707, CLIP≈28.18. In contrast to STRIDE (Table S20), feature-level perturbation *without* PCA either fails to reduce InBSim (at safe α=0.1, InBSim *rises* above baseline at every f_α) or reduces it only at substantial CLIP cost ($\alpha \geq 0.2$ collapses CLIP below 26).

	Perturbation strength α		
f_α	0.1	0.2	0.3
0.2	.700/26.90	.655/22.31	.589/22.41
0.5	.704/27.03	.656/22.33	.585/22.41
1.0	.708/27.18	.654/22.40	.585/22.51
2.0	.714/27.41	.655/22.45	.581/22.64
5.0	.745/27.51	.645/23.37	.547/23.24
8.0	.737/27.51	.638/24.85	.548/24.10
10.0	.736/27.54	.651/25.92	.533/24.55
20.0	.724/27.58	.631/27.02	.489/25.48
30.0	.714/27.57	.607/26.65	.468/25.20
50.0	.680/27.57	.532/26.23	.426/24.94

Table S22: **SD3.5 Large Turbo input-latent noise ablation over α and f_α.** Harrington-style blend $\mathbf{z}' = (1-\alpha)\,\mathbf{z} + \alpha\,\mathbf{z}_{\text{pink}}$ with frequency filter $1/(1+f)^{f_\alpha}$ generated via 2D FFT, applied to the initial latent of SD3.5 Large Turbo (4 denoising steps, guidance scale 0.0). DrawBench 100 prompts $\times$ 4 images. Each cell shows InBSim↓ / CLIP↑. Baseline (199 prompts): InBSim≈0.707, CLIP≈28.18.

	Perturbation strength α	
f_α	0.1	0.2
0.2	.728/27.72	.726/27.70
0.5	.726/27.72	.725/27.72
1.0	.728/27.75	.725/27.72
2.0	.730/27.74	.720/27.73
3.0	.727/27.76	.715/27.70
4.0	.724/27.69	.697/27.69
5.0	.717/27.77	.679/27.51
8.0	.704/27.64	.614/26.77
10.0	.681/27.62	.570/24.76
20.0	.599/27.14	.516/20.00
30.0	.541/25.92	.502/18.02
50.0	.486/24.08	.528/15.92

Table S23: **SD3.5 Large Turbo CADS ablation over τ_1 and s.** Linear annealing schedule $\gamma(t) = (\tau_2 - t)/(\tau_2 - \tau_1)$ with τ_2=0.9, ψ=1.0, noise injected into the concatenated CLIP+T5 sequence embeddings at every step. 4 denoising steps, guidance scale 0.0. DrawBench 199 prompts $\times$ 4 images. Each cell shows InBSim↓ / CLIP↑. Baseline: InBSim≈0.707, CLIP≈28.18. CADS reduces InBSim aggressively, but at every tested configuration, CLIP collapses below 23.

	Schedule start τ_1 (with τ_2=0.9)		
s	0.4	0.6	0.8
0.15*	.285/21.47	.304/22.04	**.318/22.67**
0.25	.250/19.85	.280/21.34	.317/22.64
0.4	.203/17.48	.246/19.85	.317/22.66
0.5	.178/15.87	.227/18.87	.317/22.64
0.8	.150/13.48	.180/16.08	.317/22.66
1.0	.149/13.03	.163/14.92	.317/22.64

*Operating point (τ_1=0.8, τ_2=0.9, s=0.15, ψ=1.0) reported in the main results as the least-degraded CADS configuration.

Table S24: **SD3.5 Large Turbo STRIDE PCA-rank ablation over K and α.** f_α=20, P=2, singular-value scaling ($S/\bar{S}$), L0–11, step 0 only, 4 denoising steps, guidance scale 0.0. DrawBench 199 prompts × 4 images. Each cell shows InBSim↓ / CLIP↑ / V_D/p_{DINO}↑ / HPS↑. Baseline (DrawBench, 199 prompts): 0.707 / 28.18 / 2.067 / 0.291. At α=0.1, K=1024 yields the largest diversity improvement at −0.11 CLIP; at α=0.2, the full-rank K=1024 configuration collapses (CLIP falls below 22).

	Perturbation strength α	
K	0.1	0.2
4	0.698 / 28.17 / 2.097 / 0.291	0.694 / 28.17 / 2.111 / 0.292
8	0.700 / 28.15 / 2.090 / 0.291	0.696 / 28.12 / 2.104 / 0.291
16	0.695 / 28.12 / 2.108 / 0.291	0.696 / 28.14 / 2.108 / 0.291
32	0.696 / 28.15 / 2.103 / 0.291	0.691 / 28.17 / 2.119 / 0.291
64	0.695 / 28.22 / 2.108 / 0.291	0.689 / 28.13 / 2.129 / 0.291
128	0.694 / 28.18 / 2.111 / 0.292	0.685 / 28.08 / 2.139 / 0.291
256	0.690 / 28.17 / 2.120 / 0.291	0.676 / 28.05 / 2.165 / 0.292
512	0.676 / 28.02 / 2.165 / 0.291	<u>0.633 / 27.89 / 2.326 / 0.288</u>
1024⋆	**0.639 / 28.07 / 2.300 / 0.288**	0.482 / 21.96 / 2.867 / 0.257

⋆Selected Value: α=0.1, K=1024. <u>Underlined</u>: matches selected diversity at lower cost (α=0.2, K=512; +76% vs. +167% overhead; Table S11).

Table S25: **SPELL radius calibration probe on SD3.5 Large Turbo.** Per-prompt median pairwise $\|\hat{\mathbf{x}}_0^{(i)} - \hat{\mathbf{x}}_0^{(j)}\|_2$ measured at each of the 4 denoising steps (B=4, 1024×1024, CFG=0, no repulsion). Distances grow from step 0 (pure noise) to step 3 (clean image) as the predicted clean endpoints sharpen and diverge. Following the paper's protocol (Kirchhof et al., 2024, App. D), we use the grand median of the final-step column as the centre of the radius sweep, with $r \in [0,\ 2 \times \text{grand median}]$.

Prompt	**Step 0** σ=1.000	**Step 1** σ=0.858	**Step 2** σ=0.602	**Step 3 (final)** σ=0.009
0	346.6	421.8	439.6	440.3
1	351.9	438.0	456.6	457.4
2	268.2	323.4	341.3	342.2
3	349.6	404.7	424.8	425.6
4	280.0	336.1	348.7	349.7
5	320.1	386.1	411.6	412.7
6	323.1	369.2	381.9	382.4
7	330.6	367.9	380.0	380.6
8	375.5	430.9	444.4	445.1
9	216.0	296.5	312.6	313.5
10	234.2	284.4	302.4	303.4
11	267.0	344.6	362.1	362.9
12	307.5	348.7	370.5	371.3
13	356.6	411.6	435.4	436.2
14	260.1	316.4	330.6	331.4
15	287.3	380.5	410.7	411.8
16	230.6	294.9	313.1	314.4
17	534.9	615.9	634.8	635.4
18	312.0	397.5	410.0	410.8
19	276.8	356.2	382.9	383.9
Grand median	**310.1**	**373.6**	**391.0**	**391.9**
Grand mean	311.4	376.3	396.7	397.5

Final-step grand median $M = 391.9$. Selected sweep:
$r \in \{0.5M,\ 1.0M,\ 1.5M,\ 2.0M\} = \{196,\ 392,\ 588,\ 784\}$.

Table S26: **SD3.5 Large Turbo SPELL ablation over r (repellency radius) and λ (over-compensation factor).** Sparse Repellency applied at all 4 denoising steps in dynamic batch-coupled mode (no static reference set), B=4, guidance scale 0.0, 1024×1024 resolution. DrawBench 199 prompts $\times$ 4 images. Each cell shows InBSim↓ / CLIP↑ / V_D/p_{DINO}↑ / HPS↑. Baseline: $0.707/28.18/2.067/0.291$. The radius grid $r \in \{196, 392, 588, 784\} = \{0.5M, 1.0M, 1.5M, 2.0M\}$ follows the calibration probe (final-step grand median M=391.9, Table S25); λ values are those explicitly studied in the original paper Kirchhof et al. (2024). We adopt the bolded configuration (r=588, λ=1.0) as the reported SPELL operating point in the main results.

r	λ	InBSim↓	CLIP↑	V_D/p_{DINO}↑	HPS↑
r=196 ($0.5M$): minimal effect, near-baseline diversity					
196	1.0	0.704	28.15	2.079	0.291
196	1.6	0.700	28.12	2.087	0.290
196	2.0	0.713	28.19	2.047	0.291
r=392 ($1.0M$): moderate diversity, modest CLIP cost					
392	1.0	0.676	27.63	2.182	0.287
392	1.6	0.660	27.13	2.252	0.281
392	2.0	0.657	26.99	2.270	0.279
r=588 ($1.5M$): strongest usable diversity					
588	**1.0***	**0.632**	**26.47**	**2.352**	**0.272**
588	1.6	0.600	25.47	2.473	0.262
588	2.0	0.578	24.91	2.550	0.257
r=784 ($2.0M$): degenerate regime, CLIP collapses					
784	1.0	0.589	25.29	2.512	0.261
784	1.6	0.560	24.33	2.631	0.253
784	2.0	0.552	24.07	2.660	0.252

*Selected SPELL operating point for the main results: strongest diversity gain (-10.6% InBSim) retaining CLIP > 26.

A.4 WHY UNSTRUCTURED FEATURE NOISE RAISES CLIP: DISPLACEMENT AND ATTENTION ANALYSIS

This section expands the observation in Sec. 5 that STRIDE No-PCA reports higher CLIP than STRIDE in Tables 1 and 4 while producing *less* diversity than the unperturbed baseline.

Observation. On FLUX.1-schnell both variants raise CLIP above the unperturbed baseline (DrawBench: 28.47 baseline; 28.61 STRIDE; 29.04 No-PCA), while on SD3.5 Turbo both remain at baseline level. We interpret No-PCA's higher CLIP as a symptom of reduced diversity rather than a genuine alignment benefit, because it accompanies InBSim above baseline (0.673 vs. 0.662 on DrawBench) and because, at matched diversity on the frontier, No-PCA is worse than STRIDE on both CLIP and HPS (Table S29).

Hypothesis. Perturbations injected into the text-conditioned dual-stream blocks displace features from the model's learned activation distribution, and the corrective dynamics of subsequent layers re-ground the perturbed features against the text condition. The projection direction determines how much correction occurs and what the injected energy becomes: PCA-directed noise lies largely within the dominant activation subspace and passes through largely uncorrected, whereas unstructured noise is largely off-subspace and is substantially corrected away, with part of the absorbed energy expressed as stronger re-grounding on the prompt (higher CLIP) and a tighter output distribution.

Measurement 1: feature displacement. At identical α (0.5, FLUX.1-schnell, hooked blocks L0–7), the unprojected perturbation sustains over twice the relative feature displacement of the PCA-projected one (Table S27), yet yields less output variation than the unperturbed baseline: more energy enters and more correction occurs, but less variety emerges. This also addresses a natural confound: projecting onto K of the P^2D directions reduces the injected norm, so at the same α No-PCA injects *more* energy, not less, and its failure to diversify is not explained by a lower value of α.

Measurement 2: text-attention pathway. For each prompt, we compute per-layer attention-saliency maps (image-token attention onto text tokens) for the baseline, STRIDE, and No-PCA under identical seeds and identical α, and record the deviation from baseline (per-layer $|\Delta|_{\max}$, averaged over 13 layers and 4 generations per prompt), aggregated over all 199 DrawBench prompts. A seed-only control (baseline vs. baseline at a different seed) provides the natural-variation yardstick (Table S28). PCA-directed perturbation shifts this pathway by only $0.54\times$ the seed-to-seed variation, stably across prompts (± 0.06), while unstructured noise produces 1.5–$1.7\times$ larger deviations, comparable to a full seed change. The α-sweeps corroborate the asymmetry: at $K{=}1024$ the PCA variant reduces InBSim monotonically from $\alpha{=}0.1$ onward (Table S17), whereas No-PCA raises InBSim at every $\alpha \geq 0.2$ at $f_\alpha{=}5$ (Table S14) and only reduces it at settings where alignment and quality degrade.

Scope of the claim. The feature-displacement and attention-saliency measurements are direct observations; the proposed causal relationship between them is a hypothesis rather than an established property of the model.

Table S27: **Relative feature displacement at identical** α (FLUX.1-schnell, $\alpha{=}0.5$, mean over the hooked blocks L0–7). The unprojected perturbation displaces features $2.24\times$ more than the PCA-projected one, yet produces less output variation than the baseline.

Relative feature displacement	STRIDE	No-PCA	No-PCA / STRIDE
$\|\Delta\mathbf{h}\|_2/\|\mathbf{h}_{\text{base}}\|_2$	0.243 ± 0.004	0.546 ± 0.005	$2.24\times$

Table S28: **Attention-saliency shift relative to baseline** (FLUX.1-schnell, DrawBench, 199 prompts; per-layer $|\Delta|_{\max}$ of image-to-text attention saliency, averaged over 13 layers and 4 generations per prompt; identical seeds and α). The seed-only control is baseline vs. baseline at a different seed.

Deviation from baseline	**STRIDE**	**No-PCA**	**Seed control**
Attention-saliency shift (mean $\pm$ std)	0.037 ± 0.004	0.056 ± 0.004	0.068 ± 0.004
Ratio to seed control	0.54 ± 0.06	0.82 ± 0.07	1.00

A.5 DIVERSITY–QUALITY TRADE-OFF ANALYSIS

Matched-diversity comparison. Table S29 compares CLIP and HPS at matched InBSim on the DrawBench sweeps. On the HPS axis in isolation, pink latent noise and CADS (adapted and tuned for the single-step setting) sit 0.001–0.005 HPS above STRIDE at matched diversity ($\leq 1.8\%$), but concede 0.6–1.4 CLIP. No baseline Pareto-dominates STRIDE in the joint (diversity, alignment, quality) space on either backbone.

Table S29: **CLIP / HPS at matched diversity on DrawBench.** STRIDE values are absolute; the baseline columns give the difference from STRIDE (CLIP / HPS) at the same InBSim, read off each method's sweep. n/r: no sweep point reaches this diversity level; †CADS collapses on SD3.5 Turbo (Table S23).

Backbone	**Matched InBSim**	**STRIDE (CLIP / HPS)**	**Pink latent**	**CADS**	**No-PCA**
FLUX.1-schnell	0.62	29.20 / 0.288	-1.10 / $+0.003$	-1.38 / $+0.003$	-0.57 / -0.012
FLUX.1-schnell	0.58	28.71 / 0.285	-1.10 / $+0.002$	-1.34 / $+0.005$	n/r
FLUX.1-schnell	0.64	29.50 / 0.283	-1.33 / $+0.008$	-1.44 / $+0.009$	-0.65 / -0.004
FLUX.1-schnell	0.60	28.94 / 0.284	-1.10 / $+0.005$	-1.34 / $+0.007$	n/r
FLUX.1-schnell	0.56	28.48 / 0.281	-1.15 / $+0.003$	-1.34 / $+0.009$	n/r
FLUX.1-schnell	0.52	27.78 / 0.276	-1.13 / $+0.001$	n/r	n/r
FLUX.1-schnell	0.48	27.12 / 0.269	-1.90 / -0.004	n/r	n/r
SD3.5 Turbo	0.65	28.08 / 0.289	-0.70 / $+0.002$	n/r†	-0.84 / -0.002
SD3.5 Turbo	0.60	27.77 / 0.287	-0.62 / $+0.003$	n/r†	-1.16 / -0.003
SD3.5 Turbo	0.68	28.24 / 0.289	-0.67 / $+0.003$	n/r†	-0.67 / 0.000
SD3.5 Turbo	0.63	27.87 / 0.285	-0.58 / $+0.006$	n/r†	-0.86 / 0.000
SD3.5 Turbo	0.58	27.42 / 0.280	-0.68 / $+0.007$	n/r†	-0.92 / $+0.002$
SD3.5 Turbo	0.55	26.94 / 0.276	-0.84 / $+0.008$	n/r†	-0.62 / $+0.005$
SD3.5 Turbo	0.50	26.06 / 0.269	-1.51 / $+0.003$	n/r†	-0.38 / $+0.008$
SD3.5 Turbo	0.45	25.17 / 0.263	n/r	n/r†	-0.08 / $+0.009$

Table S30: **CLIP / HPS at matched diversity on DrawBench: multi-pass baselines.** As in Table S29, STRIDE values are absolute and the method columns give the difference from STRIDE at the same InBSim. DDC is a distilled model with no strength parameter, so it exists at a single diversity level and can only be matched there. Both methods and STRIDE are evaluated on the same 199-prompt cohort. n/r: no sweep point reaches this diversity level.

Backbone	**Matched InBSim**	**STRIDE (CLIP / HPS)**	**DDC**	**SPELL**
SD3.5 Turbo	0.68	28.24 / 0.289	n/r	-0.53 / -0.001
SD3.5 Turbo	0.65	28.08 / 0.288	n/r	-1.28 / -0.011
SD3.5 Turbo	0.63	27.87 / 0.285	n/r	-1.61 / -0.013
SD3.5 Turbo	0.60	27.77 / 0.286	$+0.07$ / -0.012	-2.29 / -0.024
SD3.5 Turbo	0.58	27.42 / 0.280	n/r	-2.44 / -0.023

Quality-filtering robustness. To validate that the diversity gain is not carried by low-quality outputs, we progressively remove the lowest-HPS images per prompt at the main FLUX.1-schnell configuration (K=1024, α=0.5) on a 21-prompt $\times$ 100-image set. After discarding half of all images, STRIDE retains over 70% of its per-prompt Vendi advantage over the baseline ($+33.7\% \rightarrow$

$+23.9\%$), the InBSim advantage persists ($-0.065 \rightarrow -0.046$), and the CLIP advantage widens ($+0.46 \rightarrow +0.57$).

A.6 TRANSFER TO ADDITIONAL BACKBONES

We apply STRIDE to two additional backbones on DrawBench (199 prompts $\times$ 4 images), to assess transfer across architecture, model scales, and distillation methods. We retain patch size P=2, singular-value scaling, the few-step setting f_α=20, and injection into the first approximately 40% of dual-stream blocks. We also transfer the timestep-gating rule used in SD3 to both backbones, applying perturbations only at sampling step 0. For each backbone and fixed rank K, we calibrate only α through a seven-configuration sweep on a 30-prompt subset of DrawBench, requiring less than one GPU-hour per backbone. **FLUX.2-klein-4B** (4-step; a newer architectural generation, roughly three times smaller than FLUX.1) and **HunyuanImage-2.1-distilled** (Tencent; 17B MeanFlow-distilled DiT, 8 sampling steps, native 2048×2048, evaluated at its default guidance scale of 3.25) add a third vendor and an additional distillation family. Table S31 reports the results. On FLUX.2-klein-4B, STRIDE reduces InBSim from 0.736 to 0.686 (-6.8%) and raises V_D/p by 8.5% with CLIP slightly up ($+0.12$) and HPS unchanged (-0.0004; paired test in Table S7). The configurations $(K, \alpha) \in \{(256, 2.0), (512, 1.5), (1024, 1.0)\}$ yield similar quality–diversity trade-offs, supporting the selection of K according to computational constraints followed by one-dimensional calibration of α. On HunyuanImage-2.1, STRIDE reduces InBSim from 0.694 to 0.630 (α=14) and 0.604 (α=16), at CLIP costs of 0.43 and 0.67 and an HPS cost of 0.003 at both settings.

Table S31: **Transfer to additional backbones on DrawBench** (199 prompts $\times$ 4 images). All structural hyperparameters are carried over unchanged from the main configurations; only α is calibrated per backbone (30-prompt sweep). HunyuanImage-2.1 runs at its default guidance scale of 3.25 in all rows. These backbones were evaluated on DrawBench only.

Backbone	Method	InBSim↓	VD/p↑	CLIP↑	HPS↑
FLUX.2-klein-4B (4-step)	Baseline	0.736	1.977	28.08	0.2954
	+STRIDE (K=512, α=1.5)	0.686	2.146	28.20	0.2950
	+STRIDE (K=1024, α=1.0)	0.689	2.141	28.20	0.2946
	+STRIDE (K=256, α=2.0)	0.689	2.138	28.10	0.2946
HunyuanImage-2.1 (8-step, 2048^2)	Baseline (guidance 3.25)	0.694	2.117	26.43	0.292
	+STRIDE (K=1024, α=14)	0.630	2.324	26.00	0.289
	+STRIDE (K=1024, α=16)	0.604	2.400	25.76	0.289

A.6.1 APPLYING STRIDE TO A NEW BACKBONE

Applying STRIDE to a new compatible backbone follows three steps, which we used without modification for FLUX.2-klein-4B and HunyuanImage-2.1-distilled:

1. **Identify the injection blocks.** Locate the dual-stream, joint text–image transformer blocks and apply STRIDE to approximately the first 40% of them.
2. **Transfer the structural hyperparameters.** Retain patch size P=2, singular-value scaling, perturbation at sampling step 0 only, and the regime-specific frequency exponent: f_α=5 for one-step models and f_α=20 for few-step models.
3. **Fix the rank and calibrate the perturbation strength.** Select a practical rank K according to computational constraints, then calibrate only α. In our experiments, calibration used seven values on a 30-prompt subset of DrawBench and required less than one GPU-hour per backbone–rank pair. Rank and strength exhibit partial compensatory effects (Tables S17, S24, and S31), making calibration one-dimensional once K is fixed.

Further adjustments could include recalibrating f_α or changing the injection block range, although neither was required for intial calibration on these two backbones. STRIDE is designed for and validated on architectures with dual-stream, joint text–image transformer blocks. Its applicability to other architectures, remains a interesting future work.

A.7 Comparison with and application to a multi-step model (FLUX.1-dev)

STRIDE is scoped to distilled one-step and few-step models; a natural question is whether the same intervention helps, or is even needed, under a full multi-step schedule. Table S32 compares the distilled backbones with and without STRIDE against the 50-step FLUX.1-dev base model at its recommended guidance of 3.5 (DrawBench, 199 prompts $\times$ 4 images; identical batch settings for timing). Two observations follow. First, as an external anchor for the magnitude of the gain, FLUX.1-schnell+STRIDE at one step exceeds the 50-step base model on every diversity metric and on CLIP, at roughly $1/12$ of the wall-clock, with an HPS trade-off of 0.283 vs. 0.298; at $K{=}512$, HPS is within 0.004 of the schnell baseline (0.289 vs. 0.293) while still matching FLUX.1-dev's diversity (0.636 vs. 0.636). FLUX.1-dev is evaluated at its recommended guidance; lower guidance would increase its diversity at its own alignment cost, moving along the same trade-off. Second, applying a single step-0 injection to FLUX.1-dev itself produces only a small, non-significant gain at $K{=}1024$ ($p_{\text{holm}}{=}0.15$, $n{=}199$, paired) and no gain at lower ranks, and leaves its HPS unchanged: over 50 steps the corrective dynamics largely absorb an early perturbation, consistent with Limitation (i) in Sec. 5. Larger α or an annealed schedule (Future Work (iv)) may recover gains in the multi-step regime; we leave this to future work.

Table S32: **Distilled backbones vs. the 50-step FLUX.1-dev base model on DrawBench** (199 prompts $\times$ 4 images; FLUX.1-dev at its recommended guidance 3.5; STRIDE at $K{=}1024$ with the per-backbone locked configuration; a single step-0 injection for FLUX.1-dev). Wall-clock is measured under identical batch settings for all rows and is therefore not comparable with the batch-1 protocol of Table S11. †Not statistically significant ($p_{\text{holm}}{=}0.15$, $n{=}199$, paired).

Model (steps)	Method	InBSim↓	VD/p↑	CLIP↑	HPS↑	s/img
FLUX.1-schnell (1)	—	0.662	2.236	28.47	0.293	0.82
FLUX.1-schnell (1)	+STRIDE	0.590	2.495	28.61	0.283	2.14
SD3.5 Large Turbo (4)	—	0.707	2.067	28.18	0.292	1.13
SD3.5 Large Turbo (4)	+STRIDE	0.639	2.300	28.07	0.289	3.14
FLUX.1-dev (50)	—	0.636	2.288	27.76	0.298	24.90
FLUX.1-dev (50)	+STRIDE†	0.625	2.322	27.76	0.298	26.33

A.8 Qualitative Comparisons

We provide qualitative side-by-side comparisons between the baseline, the strongest competing methods, and STRIDE on each model. Each column shows 4 images generated from the same prompt with different seeds (matching the 4-images-per-prompt protocol used in our quantitative evaluation).

A.8.1 FLUX.1-schnell

Figure S3 shows STRIDE versus the baseline, CADS, and Input-Noise on three DrawBench prompts. STRIDE produces visibly more varied compositions, viewpoints, and styles while preserving prompt fidelity, whereas CADS often degrades fine detail.

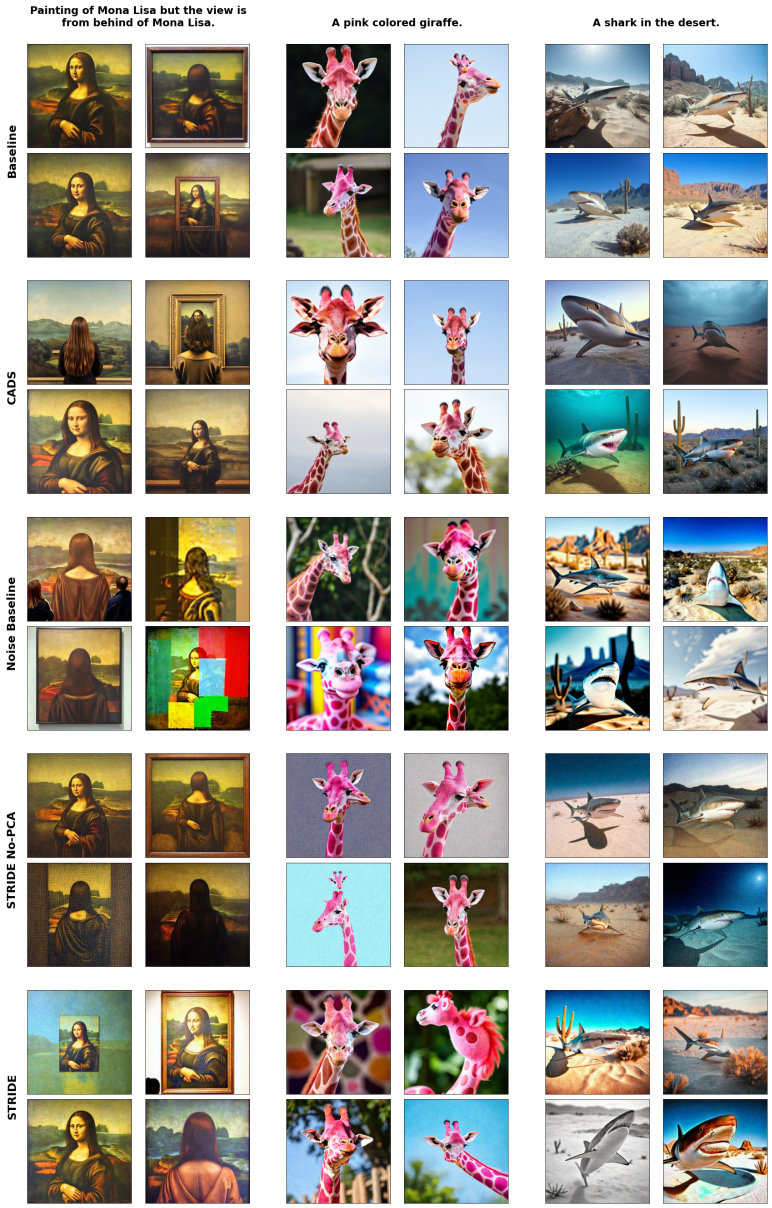


Figure S3: Qualitative comparison on FLUX.1-schnell. Each column corresponds to a different DrawBench prompt; each block of 4 images shows generations from a single method.

A.8.2 SD3.5 LARGE TURBO

Figure S4 shows the same comparison on SD3.5 Large Turbo (4-step). CADS collapses visibly on this backbone (loss of semantic coherence), while STRIDE expands compositional variety while keeping the prompt's content recognisable; DDC reaches comparable diversity at the cost of doubled inference-time VRAM.

Figure S5 shows the qualitative behaviour of STRIDE on SD3.5 Large Turbo across a sweep of α and f_α for a fixed DrawBench prompt. At α=0.1, the model produces smooth compositional variation up to $f_\alpha \approx 20$ before the perturbation begins to overwhelm the output ($f_\alpha \geq 30$); at α=0.2, the perturbation is already too aggressive at small f_α, with high-frequency artefacts visible from f_α=3 onward. The locked SD3.5 Turbo configuration (α=0.1, f_α=20) corresponds to the rightmost stable cell of the top row, balancing maximum diversity against quality preservation.

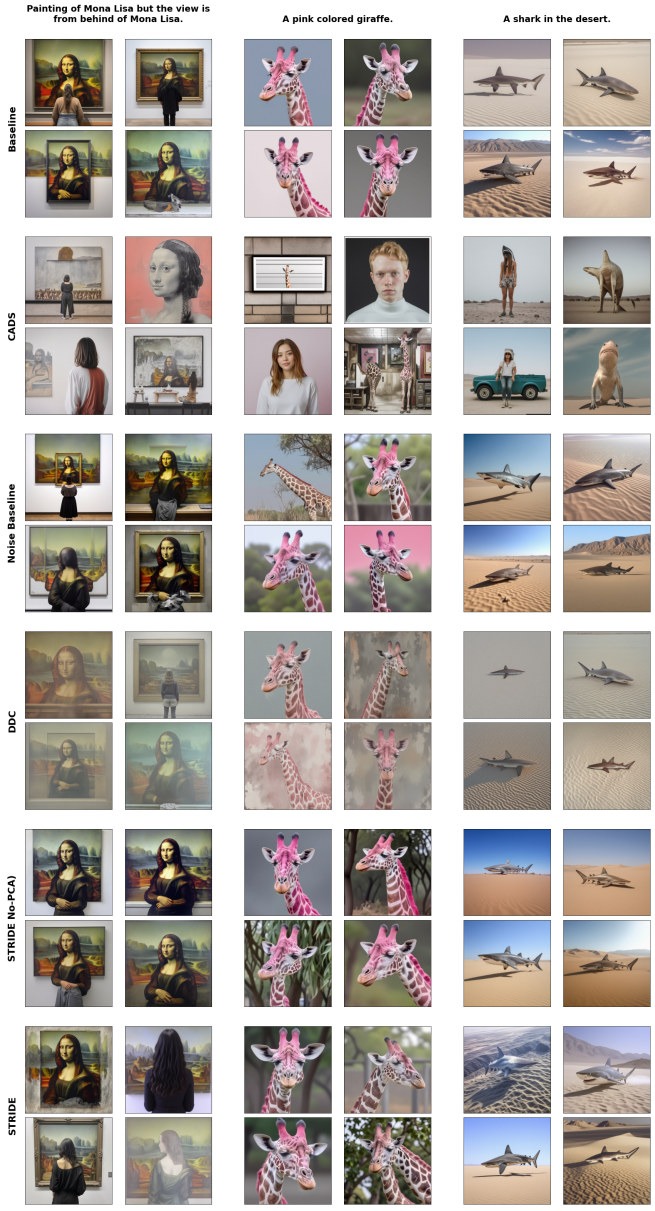


Figure S4: Qualitative comparison on SD3.5 Large Turbo. Each column corresponds to a different DrawBench prompt; each block of 4 images shows generations from a single method.

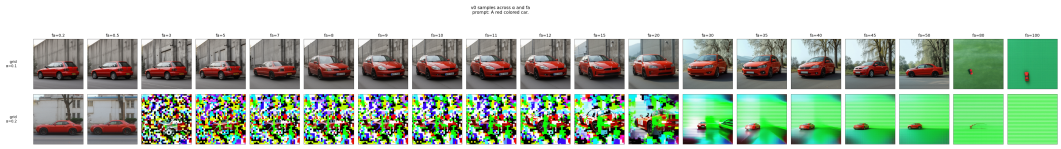


Figure S5: **Qualitative ablation on SD3.5 Large Turbo** for a fixed DrawBench prompt. Columns vary the frequency exponent f_α; rows vary the perturbation strength α.

A.9 Extended Limitations and Future Work

This appendix expands Section 5.

A.9.1 Extended Limitations

Fidelity regression at aggressive α and f_α. At our locked configurations (α=0.5, f_α=5 on FLUX.1-schnell; α=0.1, f_α=20 on SD3.5 Large Turbo), the CLIP cost relative to the unguided baseline is within 0.5 on COCO. As either parameter increases beyond the locked operating point, CLIP degrades monotonically, with the diversity-quality Pareto frontier traced in Section 4 (Figures 2a and 2b). Users prioritising strict text alignment over diversity should remain at or below the locked defaults.

Per-layer PCA cost. We compute truncated SVD (`torch.pca_lowrank`, q=K, two iterations) on the centered patch features activations of each hooked block. Costs scale with K and the number of hooked blocks (eight blocks at K=1024 for FLUX.1-schnell, twelve blocks at K=1024 for SD3.5 Large Turbo). The overhead is non-trivial relative to the underlying single- or four-step generation budget (+195% on FLUX.1-schnell and +167% on SD3.5 Turbo at K=1024, +90%/+76% at K=512, with no increase in peak VRAM; Table S11); whether a precomputed basis can recover this cost is the subject of Future Work (i).

Architecture-specific tuning. Both K and the hooked-block set must be tuned per backbone. We use K=1024 at P=2 with the first eight dual-stream blocks (L0–7 of 19) for FLUX.1-schnell, and K=1024 at P=2 with the first twelve blocks (L0–11 of 38) for SD3.5 Large Turbo, both gated to step 0 only. Tuning was coarse: a small grid of K values, layer-set windows, and timestep choices, reported in Tables S12, S18, and S20. Transferring STRIDE to a new architecture therefore requires re-identifying the early structure-forming blocks rather than reusing locked values. App. A.6.1 gives the procedure we used for FLUX.2-klein-4B and HunyuanImage-2.1, where all structural hyperparameters were transferred unchanged and only α was recalibrated.

Scope of evaluation. Our main results cover two adversarially-distilled MM-DiT families: FLUX.1-schnell (Black Forest Labs, 2024) (1-step) and SD3.5 Large Turbo (Stability AI, 2024) (4-step). DMD-family students (Yin et al., 2024b;a), consistency distillation (LCM, Hyper-SD), and score-identity distillation (Zhou et al., 2025) are untested. We conjecture that the conservation effect (the tendency of distilled models to produce *less diverse* outputs when injected with unstructured feature noise) is a property of the distilled denoising pathway rather than of any particular distillation objective, but verifying this requires running our ablation on other distillation families. The section App. A.6 reports transfer to FLUX.2-klein-4B (4-step) and HunyuanImage-2.1 (8-step, MeanFlow distillation), which partially widens this coverage. STRIDE is designed for and validated on architectures with dual-stream joint text–image transformer blocks (SD3, FLUX.1/2, HunyuanImage); its applicability to other designs has not been evaluated.

Behavioral evidence for the conservation effect. Our central observation, that unstructured feature noise *increases* intra-batch similarity, rests on behavioral evidence: pink noise at identical α applied with and without PCA projection produces opposite-direction shifts in InBSim, Table 1). A mechanistic account attributing this output-space inversion to specific transformer components would strengthen the scientific claim and is the subject of Future Work.

A.9.2 Extended Future Work

(i) Approximate or amortised PCA. If the top-K principal directions of a block were approximately stable across prompts, per-image PCA could be replaced by a basis computed once over a representative corpus, collapsing inference cost to a single matrix multiplication per hook. Our cross-prompt alignment analysis (Table S2) shows this stability holds for token-PCA but not for spatial-PCA, which is why STRIDE uses online computation. Whether prompt-clustered or domain-specific bases could recover the precomputability benefit while preserving the spatial-PCA structure is open.

(ii) Composition with complementary methods. STRIDE modifies *how* the model realises a prompt via per-sample feature perturbation. Conditioning-level methods (e.g., CADS (Sadat et al., 2023)) modify the conditioning signal across timesteps. Latent-level methods (Harrington et al., 2025) perturb the initial noise distribution. Post-hoc selection methods (e.g., SGI (Parmar et al., 2025)) operate at the batch level. Each targets a distinct source of sample variation, and the mechanisms are not in conflict; quantifying the gains from composing STRIDE with each is an open empirical question.

(iii) Mechanistic analysis of the conservation effect. Current evidence for the conservation effect is behavioral. A mechanistic account, attributing the off-manifold-noise inversion to specific attention patterns, normalisation layers, or representation-geometry properties, would strengthen the scientific claim and could suggest targeted interventions cheaper than online PCA.

(iv) Multi-step extension with annealed α. The conservation effect is defined and measured at $T \leq 4$. Whether STRIDE composes with multi-step sampling under an annealed α schedule (decaying as denoising progresses) requires running our ablation on flow-matching teachers at higher step counts.