BREAKING TO PRESERVE: THE GEOMETRY OF FEW-STEP DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step distillation is commonly viewed as compression, but this perspective overlooks what the student must acquire beyond the base model. Under six-dimensional input perturbations, six directions capture 99% of response variation in the SD1.5 base model and standard LoRA, compared with 26 for LCM-LoRA and 15 for DMD2 at one perturbation scale. Across six backbones, 90.5–99.2% of the energy in teacher–base response changes lies outside the base span; we call this out-of-span component the teacher margin. Exact matching requires recovering this margin. Any base-confined student therefore faces a matching-error floor, and feature spread alone neither recovers the margin direction nor preserves prompt following. Our breaking–fidelity decomposition reveals distinct learning requirements that equal total errors can obscure and motivates diagnostics of spread, reference alignment, and concentration. For a given starting recipe, these diagnostics predict which control is missing, explaining why the preferred complementary intervention reverses across recipes. On FLUX, complete samplewise and covariance students improve alignment with the teacher's added responses over trained distribution-matching (DM) at all 12 measured pairs and improve generation, while mean base-component error falls and complete response error rises. Translating these diagnostics into training objectives, we compare two implementations; covariance optimizes the diagnosed quantities more directly and outperforms its samplewise counterpart on five of six metrics on each backbone. Across five backbones, our one-step SD1.5 student improves all six metrics over official DMD2, and one-step klein beats its four-step counterpart on Patch FID-T.

1 INTRODUCTION

Few-step distillation changes more than sampling speed. What must fewer calls learn to replace many? We examine *responses*: output changes induced by common input perturbations. Figure 1 shows that loading a roughly 20 MB LCM-LoRA onto SD1.5 (Luo et al., 2023b; Rombach et al., 2022) raises its response dimension from 6 to 26. This dimension counts directions explaining 99% of variation under finite perturbations in a six-dimensional input family. Standard LoRA fine-tuning (Hu et al., 2022) keeps this count at 6, while DMD2 (Yin et al., 2024a) reaches 15.

Distillation compresses repeated computation into fewer evaluations (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024a). At identical weights, the multi-step teacher and one-call base implement different endpoint maps. The *base-relative teacher margin* comprises teacher responses outside the base span. Across six UNet, DiT, and MMDiT backbones, 90.5 to 99.2% of response-change energy between teacher and base lies outside the retained base span. Exact teacher-response matching requires acquiring this margin, so preserving the teacher requires leaving the base.

Base-confined responses cannot match the orthogonal margin, leaving an error floor set by its energy. Equal error, rank, norm, and teacher cosine can hide a missing margin or a lost base component. Our exact decomposition separates *target-aligned breaking*, which matches the margin, from *fidelity*, which matches the teacher's base component and suppresses extraneous responses. A spectral-tail bound quantifies the cost of missing margin directions.

These requirements motivate complementary image-feature controls. We diagnose *spread*, *reference alignment*, and *concentration* through variation breadth, agreement with teacher-reference geometry, and dominance by a shared component. To test whether spread alone preserves prompt following, we

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.

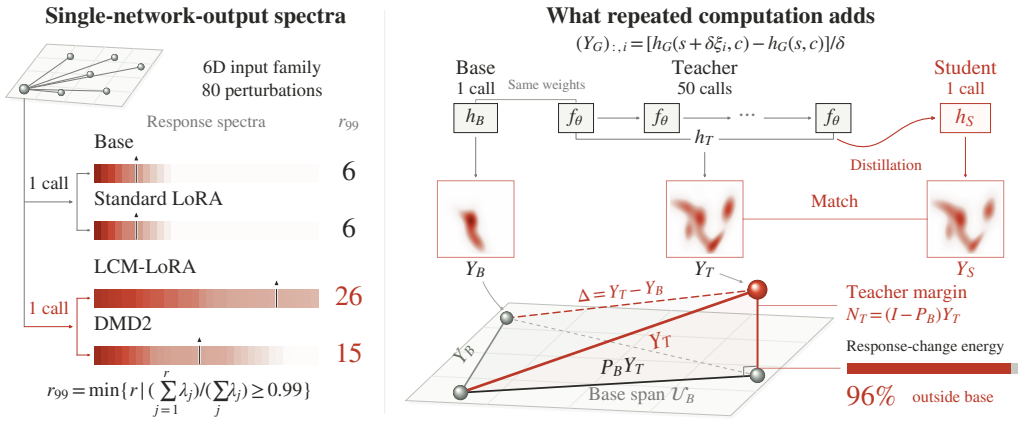


Figure 1: **What matching teacher responses requires.** (Left) One-call response dimension (official LCM-LoRA at perturbation scale 0.1). (Right) Teacher margin $N_T = (I - P_B)Y_T$ and response-change energy between teacher and base outside the base span. Protocols in Appendices A.1 and A.2.

introduce FeatNN, a feature nearest-neighbor baseline. Pooling references across prompts preserves broad centered variation while nearly erasing prompt following.

Diagnostics favor alignment for LCM, which has stronger spread, and expansion for FeatNN, which has stronger reference alignment. Samplewise control swaps confirm this reversal. Our final recipe combines distribution matching (DM), covariance expansion, within-prompt covariance alignment, and shared adversarial refinement.

Paired FLUX measurements show improved teacher-margin alignment for complete samplewise and covariance students over trained DM at all 12 measured pairs across four prompts. With DM and GAN fixed, covariance expansion produces the larger increase in the base-external norm ratio, while alignment yields the larger angular gain and reduction in base-component error. Both controls improve generation individually and jointly across three training seeds (Appendices A.5 and F.7). With shared control weights across five backbones, our one-step SD1.5 student improves all six metrics over official one-step DMD2, and one-step klein beats the official four-step model on Patch FID-T.

We make three contributions.

- **Exact teacher-response matching requires acquiring the teacher margin.** One-call spectra and six-backbone measurements reveal what repeated calls add beyond the base.
- **An exact decomposition separates breaking and fidelity.** We characterize the minimal target space and the tradeoff between rank and accuracy in matching it.
- **Geometry-guided controls improve few-step distillation.** Control swaps validate diagnostic predictions. Paired responses track teacher-margin alignment. Covariance improves on samplewise controls on SD1.5 and FLUX and supplies the five-backbone recipe.

2 Related Work

Denoising diffusion and score-based SDEs supply the teacher computation that few-step distillation compresses (Ho et al., 2020; Song et al., 2021b), with flow matching as a parallel ODE view (Lipman et al., 2023). Trajectory-based supervision compresses deterministic paths (Salimans & Ho, 2022), learns shared endpoints (Song et al., 2023; Luo et al., 2023a), transfers this through lightweight adapters (Luo et al., 2023b), broadens the map to arbitrary times (Kim et al., 2024), or learns interval-average velocities from scratch (Geng et al., 2025). Distribution-level supervision matches generated distributions (Yin et al., 2024b;a) or adversarial image and latent features (Sauer et al., 2025; 2024). Improvements to starting points and supervision include trajectory-consistent initialization and diffusion-pretrained GAN conversion (Hu et al., 2026; Zheng & Yang, 2025); revised consistency

losses, noise schedules, and continuous-time forms (Song & Dhariwal, 2024; Lu & Song, 2025); and complementary signals from latent adversarial supervision, intermediate latents, intra-segment guidance, and score regularization (Chen et al., 2025; Li & Bai, 2026; Ge et al., 2026; Zheng et al., 2026).

Less studied is what distillation changes in the model. Progressive/adversarial updates and trajectory-segmented consistency maps encode the change implicitly (Lin et al., 2024; Ren et al., 2024); WaDi models dominant weight-direction changes with low-rank rotations (Wang et al., 2026). At the functional level, DMD students reproduce teacher noise–data pairings despite distributional supervision (Li et al., 2026), concept-control structure survives distillation despite early structural commitment (Gandikota & Bau, 2026), and Jacobian-vector alignment restores noise sensitivity (Huang et al., 2026). We instead take the single-call base as a functional reference, separating teacher responses within its span from those added by repeated computation.

Feature geometry guides variation without collapse. Alignment–uniformity separates pair agreement from embedding spread (Wang & Isola, 2020), VICReg adds variance and covariance terms (Bardes et al., 2022), CKA measures representation similarity (Kornblith et al., 2019), and REPA aligns denoising representations with pretrained visual features (Yu et al., 2025). We instead read generated-image features to pick complementary expansion and alignment controls.

3 Fidelity Requires Breaking

Repeated computation changes the function without changing the weights. The opening assay measures single-call variation; we now compare the base's one-step estimate with the teacher's full trajectory under common probes.

3.1 What repeated computation adds

Let h_B, h_T, h_S denote output maps from a state s and condition c. The base map h_B returns a clean-sample (x_0) estimate from one network evaluation; h_T and h_S return the 50-step teacher and distilled one-step student endpoints, respectively. At a shared (s, c), common unit-norm perturbation probes $\xi_1, \ldots, \xi_n$ with finite-difference step $\delta > 0$ give

$$(Y_G)_{:,i} = \frac{h_G(s + \delta\xi_i, c) - h_G(s, c)}{\delta}, \qquad Y_G \in \mathbb{R}^{p\times n}, \quad G \in \{B, T, S\}. \tag{1}$$

Each column is an output change flattened into p coordinates. Projecting onto the base span gives $Y_T = P_B Y_T + N_T$; its orthogonal remainder N_T is the *base-relative teacher margin*:

$$\mathcal{U}_B = \mathrm{col}(Y_B), \quad P_B = \mathrm{Proj}_{\mathcal{U}_B}, \quad N_T = (I - P_B)Y_T, \quad \mathcal{U}_N = \mathrm{col}(N_T). \tag{2}$$

As proved in Appendix B.1, $\mathcal{U}_N$ is the unique smallest base-orthogonal addition satisfying $\mathrm{col}(Y_T) \subseteq \mathcal{U}_B \oplus \mathcal{U}_N$.

Exact response matching requires $(I - P_B)Y_S = N_T$: a nonzero margin forces departure from the base. Define $\mathcal{L}_{\mathrm{resp}} = \|Y_S - Y_T\|_F^2/(2n)$. For $\mathrm{col}(Y_S) \subseteq \mathcal{U}_B$, orthogonality gives

$$\mathcal{L}_{\mathrm{resp}} = \frac{\|Y_S - P_B Y_T\|_F^2}{2n} + \frac{\|N_T\|_F^2}{2n} \geq \frac{\|N_T\|_F^2}{2n}. \tag{3}$$

Even the optimum $Y_S = P_B Y_T$ leaves the entire margin unmatched. Table B.1 illustrates the gap on FLUX.1-dev. Starting from $Y_S = Y_B$, a base-only correction lowers $\mathcal{L}_{\mathrm{resp}}$ from 2,493 to 2,427, whereas a margin-only correction reaches 66. Most of the response gap remains because base-only corrections cannot recover the teacher margin.

3.2 How large is the margin?

Across six backbones (five in Section 5.1 plus SD3.5-Medium), we use $\delta = 0.05$ and $n \in \{32, 64, 128\}$, following the protocol in Appendix A.2. $\sigma_j(A)$ denotes the j-th largest singular value of A. The numerical base projector $\widehat{P}_B$ retains left singular vectors whose singular values exceed $0.01\sigma_1(Y_B)$, yielding the measured margin $Z_T = (I - \widehat{P}_B)Y_T$.

For $\Delta = Y_T - Y_B$, the outside-base change-energy share is

$$\rho_\Delta = \frac{\|(I - \widehat{P}_B)\Delta\|_F^2}{\|\Delta\|_F^2}. \tag{4}$$

It is 90.5–99.2%, averaging 96.0% at $n = 64$, with isotropic reference $1 - r_B/p$ for the retained base rank r_B (Appendix C.1). Most teacher–base change lies outside the retained base span. Appendix Figure A.1a shows 69–118 margin directions above the spectral cutoff with 128 probes. Appendix Figure A.1b shows early target formation. By five teacher sampling steps, the margin-to-teacher Frobenius-norm ratio reaches 0.950–0.998 across all six backbones.

In a separate SD1.5 full-UNet-output assay, we count directions in each complete, unprojected response matrix. The tail-based threshold is $\max(3f, 0.01\sigma_1)$, where f is the median of the trailing half of its singular spectrum. Applying this shared rule in Table A.6 resolves one significant direction for standard LoRA and 36 for the multi-step teacher.

All six one-step students in Table A.7 occupy the retained margin space, with thresholded ranks of 26–61. What has the student preserved, and what has it acquired?

3.3 Same summaries, different learned content

Two students can score identically yet need opposite corrections. For orthogonal unit e_1, e_2, let $n = 1$, $Y_B = e_1$, $Y_T = e_1 + e_2$, and $Y_{S_a} = e_1$, $Y_{S_b} = e_2$. S_a misses the margin; S_b loses the base component. Both have rank one, unit norm, squared teacher error one, and teacher cosine $1/\sqrt{2}$.

Orthogonal projection separates these errors. Set $P_N = \mathrm{Proj}_{\mathcal{U}_N}$ and $P_\perp = I - P_B - P_N$.

Theorem 1 (Exact breaking–fidelity characterization)**.** *For every student response matrix Y_S,*

$$\mathcal{L}_{\mathrm{resp}} = \underbrace{\frac{\|P_N(Y_S - Y_T)\|_F^2}{2n}}_{\mathcal{L}_{\mathrm{br}}} + \underbrace{\frac{\|P_B(Y_S - Y_T)\|_F^2 + \|P_\perp Y_S\|_F^2}{2n}}_{\mathcal{L}_{\mathrm{fid}}}, \tag{5}$$

Consequently, $Y_S = Y_T$ if and only if $\mathcal{L}_{\mathrm{br}} = \mathcal{L}_{\mathrm{fid}} = 0$.

Proof. Decompose $E = Y_S - Y_T$ into its three orthogonal projections. Since $P_\perp Y_T = 0$, Pythagoras gives the identity; both terms vanish exactly when $E = 0$. □

The *target-aligned breaking* term measures teacher-margin error; *fidelity* measures base-component error and responses outside the combined span. The example yields $(\mathcal{L}_{\mathrm{br}}, \mathcal{L}_{\mathrm{fid}}) = (1/2, 0)$ for S_a and $(0, 1/2)$ for S_b. Neither error component can compensate for the other.

3.4 How many directions, and how well aligned?

With room for only k margin directions, which should the student learn? The teacher spectrum identifies the best choice and the cost of leaving other directions unmatched.

Proposition 3.1 (Spectral cost of missing target directions)**.** *If $\mathrm{rank}(P_N Y_S) \le k$, then*

$$\mathcal{L}_{\mathrm{br}} \ge \frac{1}{2n} \sum_{j>k} \sigma_j(N_T)^2. \tag{6}$$

Proof. Apply the Eckart–Young theorem (Eckart & Young, 1936) to $P_N Y_S - N_T$. □

The construction $Y^{(k)} = P_B Y_T + (N_T)_k$, with $(N_T)_k$ the rank-k truncated SVD, attains the bound at zero fidelity error, preserving the base component and learning the k strongest margin directions.

Corollary 1 (Target-space rank)**.** *Let $q = \mathrm{rank}(N_T) > 0$ and $\gamma = \sigma_q(N_T) > 0$. Then*

$$\mathcal{L}_{\mathrm{br}} < \frac{\gamma^2}{2n} \quad \Longrightarrow \quad \mathrm{rank}(P_N Y_S) = q. \tag{7}$$

A target-space rank below q leaves at least $\gamma^2/(2n)$ breaking error.

With $r_B = \mathrm{rank}(Y_B)$, if $\mathcal{U}_B \subseteq \mathrm{col}(Y_T)$ then matching both components forces $\mathrm{rank}(Y_S) \geq r_B + q$: $P_B Y_S = P_B Y_T$ and $\mathcal{L}_{\mathrm{br}} = 0$ suffice (Appendix B.3).

Rank counts directions, not probe-wise agreement. On FLUX.1-dev, a random correction in the retained margin span, scaled to the retained teacher margin's Frobenius norm and added to Y_B, raises $\mathcal{L}_{\mathrm{resp}}$ from 2,493 to 4,927. Table B.2 reports approximately zero teacher-aligned reconstruction error on FLUX and lower errors than random reconstruction on all six backbones. For nonzero $A_S = (I - \widehat{P}_B)Y_S$ and Z_T, define *teacher-margin alignment*:

$$\mathrm{Align}_{\mathrm{margin}}(S) = \frac{\langle A_S, Z_T \rangle_F}{\|A_S\|_F \, \|Z_T\|_F}. \tag{8}$$

This cosine compares base-external changes on corresponding probes. Together with their base-external norm ratio $\rho = \|A_S\|_F / \|Z_T\|_F$, it determines normalized squared mismatch, as shown in Appendix B.2.

The decomposition motivates image-feature controls. Control swaps test diagnostics; shared-probe FLUX measurements track teacher-margin alignment in samplewise and covariance students.

4 Learning the Right Departure

4.1 Variation needs a conditional target

A student can retain variation and lose the prompt. Our SD1.5 feature nearest-neighbor baseline, *FeatNN*, uses the teacher-reference matching objective in Appendix D.1. Table 1 exposes the failure. Changing only reference selection from prompt-associated references to random pooled draws leaves centered spread broad (19.32 to 18.16). Yet GenEval (Ghosh et al., 2023), which measures compositional prompt following, falls from 0.2076 to 0.0042, while Patch FID-T rises from 83.47 to 449.25. Both prompt following and generation quality deteriorate.

Concentration rises from 0.4127 to 0.9983. Almost all uncentered feature energy lies in one direction. For mean $\bar{f}$, covariance C, and second moment M, the identity $M = C + \bar{f}\bar{f}^\top$ explains how a dominant shared component coexists with broad centered spread. Appendix C.4 contrasts prompt-specific conditional targets with the shared target of pooled references.

Table 1: **Variation needs a target and complementary controls.** Bold marks diagnostic strengths among the first three SD1.5 recipes; pooled FeatNN changes only reference selection.

Method	Spread	Ref. align.	Conc.	P-FID-T↓	GenEval↑
LCM official (Luo et al., 2023a)	**23.76**	0.8897	0.4416	90.90	0.2141
DMD2 official (Yin et al., 2024a)	18.99	0.8960	**0.3332**	18.75	0.3569
FeatNN	19.32	**0.9184**	0.4127	83.47	0.2076
FeatNN (pooled)	18.16	0.4676	0.9983	449.25	0.0042

4.2 Diagnose the starting recipe

Frozen DINOv2 features (Oquab et al., 2024), averaged over each image's patch tokens, yield diagnostics of *spread* (c_{div}), the breadth of centered variation; *reference alignment* (ra), overlap with the reference's leading cross-prompt subspace; and *concentration* (cm^2), the leading direction's share of uncentered energy. The diagnostic cohort contains 32 prompts with eight images each. Let λ_i, μ_i be the decreasing eigenvalues of C, M over all 256 features. Let U_r, U_r^{ref} be the leading $r = 8$ covariance directions of the 32 prompt means (Appendix C.1), and define

$$c_{\mathrm{div}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \qquad \mathrm{ra} = \frac{\|U_r^\top U_r^{\mathrm{ref}}\|_F^2}{r}, \qquad \mathrm{cm}^2 = \frac{\mu_1}{\sum_i \mu_i}. \tag{9}$$

The baseline objectives and training setups in Appendix D.1 define our *starting recipes*. Table 1 contrasts LCM's stronger spread with FeatNN's stronger reference alignment. These profiles suggest

that, *with the same adversarial complement, LCM should benefit more from alignment, FeatNN more from expansion.* The FLUX distribution-matching (DM) baseline has low spread (6.87), low reference alignment (0.4823), and high concentration (0.7842), motivating both.

4.3 From complementary controls to covariance objectives

Samplewise controls test complementarity in control-swap and cumulative studies; covariance controls form our final DM recipe, reported as Ours in Table 4. Both augment the starting objective:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{distill}} + \lambda_{\text{exp}}\mathcal{L}_{\text{exp}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}. \tag{10}$$

Control swaps activate one feature term; the final recipe uses both.

For the frozen DINOv2 map ϕ, let $f_i = \phi(x_i)$ and $f_i^{\text{ref}} = \phi(r_i)$ for generated image x_i, teacher reference r_i, and shared prompt c_i. Over a global training batch of $m = 32$ pairs, samplewise expansion penalizes a shared direction through cross-sample cosine similarity:

$$\mathcal{L}_{\text{exp}} = \frac{1}{m(m-1)} \sum_{i \neq j} \cos(f_i, f_j). \tag{11}$$

With $\mathcal{I}_- = \{(i,j) : c_i \neq c_j\}$, samplewise alignment rewards same-prompt references and penalizes different-prompt references:

$$\mathcal{L}_{\text{align}} = -\frac{1}{m} \sum_i \cos(f_i, f_i^{\text{ref}})^2 + \frac{1}{|\mathcal{I}_-|} \sum_{(i,j)\in\mathcal{I}_-} \cos(f_i, f_j^{\text{ref}})^2. \tag{12}$$

For the final recipe, we move from pairwise similarities to distribution geometry, directly targeting global spread and within-prompt reference agreement. Let C_{gen} be the covariance centered over the full generated batch, and $C_{\text{gen},g}, C_{\text{ref},g}$ the generated and reference covariances centered separately within prompt group $g \in \mathcal{G}$. We minimize

$$\mathcal{L}_{\text{exp}}^{\text{cov}} = -\frac{(\operatorname{tr} C_{\text{gen}})^2}{\operatorname{tr}(C_{\text{gen}}^\top C_{\text{gen}}) + \epsilon_{\text{num}}}, \tag{13}$$

$$\mathcal{L}_{\text{align}}^{\text{cov}} = -\frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \frac{\operatorname{tr}(C_{\text{gen},g}^\top C_{\text{ref},g})}{\sqrt{\|C_{\text{gen},g}\|_F^2 + \epsilon_{\text{num}}}\sqrt{\|C_{\text{ref},g}\|_F^2 + \epsilon_{\text{num}}}}, \tag{14}$$

with numerical stabilizer $\epsilon_{\text{num}} = 10^{-6}$. Expansion optimizes the spread measured by c_{div}; alignment compares group covariances through normalized Frobenius inner products (Kornblith et al., 2019). Unlike cross-prompt ra, it targets variation around each prompt's mean. Shared adversarial refinement trains a hinge-loss discriminator head on frozen visual features of generated images and fixed teacher references. Following Appendix D.2, we share active weights λ_{exp}=1, λ_{align}=100, and λ_{adv}=2 across backbones.

5 Experiments

We test whether diagnostics identify the more effective control and whether generation gains accompany teacher-margin alignment, then compare loss formulations and five-backbone performance.

5.1 Setup

We initialize students on SD1.5 and SDXL (Rombach et al., 2022; Podell et al., 2024), SD3.5-Large (Esser et al., 2024), FLUX.1-dev (Labs, 2024), and klein-4B (Labs, 2025) and train all parameters with AdamW (Loshchilov & Hutter, 2019). Training uses 500,000 fixed prompt–teacher-image pairs from a filtered LAION-5B subset (Schuhmann et al., 2022), with offline teacher images.

Evaluation follows SenseFlow (Ge et al., 2026), including COCO-5K caption selection (Lin et al., 2014), metrics, and the GenEval protocol (Ghosh et al., 2023). We report Patch FID-T, patch-level FID against teacher references, CLIP Score (Radford et al., 2021), HPSv2 (Wu et al., 2023), PickScore (Pick) (Kirstain et al., 2023), and ImageReward (IR) (Xu et al., 2023). Training and evaluation prompts do not overlap. For our evaluations, each metric is computed separately for three seeds and averaged. Appendix D gives the full training and evaluation settings.

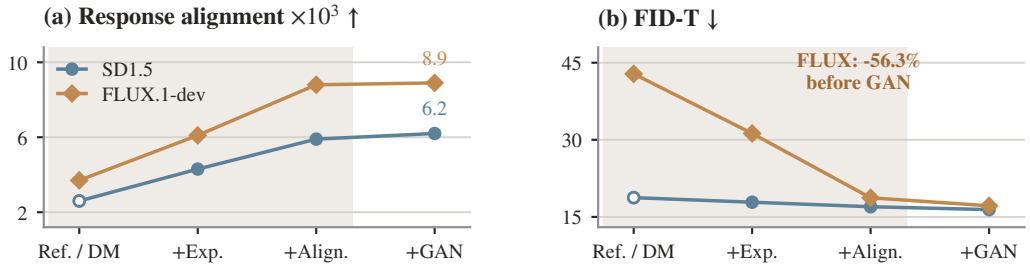


Figure 2: **Closer teacher-margin alignment accompanies better generation.** Mean teacher-margin alignment over 12 state–prompt pairs per backbone. Lines compare configurations. Open SD1.5 markers identify official DMD2 (Yin et al., 2024a); our students start from pretrained bases.

5.2 Testing the predicted complement

Diagnostics predict that LCM benefits more from alignment and FeatNN from expansion. We compare these samplewise controls within each recipe, holding GAN, data, updates, and settings fixed.

Table 2: **Control swaps on one-step SD1.5.** Each row adds the listed control and GAN; bold marks the better scores within each starting recipe.

Starting recipe	Added control	P-FID-T↓	GenEval↑
LCM (Luo et al., 2023a)	Alignment	**43.28**	**0.2874**
	Expansion	68.34	0.2358
FeatNN	Expansion	**41.75**	**0.2831**
	Alignment	65.92	0.2331

Table 2 confirms both predictions. Although the alternative controls give LCM greater spread and FeatNN stronger reference alignment, correcting the diagnosed deficiency wins on both generation metrics. Both preferences persist across weights; base-relative responses guide feature-matched choices (Appendices F.2, F.10).

5.3 From feature controls to teacher-margin alignment

Do feature controls improve teacher-margin alignment? FLUX successively adds expansion, alignment, and GAN to DM, yielding four samplewise configurations. SD1.5 compares three from-base configurations against official DMD2. Alignment averages four prompts and three initial noises, giving 12 state–prompt pairs per backbone sharing probes and a base projector (Appendix A.4).

On FLUX, expansion and alignment raise mean teacher-margin alignment from 0.0037 to 0.0088 and lower Patch FID-T from 42.83 to 18.72. Adding alignment and GAN gives the same trend across three trained SD1.5 configurations (Figure 2). GAN further reduces concentration and Patch FID-T on both backbones (Appendix F.1).

At the matched measurements in Appendix B.5, the complete samplewise SD1.5 and FLUX students have larger complete response errors than their bases. Cosine measures angular agreement; complete error also includes magnitude, base-component, and residual mismatches.

Complete samplewise and covariance students improve teacher-margin alignment over trained DM at all 12 FLUX pairs across four prompts. Mean alignment is 0.0037, 0.0089, and 0.0093, with student-to-teacher base-external norm ratios of 0.55, 0.85, and 0.87. Mean complete error rises. Increased retained-margin (e_N) and residual (e_R) errors outweigh reduced base-component error (e_B, Appendix A.5). Paired accounting resolves these changes across scales and thresholds, with FP32 retaining direction gains (Appendices A.5–A.7). With DM and GAN fixed, the combined covariance controls reduce mean FLUX Patch FID-T from 28.67 to 17.06 across three training

Table 3: **Samplewise versus covariance controls.** Both retain DM (Yin et al., 2024a) and GAN; bold marks the better score within each backbone.

Backbone / objective	P-FID-T↓	CLIP↑	HPSv2↑	Pick↑	IR↑	GenEval↑
SD1.5 (Rombach et al., 2022) / Samplewise	16.42	0.3054	0.2478	**21.0524**	-0.0851	0.3826
Covariance	**16.27**	**0.3078**	**0.2491**	21.0200	**0.0413**	**0.3854**
FLUX (Labs, 2024) / Samplewise	17.16	**0.3214**	0.3066	22.4162	1.1472	0.6551
Covariance	**17.04**	0.3212	**0.3131**	**22.4287**	**1.2158**	**0.6598**

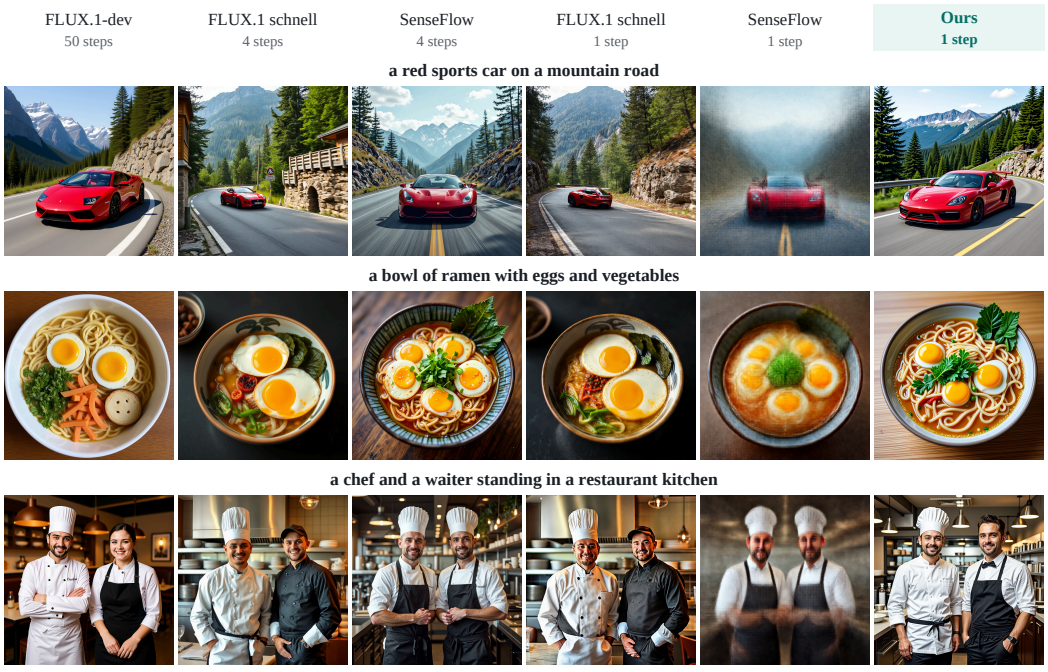


Figure 3: **Sharper one-step FLUX generation.** Our student produces sharper details and more coherent local structure than one-step schnell and SenseFlow. Rows share prompts; one-step SenseFlow directly uses released four-step weights.

seeds (Appendix F.5). Complementary feature controls thus improve teacher-margin alignment and provide generation gains beyond DM and GAN alone.

5.4 Two implementations of complementary controls

Samplewise and covariance implementations share DM, GAN, data, training budget, control weights, and the eight-prompt, four-sample batch structure and reference pairing (Appendix D.3). On SD1.5 and FLUX, covariance improves five of six mean scores in Table 3, with samplewise leading on SD1.5 PickScore and FLUX CLIP (Appendix F.3). Both formulations improve generation; covariance's broader gains and closer match to the diagnostic goals motivate our final recipe (Ours).

5.5 One-step generation across five backbones

We keep the same control weights across five backbones. Table 4 combines our evaluations with baseline scores from Ge et al. (2026), all under the common protocol. Each student step uses one network evaluation; Appendix E.1 gives complete results and sampling budgets.

Our one-step SD1.5 student improves all six metrics over official one-step DMD2. On klein, one step reaches Patch FID-T 13.67, compared with 17.26 for the official four-step model. At one step

Table 4: **One-step generation with four-step baselines.** SenseFlow protocol (Ge et al., 2026); baseline scores are measured by us (†) or quoted from SenseFlow. Bold marks the best among multiple scores per backbone and budget; dashes indicate unreported scores. One-step SenseFlow uses adapted weights; one-step klein uses four-step weights directly.

Method	Steps↓	P-FID-T↓	CLIP↑	HPSv2↑	Pick↑	IR↑	GenEval↑
SD1.5							
LCM (Luo et al., 2023a)†	1	90.90	0.2760	0.1486	19.18	-1.5904	0.2141
DMD2 (Yin et al., 2024a)†	1	18.75	0.2997	0.2369	20.87	-0.1712	0.3569
Ours	1	**16.27**	**0.3078**	**0.2491**	**21.02**	**0.0413**	**0.3854**
SDXL							
LCM (Luo et al., 2023a)	4	30.63	0.3230	0.2824	22.22	0.5693	0.5036
PCM (Wang et al., 2024)	4	17.77	0.3242	0.2920	22.54	0.6926	0.4944
Flash (Chadebec et al., 2025)	4	23.24	0.3216	0.2830	22.17	0.4295	0.4715
Lightning (Lin et al., 2024)	4	**16.57**	0.3214	0.2931	22.80	0.7799	0.5332
Hyper (Ren et al., 2024)	4	17.49	0.3254	0.3000	22.98	0.9777	0.5398
DMD2 (Yin et al., 2024a)	4	18.72	**0.3277**	0.2963	22.98	0.9324	0.5779
SenseFlow (Ge et al., 2026)	4	21.01	0.3248	**0.3010**	**23.17**	**0.9951**	**0.5784**
SenseFlow (Ge et al., 2026)	1	–	**0.3298**	0.2818	22.24	0.8570	–
Ours	1	19.77	0.3289	**0.3034**	**22.61**	**1.0584**	0.5906
SD3.5-Large							
Turbo (Sauer et al., 2024)	4	22.88	0.3262	0.2909	22.89	1.0116	0.6877
SenseFlow (Ge et al., 2026)	4	**17.48**	**0.3286**	**0.3016**	**23.01**	**1.1713**	**0.6955**
SenseFlow (Ge et al., 2026)	1	–	0.3332	0.2803	22.32	1.0651	–
Ours	1	15.94	**0.3363**	**0.3056**	**22.53**	**1.1938**	0.7112
FLUX.1-dev							
schnell (Labs, 2024)	4	–	**0.3264**	0.2962	22.77	1.0755	**0.6807**
Hyper (Ren et al., 2024)	4	23.47	0.3238	0.2963	23.09	**1.0983**	0.6193
Turbo-Alpha (Team, 2024)	4	24.52	0.3218	0.2907	22.89	1.0106	0.4724
SenseFlow (Ge et al., 2026)	4	**19.60**	0.3167	**0.2997**	**23.13**	1.0921	0.6471
schnell (Labs, 2024)†	1	20.66	0.3180	0.2915	**22.64**	1.0220	0.6572
Ours	1	**17.04**	**0.3212**	**0.3131**	22.43	**1.2158**	**0.6598**
klein-4B							
Official distilled (Labs, 2025)†	4	17.26	0.3122	0.2999	23.05	1.1226	0.8282
Official distilled (Labs, 2025)†	1	54.84	0.3061	0.2147	20.96	0.1536	0.5746
Ours	1	**13.67**	**0.3223**	**0.2837**	**22.35**	**1.0402**	**0.7882**

on FLUX, our student has better mean scores on five of six metrics than official FLUX.1-schnell, including Patch FID-T of 17.04 versus 20.66; schnell has the higher PickScore, 22.64 versus 22.43. The same recipe reaches one-step Patch FID-T 19.77 on SDXL and 15.94 on SD3.5-Large. Figure 3 shows the qualitative gains on FLUX.

Fixed-weight evaluations at one, two, and four steps reveal metric tradeoffs across five backbones. At four steps, covariance students on SDXL and SD3.5-Large and the samplewise FLUX student improve Patch FID-T over SenseFlow; klein exceeds its 50-step teacher on all five non-FID metrics. Complementary controls remain effective across backbones and budgets (Appendices E.2 and E.3).

6 Conclusion

Exact teacher-response matching requires acquiring the teacher margin. Our decomposition separates complementary errors. Swaps confirm predicted preferences for alignment in LCM and expansion in FeatNN. On FLUX, complete samplewise and covariance students improve teacher-margin alignment at all 12 measured pairs; mean base-component error falls while complete response error rises. Covariance supplies the five-backbone recipe. Break from the base to preserve the teacher.

AI USE STATEMENT

In this work, we used generative AI tools to aid manuscript writing and language polishing. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

Faster image generation increases the importance of responsible deployment. Applications should respect model and data licenses, assess inherited biases, and maintain safeguards against misleading and harmful imagery.

REPRODUCIBILITY STATEMENT

Appendices A and B provide response protocols, numerical conventions, proofs, and constructions. Appendix C develops the feature and distributional arguments. Appendix D specifies training, reference construction, preprocessing, sampling, and evaluation. Appendices E and F contain the complete generation and training-comparison results.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Springer Berlin Heidelberg, 2006. doi: 10.1007/3-540-33099-2. URL https://doi.org/10.1007/3-540-33099-2.

Clément Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15686–15695, 2025. doi: 10.1609/aaai.v39i15.33722. URL https://doi.org/10.1609/aaai.v39i15.33722.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16185–16195, 2025. doi: 10.1109/ICCV51701.2025.01502. URL https://doi.org/10.1109/ICCV51701.2025.01502.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. URL https://doi.org/10.1007/BF02288367.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Herbert Federer. *Geometric Measure Theory*. Springer Berlin Heidelberg, 1996. doi: 10.1007/978-3-642-62010-2. URL https://doi.org/10.1007/978-3-642-62010-2.

Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1304–1313, 2026. doi: 10.1109/WACV61042.2026.00133. URL https://doi.org/10.1109/WACV61042.2026.00133.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. Senseflow: Scaling distribution matching for flow-based text-to-image distillation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.),

International Conference on Learning Representations, volume 2026, pp. 92014–92045, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/94a98f4338b6e5fba81344f961bac8e5-Paper-Conference.pdf.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 75460–75482. Curran Associates, Inc., 2025. doi: 10.52202/085713-2534. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6d13e085b79d454da5910e4ca82a3d9d-Paper-Conference.pdf.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023. doi: 10.52202/075280-2270. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Paper-Datasets_and_Benchmarks.pdf.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

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

Zheyuan Hu, Chieh-Hsin Lai, Yuki Mitsufuji, and Stefano Ermon. Cmt: Mid-training for efficient learning of consistency, mean flow, and flow-map models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 155363–155404, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fb44a668c2d4bc984e9d6ca261262cbb-Paper-Conference.pdf.

Huayang Huang, Ruoyu Wang, Jinhui Zhao, Wei Deng, Daiguo Zhou, Jian Luan, Yu Wu, and Ye Zhu. Restoring initial noise sensitivity in text-to-image distillation via geometric alignment. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=dlufiUI3xI.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c204d12afa0175285e5aac65188808b4-Paper-Conference.pdf.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In A. Oh, T. Naumann,

A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023. doi: 10.52202/075280-1594. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Shucheng Li, Iolo Jones, Alexander Tong, and Michael M. Bronstein. Why are DMD students lazy? understanding the copying behavior in few-step distillation. In *High-dimensional Learning Dynamics 2026*, 2026. URL https://openreview.net/forum?id=nlXmmmS7Fc.

Zilai Li and Lujia Bai. Empurple: A free lunch for diffusion distillation based on the information bottleneck, 2026. URL https://arxiv.org/abs/2607.04276.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024. URL https://arxiv.org/abs/2402.13929.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7e9c2053258b1bdd32ff2654802cd594-Paper-Conference.pdf.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023a. URL https://arxiv.org/abs/2310.04378.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023b. URL https://arxiv.org/abs/2311.05556.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 117340–117362. Curran Associates, Inc., 2024. doi: 10.52202/079017-3725. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d4e1c24ac41ff0b82ca1b171731f0b23-Paper-Conference.pdf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

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*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687625. URL https://doi.org/10.1145/3680528.3687625.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 87–103, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0. URL https://doi.org/10.1007/978-3-031-73016-0_6.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294. Curran Associates, Inc., 2022. doi: 10.52202/068431-1833. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/41bd71e7bf7f9fe68f1c936940fd06bd-Paper-Conference.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Alimama Creative Team. Flux.1-turbo-alpha. https://huggingface.co/alimama-creative/FLUX.1-Turbo-Alpha, 2024.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83951–84009. Curran Associates, Inc., 2024. doi: 10.52202/079017-2668. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/98a29475083c502c34949f9baa1aa2ef-Paper-Conference.pdf.

Lei Wang, Yang Cheng, Senmao Li, Ge Wu, Yaxing Wang, and Jian Yang. Wadi: Weight direction-aware distillation for one-step image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5574–5584, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_WaDi_Weight_Direction-aware_Distillation_for_One-step_Image_Synthesis_CVPR_2026_paper.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

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, 2023. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/54dcf25318f9de5a7a01f0a4125c541e-Paper-Conference.pdf.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024b. doi: 10.1109/CVPR52733.2024.00632. URL https://doi.org/10.1109/CVPR52733.2024.00632.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87400–87442, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d9e42b4d7163931f3689d6d6fbaa11d0-Paper-Conference.pdf.

Bowen Zheng and Tianming Yang. Revisiting diffusion models: From generative pre-training to one-step generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78434–78453. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zheng25r.html.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0534abc9e6db91683d82186ef0d68202-Paper-Conference.pdf.

A Response Protocols and Robustness

The opening assay measures what changes within a single call. The endpoint assays identify the teacher component required for response matching beyond the base. We specify both protocols and follow the teacher margin through the trained students.

A.1 Single-network-output spectra

We observe the full SD1.5 UNet noise-prediction output at $t = 500$ on a 1,000-step DDIM grid (Song et al., 2021a). Six independent random directions define the input family. With one fixed measurement seed, 80 points have coordinates sampled uniformly from $[-1, 1]$, at perturbation scales $\varepsilon \in \{0.05, 0.1, 0.2\}$. States, perturbations, UNet forwards, and response differencing use FP32; spectrum calculations use FP64. Appendix D.7 identifies the checkpoints, prompts, and seeds. For eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ of the covariance across the measured response columns, define

$$r_{99} = \min\left\{ r : \frac{\sum_{j=1}^{r} \lambda_j}{\sum_j \lambda_j} \geq 0.99 \right\}, \qquad \sum_j \lambda_j > 0. \tag{A.1}$$

At perturbation scale 0.1, loading approximately 20 MB of LCM-LoRA changes r_{99} from 6 to 26, as shown in Table A.1. Across all three scales, the base and standard LoRA remain at 6, while the official SD1.5 DMD2 baseline used in Table 4 gives 9, 15, and 18, respectively. Table A.2 reports these results alongside three LCM-LoRA instances. These spectra describe finite-scale nonlinear responses within the fixed input family.

Table A.1: **A changed response spectrum within one network call.** SD1.5 at perturbation scale 0.1. Entries give the effective finite-response dimension for the common six-dimensional input family.

Checkpoint	r_{99}
Base	6
Standard LoRA	6
LCM-LoRA distilled	26
DMD2 distilled	15

Table A.2: **Response expansion across distillation routes, instances, and scales.** SD1.5 models share the state, prompt, six input directions, and 80 perturbation points at $t = 500$. Base uses official pretrained weights; standard LoRA (Hu et al., 2022) is the rank-32 standard adapter specified in Appendix D.7. LCM-B is the official LCM-LoRA (Luo et al., 2023b) instance, and LCM-A and LCM-C are reproduced instances. DMD2 uses the official checkpoint (Yin et al., 2024a). The opening value of 26 uses LCM-B at scale 0.1.

Scale	Base	Standard LoRA	LCM-A	LCM-B	LCM-C	DMD2
0.05	6	6	12	14	19	9
0.1	6	6	24	26	28	15
0.2	6	6	30	30	30	18

The energy distribution further characterizes this spectral expansion. Energy beyond the leading six directions accounts for 0.1%, 0.2%, and 0.5% of the total for the base and 2.4%, 5.4%, and 5.9% for LCM-A. At scales 0.1 and 0.2, LCM-A therefore also satisfies $r_{95} > 6$, while the base has $r_{95} \leq 6$, with r_{95} defined as in Equation (A.1) at level 0.95.

We extend the same perturbation design to the full FLUX network at the corresponding mid-schedule state, retaining FP32 inputs and forwards and FP64 spectrum calculations. Under this protocol, Hyper-FLUX (Ren et al., 2024), FLUX-Turbo-Alpha (Team, 2024), FLUX.1-schnell (Labs, 2024),

Table A.3: **Energy is distributed beyond the leading six directions.** Percentages refer to each model's ordered response-covariance spectrum for the base and LCM-A (Luo et al., 2023b).

Scale	Base: first 6	Base: 7–30	LCM-A: first 6	LCM-A: 7–30
0.05	99.9%	0.06%	97.6%	2.29%
0.1	99.8%	0.20%	94.6%	4.79%
0.2	99.5%	0.47%	94.1%	4.91%

and SenseFlow (Ge et al., 2026) weights show effective-response dimensions 3–5 times the base value.

A sampler-only change preserves the single-network output at an identical input while changing the composition of calls. To measure the response changes introduced by this composition, we next compare endpoints.

A.2 Common-probe endpoint responses

At each measurement location, Equation (1) uses a fixed state, condition, measurement seed, and common unit-normalized random probes ξ_i. We use $\delta = 0.05$ and $n \in \{32, 64, 128\}$. The base returns a one-step x_0 estimate from its noise prediction or flow counterpart, the teacher returns the endpoint of a 50-step trajectory, and the student returns its one-step endpoint. SD1.5 uses DDIM and latent endpoints; the other five backbones use pixel endpoints downsampled to 128×128. Each output is flattened into a response column. Finite differences are computed through input-output queries. For SD1.5, input states, perturbations, and network forwards use FP32. For the other five backbones, probe addition is formed in FP32 and the resulting network inputs, network forwards, and VAE decoding use BF16. Returned representations are cast to FP32 before differencing; finite differences, SVD, and projections use FP32. The independent full-FP32 recomputations in Appendix A.7 validate the paired FLUX measurements. For a stochastic map, a response is defined conditional on sampler randomness. Compared configurations share the probes and base projector at each measurement location.

Let $\mathrm{Proj}_{0.01}(M)$ project onto the left singular vectors of M whose singular values exceed $0.01\sigma_1(M)$. The numerical base projector, measured teacher margin, and retained margin projector are

$$\widehat{P}_B = \mathrm{Proj}_{0.01}(Y_B), \quad Z_T = (I - \widehat{P}_B)Y_T, \quad \widehat{P}_N = \mathrm{Proj}_{0.01}(Z_T). \tag{A.2}$$

The counting rule uses the same relative cutoff $\tau = 0.01$, giving

$$q_\tau(M) = \#\{j : \sigma_j(M) > 0.01\sigma_1(M)\}, \qquad q_\tau(0) = 0. \tag{A.3}$$

Thus $q_\tau(Z_T) = \mathrm{rank}(\widehat{P}_N)$, and student occupation uses $q_\tau(\widehat{P}_N Y_S)$. The exact-span objects are P_B, N_T, P_N; Appendix B.2 gives the residual-inclusive numerical identity, teacher truncation tail, and base residual.

A.3 Probe budgets, trajectories, and controls

The target persists across probe budgets. Table A.4 reports all three probe counts. Every backbone has a nonzero margin throughout the scan; 128 probes resolve 69–118 retained directions. The associated $\gamma_\tau = \sigma_{q_\tau(Z_T)}(Z_T)$ is the smallest retained singular value in each representation.

Repeated calls create the target. We sweep $K \in \{1, 2, 5, 10, 25, 50\}$ teacher sampling steps, changing the teacher time grid while holding the remaining settings fixed. At $n = 64$, Table A.5 gives the margin-to-teacher norm ratio $\eta(K) = \|Z_T(K)\|_F / \|Y_T(K)\|_F$. Two steps give sharply different ratios across backbones; by five steps, all six reach 0.950–0.998.

Full-output spectral controls. A separate SD1.5 full-UNet-output response comparison uses $t = 999$, a fixed prompt and initial state, 128 probes, and $\delta = 0.05$, with FP32 states, probes, forwards, and differences and FP64 SVD. Standard adaptation uses the rank-32 standard adapter specified in

Table A.4: **The teacher margin persists as the probe budget grows.** Margin counts use $0.01\sigma_1(Z_T)$. The right-hand group gives the smallest retained singular values at the same budgets.

	Margin count q_τ			Retained γ_τ		
Backbone	$n = 32$	64	128	$n = 32$	64	128
SD1.5 (latent)	32	43	105	182	154	109
FLUX.1-dev	32	64	101	21	18	22
klein-4B	32	56	69	14	14	19
SDXL	32	64	118	21	18	16
SD3.5-medium	32	64	110	7	7	6
SD3.5-Large	32	64	99	19	14	17

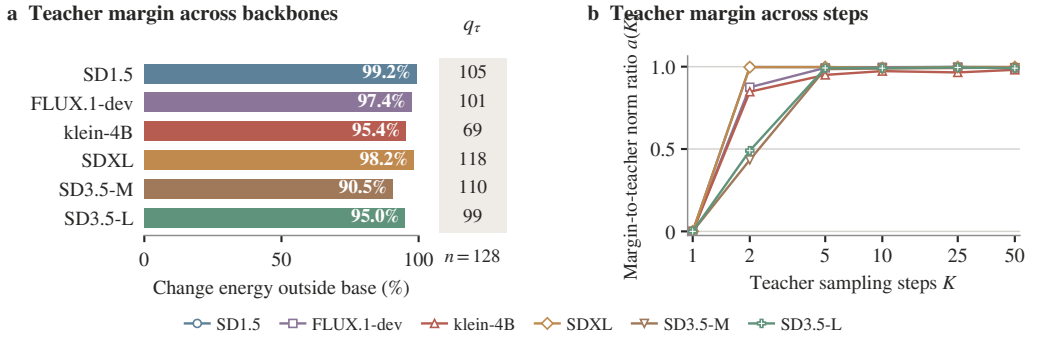


Figure A.1: **Repeated computation adds a substantial target beyond the base.** (a) Outside-base change-energy share ρ_Δ at $n = 64$ and retained margin rank q_τ at $n = 128$; $\tau = 0.01$ is the relative singular-value cutoff on Z_T. (b) Margin-to-teacher norm ratio $\eta(K) = \|Z_T(K)\|_F / \|Y_T(K)\|_F$ at $n = 64$, with K teacher sampling steps. Approximate-zero, base-consistent $K = 1$ readings are plotted at zero.

Table A.5: **Teacher sampling steps and margin-to-teacher norm ratios.** Complete scan over teacher sampling steps underlying Figure A.1b, with $n = 64$.

Backbone	$K = 1$	2	5	10	25	50
SD1.5	≈ 0	0.998	0.997	0.997	0.998	0.997
FLUX.1-dev	≈ 0	0.875	0.993	0.996	0.998	0.991
klein-4B	≈ 0	0.848	0.950	0.973	0.965	0.981
SDXL	≈ 0	0.997	0.998	0.992	0.998	0.998
SD3.5-medium	≈ 0	0.434	0.985	0.989	0.992	0.994
SD3.5-Large	≈ 0	0.490	0.992	0.991	0.995	0.991

Appendix D.7, fused into the UNet; the teacher uses 50 steps. For the descending singular spectrum of the measured full-output response matrix M, define

$$f(M) = \text{median}\{\sigma_{65}(M), \ldots, \sigma_{128}(M)\},$$
$$q_{\text{sig}}(M) = \#\{j : \sigma_j(M) > \max(3f(M), 0.01\sigma_1(M))\}. \tag{A.4}$$

Under this shared tail-based rule, standard LoRA retains one significant spectral direction and the teacher retains 36, as reported in Table A.6. Applying q_τ to the same full-output matrices retains 128 and 103, respectively. The table gives the thresholds and counting boundaries for both rules. The endpoint teacher margin counts after base projection are reported in Table A.4.

Table A.6: **A shared tail-based rule separates the full-output spectra.** The independent SD1.5 full-UNet-output assay uses $t = 999$, 128 probes, and $\delta = 0.05$, comparing standard LoRA (Hu et al., 2022) with the multi-step SD1.5 teacher (Rombach et al., 2022). Counts apply to each complete response matrix. The tail-based threshold is $\max(3f, 0.01\sigma_1)$, with f the trailing-half median; q_τ uses the relative cutoff alone.

Quantity	Standard LoRA	Multi-step teacher
Largest singular value σ_1	81.83	11524.8
Relative threshold $0.01\sigma_1$	0.818	115.2
Tail median f	7.22	251.0
Tail threshold $3f$	21.65	753.0
Combined threshold $\max(3f, 0.01\sigma_1)$	21.65	753.0
Relative-threshold count q_τ	128	103
Smallest σ retained by q_τ	6.64	121.5
Largest σ excluded by q_τ	—	114.3
Tail-based count q_{sig}	**1**	**36**
Smallest σ retained by q_{sig}	81.83	753.4
Largest σ excluded by q_{sig}	9.07	751.2
σ_{105}	7.08	103.4
σ_{106}	7.06	97.3

Students occupy the teacher margin space. Table A.7 reports nonzero projected responses for all six endpoint-assay students. SD1.5, FLUX, and klein use our evaluated one-step checkpoints; SDXL and the two SD3.5 models use released one-step students. The counts locate endpoint responses in the retained teacher margin span. Probe-wise accuracy is measured by the errors of Section 3.3 and angular alignment in Appendix A.4.

Table A.7: **All six endpoint-assay students occupy part of the teacher margin space.** SDXL and SD3.5 entries use released one-step SenseFlow students. Counts $q_\tau(\widehat{P}_N Y_S)$ use the retained teacher margin projector and the 1% spectral cutoff.

Backbone	Reported projection count
SD1.5	26
FLUX.1-dev	59
klein-4B	52
SDXL	61
SD3.5-medium	55
SD3.5-Large	54

Base-span inclusion. A separate pixel-space assay at $n = 128$ reports threshold-count slack $q_\tau([Y_B, Y_T]) - q_\tau(Y_T)$ of 1 for SD1.5 and 128 for SDXL, with SDXL base count 128. Each matrix uses its own relative threshold. The exact inclusion assumption for total-rank growth is stated in Proposition B.1.

Feature-sample checks. An auxiliary feature-sample scan compares the base, SenseFlow, and a trained FLUX checkpoint at $N \in \{16, 32, 64, 128\}$. Their ordering in spread, reference alignment, and concentration is maintained throughout this scan. The training diagnostics in Section 4.2 use the 32-prompt, eight-image cohort specified in Appendix D.3.

A.4 Teacher-margin alignment in the same training students

Each SD1.5 and FLUX cumulative comparison uses four prompts and three initial noises per prompt, yielding 12 state-prompt pairs per backbone. At each pair, we use 64 common unit probes and $\delta = 0.05$, construct $\widehat{P}_B$ from the base responses, and share the probes, teacher responses, and base projector across student configurations. Representations and FP32 differencing, SVD, and projection follow Appendix A.2. FLUX compares the final weights of four samplewise configurations retaining DM and successively adding expansion, alignment, and GAN. SD1.5 uses the official DMD2 checkpoint as a reference alongside the final weights of three from-base samplewise configurations consisting of DM with expansion, then alignment, then GAN. We compute $\mathrm{Align}_{\mathrm{margin}}$ from Equation (8) on the complete base-external residuals at each pair and report the arithmetic mean over all 12 pairs.

Mean response alignment is 0.0026, 0.0043, 0.0059, and 0.0062 on SD1.5 and 0.0037, 0.0061, 0.0088, and 0.0089 on FLUX. Appendix F.1 pairs every mean with the generation scores and DINO diagnostics of that student.

A Gaussian reference characterizes the scale of random angular alignment using 30 independent random matrices per backbone at a fixed state-prompt pair, with the same shape as the teacher responses, base projection, and cosine calculation. Table A.8 reports the random cosines' signed mean and standard deviation alongside the complete samplewise student's mean response alignment over 12 state-prompt pairs.

Table A.8: **Gaussian cosine scale and student response alignment.** The Gaussian reference uses 30 random matrices at a fixed state-prompt pair with 64 probes per backbone. Student entries average 12 state-prompt pairs for the complete samplewise recipe with DM, expansion, alignment, and GAN.

Backbone	Random mean	Random std.	Student mean
SD1.5	≈ 0	0.00094	0.0062
FLUX.1-dev	-0.0002	0.00057	0.0089

A.5 Direction, magnitude, and error in paired FLUX responses

Students and shared measurements. We compare pure distribution matching (DM), the complete samplewise recipe (SW), and the final covariance recipe (COV) as three trained FLUX students. SW and COV both retain DM, expansion, alignment, and GAN, differing in the two feature controls. Their Patch FID-T / GenEval scores are 42.83 / 0.5316 for DM, 17.16 / 0.6551 for SW, and 17.04 / 0.6598 for COV. The unmodified pretrained one-step map B defines the base response span; T is the 50-step teacher. All three students share B, T, the 64 unit probes, and the projectors at each of the four prompts and three initial noises per prompt. We label these 12 state-prompt pairs P1-N1 through P4-N3. The baseline setting is $\delta = 0.05$, $\tau = 0.01$, with RGB endpoints at 128×128, giving $p = 49152$.

For this paired FLUX analysis and its endpoint checks below, transformer forward passes and VAE decoding use BF16. Decoded floating-point images are cast to FP32, downsampled to 128×128, and differenced in FP32. SVD, projections, and component-error calculations use FP64. Differencing acts on these floating-point outputs before integer image quantization. The earlier six-backbone and cumulative assays use the precision specified in Appendices A.2 and A.4.

Table A.9 summarizes the paired responses of the three students.

Table A.9: **Paired FLUX responses from three trained students.** Mean (descriptive standard deviation) over the same four prompts and three initial noises per prompt. SW and COV are the complete samplewise and covariance recipes; DM is the trained distribution-matching student. Alignment a, norm ratio ρ, and errors are computed per pair before averaging. Appendix A.5 defines all quantities.

Student	a	ρ	E_{out}	$\mathcal{L}_{\text{resp}}$
DM	0.0037 (0.0011)	0.5500 (0.0555)	1.3015 (0.0618)	4779.6 (158.3)
SW	0.0089 (0.0026)	0.8500 (0.0240)	1.7079 (0.0392)	5162.3 (105.9)
COV	0.0093 (0.0027)	0.8700 (0.0283)	1.7415 (0.0489)	5121.1 (115.3)

Per-pair quantities and aggregation. At each pair, let $Z_T = (I - \widehat{P}_B)Y_T$, $A_S = (I - \widehat{P}_B)Y_S$, and $D_S = Y_S - Y_T$. The numerical projectors retain singular directions above τ times the respective leading singular value of Y_B and Z_T, and $\widehat{P}_\perp = I - \widehat{P}_B - \widehat{P}_N$. Define

$$a = \frac{\langle A_S, Z_T\rangle_F}{\|A_S\|_F\|Z_T\|_F}, \qquad \rho = \frac{\|A_S\|_F}{\|Z_T\|_F}, \qquad E_{\text{out}} = \frac{\|A_S - Z_T\|_F^2}{\|Z_T\|_F^2},$$
$$e_B = \frac{\|\widehat{P}_B D_S\|_F^2}{2n}, \qquad e_N = \frac{\|\widehat{P}_N D_S\|_F^2}{2n}, \qquad e_R = \frac{\|\widehat{P}_\perp D_S\|_F^2}{2n}. \tag{A.5}$$

Here a is response alignment from Equation (8). With $H = \|Z_T\|_F^2/(2n)$, the per-pair identities are

$$E_{\text{out}} = \rho^2 + 1 - 2\rho a, \qquad \mathcal{L}_{\text{resp}} = e_B + e_N + e_R, \qquad e_N + e_R = HE_{\text{out}}. \tag{A.6}$$

The residual e_R includes the teacher's discarded tail through the student–teacher difference. Appendix B.5 relates it to the student-only residual energy in the numerical fidelity term.

We compute all quantities per state-prompt pair and then average the 12 pairs with equal weight. Descriptive standard deviations use denominator 12. Prompt summaries average the three initial noises within each prompt. The 64 probes form each pair's response matrices. Table A.10 gives all 36 student rows; Table A.11 gives their shared base and teacher quantities.

Table A.10: **All 12 paired FLUX measurements.** Four prompts (P1–P4), three initial noises (N1–N3), and three trained students. Each pair shares the base, 50-step teacher, 64 unit probes, and retained projectors; $\delta = 0.05$, $\tau = 0.01$. SW and COV are the complete samplewise and covariance recipes. All columns use their original units; independently rounded error components can differ from the total by 0.1.

Pair	Student	a	ρ	E_{out}	$\mathcal{L}_{\text{resp}}$	e_B	e_N	e_R
P1-N1	DM	0.0022	0.6159	1.3766	5017.1	1663.2	3097.3	256.6
P1-N1	SW	0.0053	0.8122	1.6510	5106.9	1084.4	3718.9	303.6
P1-N1	COV	0.0055	0.8945	1.7902	5184.3	822.6	3695.6	666.1
P1-N2	DM	0.0024	0.4991	1.2467	4646.5	1615.8	2914.2	116.5
P1-N2	SW	0.0058	0.8468	1.7072	5110.7	960.5	3377.5	772.7
P1-N2	COV	0.0061	0.8403	1.6958	4957.3	834.7	3498.6	624.0
P1-N3	DM	0.0021	0.5536	1.3041	4813.9	1638.4	3025.8	149.7
P1-N3	SW	0.0050	0.8219	1.6673	4985.2	925.4	3562.2	497.6
P1-N3	COV	0.0052	0.8462	1.7073	5105.8	948.5	3835.2	322.1
P2-N1	DM	0.0031	0.4690	1.2170	4640.4	1686.4	2697.0	257.0
P2-N1	SW	0.0074	0.8662	1.7374	5187.1	970.0	3813.3	403.8
P2-N1	COV	0.0077	0.8270	1.6711	4957.8	901.6	3575.1	481.1
P2-N2	DM	0.0033	0.4911	1.2379	4647.9	1655.1	2723.5	269.3
P2-N2	SW	0.0079	0.8429	1.6971	5075.5	972.5	3665.4	437.6
P2-N2	COV	0.0082	0.9049	1.8040	5208.9	847.5	4023.6	337.8
P2-N3	DM	0.0034	0.5968	1.3521	4923.0	1635.6	2951.9	335.5
P2-N3	SW	0.0082	0.8825	1.7643	5335.6	1046.1	3645.0	644.5
P2-N3	COV	0.0086	0.8928	1.7817	5240.6	908.7	3724.8	607.1
P3-N1	DM	0.0042	0.6091	1.3659	4968.9	1637.0	3108.4	223.5
P3-N1	SW	0.0101	0.8692	1.7380	5264.6	1025.2	3845.7	393.7
P3-N1	COV	0.0105	0.8967	1.7853	5200.7	845.9	3841.8	513.0
P3-N2	DM	0.0041	0.6537	1.4219	5040.2	1575.7	2973.7	490.8
P3-N2	SW	0.0099	0.8304	1.6731	5147.1	1070.5	3355.5	721.1
P3-N2	COV	0.0103	0.9043	1.7991	5327.4	943.8	3979.0	404.6
P3-N3	DM	0.0044	0.5253	1.2713	4589.4	1510.2	2800.5	278.7
P3-N3	SW	0.0106	0.8824	1.7600	5309.2	1046.6	3514.0	748.6
P3-N3	COV	0.0111	0.8722	1.7413	5116.3	898.8	3866.7	350.8
P4-N1	DM	0.0049	0.5343	1.2803	4709.8	1607.6	2828.1	274.1
P4-N1	SW	0.0118	0.8797	1.7531	5242.9	995.0	3526.6	721.3
P4-N1	COV	0.0123	0.8831	1.7582	5147.4	887.1	3612.0	648.3
P4-N2	DM	0.0053	0.5541	1.3011	4718.5	1554.3	2934.0	230.2
P4-N2	SW	0.0128	0.8379	1.6807	5163.8	1076.6	3583.9	503.3
P4-N2	COV	0.0134	0.8402	1.6835	4966.9	872.8	3730.1	364.0
P4-N3	DM	0.0050	0.4981	1.2431	4639.7	1600.8	2885.5	153.4
P4-N3	SW	0.0121	0.8281	1.6656	5018.9	947.1	3473.9	597.9
P4-N3	COV	0.0126	0.8377	1.6806	5040.3	931.9	3335.3	773.1

Table A.11: **Shared base and teacher quantities for the paired comparison.** $H = \|Z_T\|_F^2/(2n)$, $\mathcal{L}_{\text{resp}}(B) = \|Y_B - Y_T\|_F^2/(2n)$, and ρ_Δ is the outside-base share of teacher–base change energy. B is the unmodified pretrained one-step map. Each row applies to DM, SW, and COV at that pair.

Pair	H	$\mathcal{L}_{\text{resp}}(B)$	ρ_Δ
P1-N1	2436.4	2500.1	0.9745
P1-N2	2431.0	2491.4	0.9758
P1-N3	2435.0	2502.3	0.9731
P2-N1	2427.2	2498.1	0.9716
P2-N2	2417.6	2482.4	0.9739
P2-N3	2431.3	2505.4	0.9704
P3-N1	2439.3	2499.8	0.9758
P3-N2	2436.5	2500.4	0.9744
P3-N3	2422.0	2483.3	0.9753
P4-N1	2423.1	2491.7	0.9725
P4-N2	2431.9	2489.6	0.9768
P4-N3	2444.6	2504.1	0.9762

Prompt-level direction gains and complete-error changes. SW and COV increase alignment over DM at all 12 pairs and in all four prompt groups (Table A.12). The paired alignment increments have means and descriptive standard deviations 0.0052 ± 0.0015 and 0.0056 ± 0.0016, respectively. Their base-external norm ratios are closer to one at every pair. Both E_{out} and $\mathcal{L}_{\text{resp}}$ also exceed DM at every baseline pair. The direction gain and error increase occur together in these same students.

Table A.13 resolves the complete-error changes. SW and COV reduce mean base-component error by 605.0 and 728.0. Their mean retained-margin errors rise by 678.5 and 814.8, and their mean residual errors rise by 309.2 and 254.7. The resulting complete-error increases are 382.7 and 341.5. The squared-amplitude and cross-term averages in Table A.14 explain the base-external error. The increase in $\overline{\rho^2}$ exceeds the increase in $\overline{2\rho a}$. Each product is formed per pair before averaging. At these small absolute cosines, bringing the response norm closer to the teacher's raises squared mismatch even as angular agreement improves.

Table A.12: **Direction gains recur in every prompt group.** Within each prompt, increments relative to DM average its three initial noises. Both SW and COV have positive alignment increments at all 12 individual pairs. All 12 complete errors also exceed DM for each student at the baseline setting.

Prompt	SW: Δa	COV: Δa	SW: $\Delta\mathcal{L}_{\text{resp}}$	COV: $\Delta\mathcal{L}_{\text{resp}}$
P1	0.00313	0.00337	+241.8	+256.6
P2	0.00457	0.00490	+462.3	+398.7
P3	0.00597	0.00640	+374.1	+348.6
P4	0.00717	0.00770	+452.5	+362.2

Table A.13: **Where complete response error changes.** Means over the 12 FLUX pairs and paired changes from trained DM. Each component measures student–teacher error in the indicated orthogonal space. Each column is rounded independently.

Student / change	$\mathcal{L}_{\text{resp}}$	e_B	e_N	e_R
DM	4779.6	1615.0	2911.7	252.9
SW	5162.3	1010.0	3590.2	562.1
COV	5121.1	887.0	3726.5	507.7
SW minus DM	+382.7	-605.0	+678.5	+309.2
COV minus DM	+341.5	-728.0	+814.8	+254.7

Table A.14: **Amplitude growth and the alignment cross term.** Each product is computed at each pair and then averaged over the 12 FLUX pairs. The last column reconstructs mean normalized base-external error from the displayed direction and norm-ratio readings.

Student	$\overline{\rho^2}$	$\overline{2\rho a}$	$1+\overline{\rho^2}-\overline{2\rho a}$
DM	0.3055799	0.0040690	1.3015110
SW	0.7230802	0.0151848	1.7078955
COV	0.7576917	0.0161666	1.7415251

Individual SD1.5 samplewise alignments. Table A.15 gives the 12 individual alignment readings of the complete SD1.5 samplewise student from Appendix A.4. Its mean and descriptive standard deviation are 0.0062 ± 0.0025. The three-student direction, magnitude, and error comparison above uses FLUX; this SD1.5 table records the distribution across pairs for the complete samplewise student.

Table A.15: **Individual response alignments of the complete SD1.5 samplewise student.** Four prompts with three initial noises each under Appendix A.4. The overall mean and descriptive standard deviation are 0.0062 ± 0.0025.

Prompt	N1	N2	N3	Prompt mean
P1	0.0028	0.0026	0.0032	0.00287
P2	0.0048	0.0053	0.0052	0.00510
P3	0.0071	0.0076	0.0071	0.00727
P4	0.0098	0.0094	0.0094	0.00953

A.6 FLUX ENDPOINT CHECKS ACROSS SCALES AND THRESHOLDS

At the fixed pairs P1-N1 and P4-N1, we check five (δ, τ) settings comprising the baseline $(0.05, 0.01)$, adjacent scales $(0.025, 0.01)$ and $(0.10, 0.01)$, and adjacent thresholds $(0.05, 0.005)$ and $(0.05, 0.02)$. Students share the base and teacher responses and probes within each setting. Computation uses the BF16 forward / FP32 difference / FP64 projection protocol above. Table A.16 gives all 30 student readings, and Table A.17 records the common quantities.

SW and COV retain positive alignment increments over DM at both pairs under every listed setting, with norm ratios closer to one. Table A.18 shows the scale-wise increments. At P1-N1 with $\delta = 0.025$, SW's complete error is 5002.8 versus DM's 5011.1, a decrease of 8.3. The other listed SW and COV complete errors exceed DM. Thus direction gains persist across these adjacent scales while the local complete-error ordering can change.

The threshold-only rows reuse each pair's baseline response matrices. For $\tau = 0.005, 0.01, 0.02$, the base rank remains 64 and the margin rank is 62, 59, and 58, respectively. The same base projector yields identical a, ρ, E_{out}, e_B, H, and ρ_Δ. Changing the retained margin projector redistributes e_N and e_R, with their sum and complete response error fixed. These rows describe threshold dependence of the same responses; the paired statistics above use the 12 baseline pairs.

Table A.16: **Endpoint response checks at adjacent scales and thresholds.** All 30 readings at two fixed FLUX pairs. Baseline rows reproduce Table A.10. Threshold-only rows reuse the baseline response matrices. SW and COV increase alignment over DM in every listed setting. At P1-N1 and $\delta = 0.025$, SW has complete error 5002.8 versus 5011.1 for DM.

Pair	δ	τ	Student	a	ρ	E_{out}	$\mathcal{L}_{\text{resp}}$	e_B	e_N	e_R
P1-N1	0.05	0.01	DM	0.0022	0.6159	1.3766	5017.1	1663.2	3097.3	256.6
P1-N1	0.05	0.01	SW	0.0053	0.8122	1.6510	5106.9	1084.4	3718.9	303.6
P1-N1	0.05	0.01	COV	0.0055	0.8945	1.7902	5184.3	822.6	3695.6	666.1
P1-N1	0.025	0.01	DM	0.0022	0.5924	1.3483	5011.1	1724.2	3015.4	271.5
P1-N1	0.025	0.01	SW	0.0054	0.7851	1.6078	5002.8	1083.2	3447.2	472.4
P1-N1	0.025	0.01	COV	0.0057	0.8996	1.7991	5206.7	820.9	3909.0	476.8
P1-N1	0.1	0.01	DM	0.0022	0.5923	1.3483	4975.6	1706.6	2885.3	383.7
P1-N1	0.1	0.01	SW	0.0051	0.8157	1.6570	5068.0	1050.4	3418.7	598.9
P1-N1	0.1	0.01	COV	0.0054	0.9059	1.8109	5228.3	837.5	3595.6	795.2
P1-N1	0.05	0.005	DM	0.0022	0.6159	1.3766	5017.1	1663.2	3097.3	256.6
P1-N1	0.05	0.005	SW	0.0053	0.8122	1.6510	5106.9	1084.4	3718.9	303.6
P1-N1	0.05	0.005	COV	0.0055	0.8945	1.7902	5184.3	822.6	3695.6	666.1
P1-N1	0.05	0.02	DM	0.0022	0.6159	1.3766	5017.1	1663.2	3086.8	267.1
P1-N1	0.05	0.02	SW	0.0053	0.8122	1.6510	5106.9	1084.4	3717.5	305.0
P1-N1	0.05	0.02	COV	0.0055	0.8945	1.7902	5184.3	822.6	3672.5	689.2
P4-N1	0.05	0.01	DM	0.0049	0.5343	1.2803	4709.8	1607.6	2828.1	274.1
P4-N1	0.05	0.01	SW	0.0118	0.8797	1.7531	5242.9	995.0	3526.6	721.3
P4-N1	0.05	0.01	COV	0.0123	0.8831	1.7582	5147.4	887.1	3612.0	648.3
P4-N1	0.025	0.01	DM	0.0048	0.5145	1.2598	4646.5	1614.7	2850.6	181.2
P4-N1	0.025	0.01	SW	0.0121	0.9014	1.7907	5289.9	980.4	3414.4	895.1
P4-N1	0.025	0.01	COV	0.0121	0.8842	1.7604	5106.3	869.8	3384.5	852.0
P4-N1	0.1	0.01	DM	0.0049	0.5554	1.3030	4766.5	1658.0	2810.3	298.2
P4-N1	0.1	0.01	SW	0.0122	0.9074	1.8013	5293.8	996.5	3792.5	504.8
P4-N1	0.1	0.01	COV	0.0126	0.8945	1.7777	5119.1	878.1	3883.3	357.7
P4-N1	0.05	0.005	DM	0.0049	0.5343	1.2803	4709.8	1607.6	2828.1	274.1
P4-N1	0.05	0.005	SW	0.0118	0.8797	1.7531	5242.9	995.0	3526.6	721.3
P4-N1	0.05	0.005	COV	0.0123	0.8831	1.7582	5147.4	887.1	3612.0	648.3
P4-N1	0.05	0.02	DM	0.0049	0.5343	1.2803	4709.8	1607.6	2818.8	283.4
P4-N1	0.05	0.02	SW	0.0118	0.8797	1.7531	5242.9	995.0	3526.6	721.3
P4-N1	0.05	0.02	COV	0.0123	0.8831	1.7582	5147.4	887.1	3579.1	681.2

Table A.17: **Shared quantities for the endpoint checks.** r_B and q are the retained base and margin ranks. The RGB representation has $p = 49152$ coordinates. Changing τ at fixed $\delta = 0.05$ retains the same base span and changes the retained margin rank.

Pair	δ	τ	p	r_B	q	ρ_Δ	H	$\mathcal{L}_{\text{resp}}(B)$
P1-N1	0.05	0.01	49152	64	59	0.9745	2436.4	2500.1
P1-N1	0.025	0.01	49152	64	59	0.9727	2437.8	2506.2
P1-N1	0.1	0.01	49152	64	59	0.9747	2424.6	2487.5
P1-N1	0.05	0.005	49152	64	62	0.9745	2436.4	2500.1
P1-N1	0.05	0.02	49152	64	58	0.9745	2436.4	2500.1
P4-N1	0.05	0.01	49152	64	59	0.9725	2423.1	2491.7
P4-N1	0.025	0.01	49152	64	59	0.9773	2406.6	2462.4
P4-N1	0.1	0.01	49152	64	59	0.9672	2385.7	2466.6
P4-N1	0.05	0.005	49152	64	62	0.9725	2423.1	2491.7
P4-N1	0.05	0.02	49152	64	58	0.9725	2423.1	2491.7

Table A.18: **Paired changes under the perturbation-scale scan.** SW and COV increments are relative to DM at each setting, with $\tau = 0.01$. Direction gains persist at both fixed pairs while the local complete-error ordering changes for SW at P1-N1 and the smallest scale.

Pair	δ	SW: Δa	COV: Δa	SW: $\Delta\mathcal{L}_{\text{resp}}$	COV: $\Delta\mathcal{L}_{\text{resp}}$
P1-N1	0.05	0.0031	0.0033	+89.8	+167.2
P1-N1	0.025	0.0032	0.0035	-8.3	+195.6
P1-N1	0.1	0.0029	0.0032	+92.4	+252.7
P4-N1	0.05	0.0069	0.0074	+533.1	+437.6
P4-N1	0.025	0.0073	0.0073	+643.4	+459.8
P4-N1	0.1	0.0073	0.0077	+527.3	+352.6

Table A.19: **Tail energy and retained ranks from the same response records.** $K_{\text{tail}} = \sum_{j>q} \sigma_j^2/(2n)$; the last column is the sharp cutoff bound $(n-q)\tau^2\sigma_1^2/(2n)$. Each threshold reuses the identical base and teacher matrices at $n = 64$, $\delta = 0.05$.

Pair	τ	q	σ_1	K_{tail}	Cutoff bound
P1-N1	0.01	59	126.34797	0.02073431	0.06235864
P1-N1	0.005	62	126.34797	0.00140306	0.00623586
P1-N1	0.02	58	126.34797	0.04456507	0.29932145
P4-N1	0.01	59	125.77752	0.02052432	0.06179681
P4-N1	0.005	62	125.77752	0.00134727	0.00617968
P4-N1	0.02	58	125.77752	0.04297597	0.29662471

A.7 Forward precision and probe correspondence

At P1-N1 and P4-N1, we recompute the complete base, teacher, and three student response matrices at $\delta \in \{0.025, 0.05, 0.10\}$. The FP32 pipeline loads model and decoder weights into FP32, retains states, perturbations, text embeddings, transformer operations, and VAE decoding in FP32, disables autocast and TF32, and uses the same initial noises and probe tensors as the BF16 pipeline. Both pipelines perform floating-point RGB resizing and differencing in FP32 and SVD in FP64. The base and margin projectors are recomputed separately for each pipeline at $\tau = 0.01$.

Tables A.20–A.22 report response norms, paired alignments, and principal spectra. Across the six pair–scale settings, response-norm changes are below 0.9%; retained base and margin ranks stay at 64 and 59. The largest absolute alignment change is 0.0002. FP32 SW–DM increments range from 0.0029 to 0.0072, and COV–DM increments from 0.0032 to 0.0076. The direction gains therefore persist with full-precision perturbations, network computation, and decoding.

Table A.20: **Forward-precision comparison of base and teacher norms.** Both pairs are recomputed at each scale with either BF16 or full FP32 input, network, and VAE computation. Norms are Frobenius norms of complete response matrices; Z_T uses the base projector recomputed in the same precision. Both pipelines use FP32 differences and FP64 decompositions.

Pair	δ	Forward	$\|Y_B\|_F$	$\|Y_T\|_F$	$\|Z_T\|_F$
P1-N1	0.05	BF16	61.14	562.38	558.44
P1-N1	0.05	FP32	60.95	560.70	556.77
P1-N1	0.025	BF16	63.90	562.77	558.60
P1-N1	0.025	FP32	63.52	559.39	555.25
P1-N1	0.1	BF16	55.89	561.50	557.09
P1-N1	0.1	FP32	55.83	560.93	556.53
P4-N1	0.05	BF16	59.97	561.55	556.92
P4-N1	0.05	FP32	59.79	559.87	555.25
P4-N1	0.025	BF16	41.26	559.90	555.02
P4-N1	0.025	FP32	41.01	556.54	551.69
P4-N1	0.1	BF16	68.26	557.73	552.60
P4-N1	0.1	FP32	68.19	557.17	552.05

Table A.21: **Student responses under BF16 and full FP32 computation.** $A_S = (I - \widehat{P}_B)Y_S$. The paired SW–DM and COV–DM alignment increments remain positive at both pairs and all three scales.

Pair	δ	Student	$\|A_S\|_{16}$	$\|A_S\|_{32}$	a_{16}	a_{32}	ρ_{16}	ρ_{32}
P1-N1	0.05	DM	343.93	342.64	0.0022	0.0022	0.6159	0.6154
P1-N1	0.05	SW	453.55	451.87	0.0053	0.0052	0.8122	0.8116
P1-N1	0.05	COV	499.51	497.64	0.0055	0.0055	0.8945	0.8938
P1-N1	0.025	DM	330.90	328.27	0.0022	0.0023	0.5924	0.5912
P1-N1	0.025	SW	438.54	435.04	0.0054	0.0053	0.7851	0.7835
P1-N1	0.025	COV	502.54	498.51	0.0057	0.0056	0.8996	0.8978
P1-N1	0.1	DM	329.99	329.63	0.0022	0.0022	0.5923	0.5923
P1-N1	0.1	SW	454.41	453.96	0.0051	0.0051	0.8157	0.8157
P1-N1	0.1	COV	504.69	504.16	0.0054	0.0054	0.9059	0.9059
P4-N1	0.05	DM	297.57	296.45	0.0049	0.0048	0.5343	0.5339
P4-N1	0.05	SW	489.91	488.06	0.0118	0.0117	0.8797	0.8790
P4-N1	0.05	COV	491.83	489.95	0.0123	0.0122	0.8831	0.8824
P4-N1	0.025	DM	285.56	283.29	0.0048	0.0047	0.5145	0.5135
P4-N1	0.025	SW	500.29	496.30	0.0121	0.0119	0.9014	0.8996
P4-N1	0.025	COV	490.73	486.81	0.0121	0.0120	0.8842	0.8824
P4-N1	0.1	DM	306.89	306.55	0.0049	0.0049	0.5554	0.5553
P4-N1	0.1	SW	501.44	500.93	0.0122	0.0121	0.9074	0.9074
P4-N1	0.1	COV	494.33	493.81	0.0126	0.0125	0.8945	0.8945

Table A.22: **Principal response spectra under each forward precision.** $s_j = \sigma_j/\sigma_1$ and $q = \#\{j : s_j > 0.01\}$. Base refers to Y_B; margin refers to Z_T.

Pair	δ	Forward	Matrix	σ_1	s_8	s_{32}	s_{59}	s_{60}	q
P1-N1	0.05	BF16	Base	15.56	0.7937	0.3595	0.1475	0.1427	64
P1-N1	0.05	BF16	Margin	126.35	0.8758	0.4572	0.0138	0.0090	59
P1-N1	0.05	FP32	Base	15.57	0.7928	0.3585	0.1473	0.1419	64
P1-N1	0.05	FP32	Margin	126.16	0.8752	0.4570	0.0138	0.0090	59
P1-N1	0.025	BF16	Base	16.26	0.7663	0.3523	0.1467	0.1413	64
P1-N1	0.025	BF16	Margin	128.73	0.8563	0.4410	0.0133	0.0089	59
P1-N1	0.025	FP32	Base	16.19	0.7659	0.3526	0.1466	0.1411	64
P1-N1	0.025	FP32	Margin	128.52	0.8542	0.4402	0.0132	0.0088	59
P1-N1	0.1	BF16	Base	14.48	0.7701	0.3675	0.1476	0.1390	64
P1-N1	0.1	BF16	Margin	125.17	0.8747	0.4558	0.0144	0.0089	59
P1-N1	0.1	FP32	Base	14.49	0.7694	0.3666	0.1482	0.1383	64
P1-N1	0.1	FP32	Margin	125.19	0.8750	0.4568	0.0144	0.0089	59
P4-N1	0.05	BF16	Base	15.37	0.7838	0.3494	0.1462	0.1408	64
P4-N1	0.05	BF16	Margin	125.78	0.8893	0.4738	0.0135	0.0090	59
P4-N1	0.05	FP32	Base	15.36	0.7847	0.3486	0.1461	0.1404	64
P4-N1	0.05	FP32	Margin	125.58	0.8878	0.4736	0.0135	0.0090	59
P4-N1	0.025	BF16	Base	10.33	0.8012	0.3663	0.1511	0.1437	64
P4-N1	0.025	BF16	Margin	124.55	0.8542	0.4642	0.0142	0.0093	59
P4-N1	0.025	FP32	Base	10.30	0.8015	0.3646	0.1510	0.1428	64
P4-N1	0.025	FP32	Margin	123.98	0.8524	0.4641	0.0142	0.0093	59
P4-N1	0.1	BF16	Base	17.13	0.7918	0.3719	0.1461	0.1408	64
P4-N1	0.1	BF16	Margin	127.75	0.8339	0.4498	0.0139	0.0086	59
P4-N1	0.1	FP32	Base	17.18	0.7912	0.3698	0.1459	0.1399	64
P4-N1	0.1	FP32	Margin	128.24	0.8305	0.4480	0.0138	0.0086	59

For a correspondence control, let Π be a permutation matrix with no fixed columns. We replace each measured student response by $Y_S\Pi$ while fixing Y_T and the base projector. Since $A_S\Pi(A_S\Pi)^\top = A_S A_S^\top$ and $\Pi\mathbf{1} = \mathbf{1}$, this operation preserves response norms, singular spectra, column means, and the shared reference-output subtraction. It changes the association between a student's response and the teacher response to the same probe. One common permutation is applied to all students in a comparison, retaining their cross-configuration column relationship.

For each FP32 pair and scale, we draw 999 derangements with seed 7001 and compute the SW–DM and COV–DM alignment increments. Table A.23 reports the null distributions and the corresponding-probe increments. Every observed increment exceeds all 999 permuted increments, giving the plus-one Monte Carlo tail fraction $1/1000$. This control identifies probe correspondence as the source of the measured direction improvement at these fixed pairs, while preserving the spectra and shared offsets that an ambient Gaussian reference does not retain.

Table A.23: **Correspondence-breaking control with measured-response permutations.** FP32 responses, 999 probe-column derangements per pair and scale, and one common permutation across all students. Each observed increment compares corresponding probes; the null compares the permuted student responses with the fixed teacher. The interval gives the 2.5th and 97.5th percentiles of the null distribution; p is the one-sided Monte Carlo tail fraction with the plus-one correction.

Pair	δ	Contrast	Observed	Null mean	Null SD	Null 95% interval	p
P1-N1	0.025	SW–DM	0.0030	0.000184	0.000763	[-0.001287, 0.001733]	0.001
P1-N1	0.025	COV–DM	0.0033	0.000181	0.000792	[-0.001412, 0.001740]	0.001
P1-N1	0.05	SW–DM	0.0030	0.000098	0.000812	[-0.001587, 0.001618]	0.001
P1-N1	0.05	COV–DM	0.0033	0.000073	0.000785	[-0.001483, 0.001657]	0.001
P1-N1	0.1	SW–DM	0.0029	-0.000093	0.000679	[-0.001444, 0.001217]	0.001
P1-N1	0.1	COV–DM	0.0032	-0.000078	0.000704	[-0.001446, 0.001315]	0.001
P4-N1	0.025	SW–DM	0.0072	-0.000177	0.000721	[-0.001541, 0.001233]	0.001
P4-N1	0.025	COV–DM	0.0073	-0.000069	0.000714	[-0.001475, 0.001344]	0.001
P4-N1	0.05	SW–DM	0.0069	-0.000047	0.000721	[-0.001460, 0.001336]	0.001
P4-N1	0.05	COV–DM	0.0074	0.000129	0.000735	[-0.001323, 0.001500]	0.001
P4-N1	0.1	SW–DM	0.0072	0.000111	0.000748	[-0.001415, 0.001572]	0.001
P4-N1	0.1	COV–DM	0.0076	0.000031	0.000796	[-0.001581, 0.001459]	0.001

A.8 Joint representation of spectra and response accounting

We represent responses in a separate coordinate system for each state and prompt pair, perturbation scale, and precision setting. At a setting with m maps, concatenate their endpoint-derived responses as $A = [Y_B, Y_T, Y_{S_1}, \ldots, Y_{S_{m-2}}] \in \mathbb{R}^{49152 \times 64m}$. A thin QR factorization $A = QR$ supplies compact matrices $\widetilde{Y}_G = Q^\top Y_G$ from the corresponding 64-column blocks of R. The compact dimension is set by the joint numerical rank of A, using relative tolerance 10^{-10}. This storage criterion is separate from the 0.01 cutoff defining the base and margin projectors. Eight maps may require 512 coordinates even though each has only 64 columns. The eight-map settings here have joint rank 512; the six-map replica settings have rank 384, and five-map scale checks have rank 320.

Endpoint reconstruction example. At P1-N1 and $\delta = 0.05$, the two precision pipelines yield 16 endpoint matrices of shape 49152×65. The first column is the unperturbed floating RGB endpoint, and the remaining columns are the endpoints at the 64 shared perturbed inputs. They are the channel-first 128×128 RGB values after resizing and before integer quantization. Subtract column zero and divide by δ in FP32, then cast to FP64 for QR and SVD. Both precisions include B, T, DM, SW, and all four factorial cells, with COV shared by the paired and factorial analyses.

We determine the joint rank and compare the original responses with their compact coordinates. Table A.24 reports $\|A - QR\|_F / \|A\|_F$ and $\|A^\top A - R^\top R\|_F / \|A^\top A\|_F$. The compact representation retains the full response span in this example. Since $Q^\top Q = I$, left projection in compact coordinates, singular values, and all Frobenius inner products agree with their full-output counterparts to the reported errors.

Table A.24: **Endpoint-to-coordinate preservation at P1-N1.** Each pipeline concatenates eight matrices with 64 common probes in 49,152 output coordinates. The joint rank uses relative cutoff 10^{-10}; response analysis separately uses $\tau = 0.01$. Errors compare the full endpoint-derived responses with their QR representation.

Precision	Joint rank	σ_{512}/σ_1	Projection residual	Gram error
BF16	512	1.58e-05	8.15e-16	1.02e-15
FP32	512	1.59e-05	8.03e-16	9.29e-16

Reuse and normalization conventions. A checkpoint's response matrix is shared wherever that checkpoint recurs. The both-controls cell is COV, and threshold sweeps reuse the same responses. At a fixed input and precision, base and teacher maps also remain common across training seeds.

Distinct precision or perturbation-scale records have their own endpoint differences and spectra. Singular values are divided by their leading value only for display in Table A.22; the stored matrices and error calculations keep their original scale. Factorial contrasts are differences between individual checkpoint records, with their pair-dependent variation retained.

We construct both projectors with the strict relative cutoff and compute every component from the same difference between student and teacher responses. For each training seed and prompt cohort, we calculate $a, \rho, e_B, e_N, e_R, U_S, C_S, H, K_{\mathrm{tail}}, q$ and the sharp tail bound. Probe permutation summaries use these same matrices with one shared column permutation across students. We verify the discarded spectrum identity, the inner product bound, the component sums, and the identity relating angle and amplitude.

B Proofs and Response Constructions

The teacher margin specifies a required component of exact response matching on common probes. We develop its minimality, exact error decomposition, spectral accuracy cost, attainability, and response constructions.

B.1 The base-confined floor and canonical decomposition

Fix $Y_B, Y_T \in \mathbb{R}^{p\times n}$, with $n \geq 1$, and use the exact projectors of Section 3. For any $u = Y_B a$ and $v = (I - P_B)Y_T b$, $u^\top v = 0$. Hence $\mathcal{U}_N \perp \mathcal{U}_B$, $P_B P_N = 0$, and $P_\perp = I - P_B - P_N$ is the orthogonal projector onto their joint orthogonal complement. Each teacher column decomposes as its base projection plus its margin, giving

$$Y_T = P_B Y_T + N_T, \qquad P_N Y_T = N_T, \qquad P_\perp Y_T = 0. \tag{B.1}$$

The margin space is the unique smallest base-orthogonal space needed to contain the teacher together with the base. Indeed, if $V \perp \mathcal{U}_B$ and $\mathrm{col}(Y_T) \subseteq \mathcal{U}_B \oplus V$, projecting this inclusion with $I - P_B$ gives $\mathcal{U}_N \subseteq V$. Conversely, $V = \mathcal{U}_N$ satisfies the inclusion by Equation (B.1). This minimality fixes the teacher margin and its orthogonal projector.

For a base-confined matrix Y, the vectors in $Y - P_B Y_T$ are orthogonal to those in N_T. Therefore

$$\min_{\mathrm{col}(Y)\subseteq\mathcal{U}_B} \frac{\|Y - Y_T\|_F^2}{2n} = \frac{\|N_T\|_F^2}{2n}, \qquad Y^\star = P_B Y_T. \tag{B.2}$$

The optimizer is unique among response matrices. Every base-confined correction leaves the teacher-margin error unchanged because its projection onto $\mathcal{U}_N$ is zero.

For unrestricted Y_S, write $E = Y_S - Y_T$. Orthogonality and $P_\perp Y_T = 0$ give

$$\|E\|_F^2 = \|P_B E\|_F^2 + \|P_N E\|_F^2 + \|P_\perp Y_S\|_F^2 = 2n(\mathcal{L}_{\mathrm{br}} + \mathcal{L}_{\mathrm{fid}}). \tag{B.3}$$

This proves Theorem 1. If $E = 0$, both losses vanish. Conversely, simultaneous vanishing forces $E = 0$. Each loss is bounded above by $\mathcal{L}_{\mathrm{resp}}$; separate bounds $\varepsilon_{\mathrm{br}}$ and $\varepsilon_{\mathrm{fid}}$ imply total error at most their sum. No component can cancel another.

For the two-dimensional example, $Y_B = e_1$, $Y_T = e_1 + e_2$, and $Y_{S_a} = e_1, Y_{S_b} = e_2$. Both students have unit norm, rank one, squared error one, and teacher cosine $1/\sqrt{2}$. Their breaking–fidelity pairs are $(1/2, 0)$ and $(0, 1/2)$. Adding $\Delta_N = P_N(Y_T - Y_S)$ leaves total error $\mathcal{L}_{\mathrm{fid}}$; adding $\Delta_F = (I - P_N)(Y_T - Y_S)$ leaves $\mathcal{L}_{\mathrm{br}}$. Replacing the missing component makes each example exact.

B.2 Spectral accuracy costs and numerical teacher-margin alignment

Every missing direction has a spectral cost. Let $q = \mathrm{rank}(N_T)$ and write its compact SVD as $N_T = \sum_{j=1}^{q} \sigma_j u_j v_j^\top$. For any integer $0 \leq k \leq q$, the Eckart–Young theorem (Eckart & Young, 1936) gives

$$\min_{\mathrm{rank}(A)\leq k} \|A - N_T\|_F^2 = \sum_{j>k} \sigma_j(N_T)^2. \tag{B.4}$$

The minimizer $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^\top$ lies in $\mathcal{U}_N$, so it is also an admissible target projection $P_N Y_S$. In particular, $Y^{(k)} = P_B Y_T + A_k$ satisfies $P_B Y^{(k)} = P_B Y_T$, $P_N Y^{(k)} = A_k$, and $P_\perp Y^{(k)} = 0$. It attains the spectral tail with zero fidelity error, proving Proposition 3.1 and the main-text construction.

The full-rank threshold. Let $q = \operatorname{rank}(N_T) > 0$, $\gamma = \sigma_q(N_T) > 0$, and $D = P_N Y_S - N_T$. Since $P_N Y_S$ has columns in the q-dimensional space $\mathcal{U}_N$, its rank is at most q. If $\mathcal{L}_{\mathrm{br}} < \gamma^2/(2n)$, then

$$\sigma_q(P_N Y_S) \geq \sigma_q(N_T) - \|D\|_2 \geq \gamma - \|D\|_F > 0. \tag{B.5}$$

Thus its rank is exactly q, proving Corollary 1. Exact breaking-error zero gives the stronger equality $P_N Y_S = N_T$, while $P_N Y_B = 0$.

Deleting the last SVD term gives a rank-$(q-1)$ target projection with squared error exactly γ^2, establishing the sharp boundary of the strict threshold.

Numerical rank and discarded student energy. For $A = P_N Y_S$, let A_k be a rank-k SVD truncation and $\epsilon_k = \|A - A_k\|_F$. Combining Equation (B.4) with the triangle inequality yields

$$\sqrt{2n\mathcal{L}_{\mathrm{br}}} \geq \left[\left(\sum_{j>k} \sigma_j(N_T)^2 \right)^{1/2} - \epsilon_k \right]_+, \qquad [u]_+ = \max(u, 0). \tag{B.6}$$

This form uses both the teacher spectral tail and the discarded student energy.

Retained projectors and residual-inclusive matching. Use $\widehat{P}_B, \widehat{P}_N, Z_T$ from Equation (A.2), and set $\widehat{P}_\perp = I - \widehat{P}_B - \widehat{P}_N$. The retained target is $Z_q = \widehat{P}_N Y_T$, with teacher tail $R_T = \widehat{P}_\perp Y_T = Z_T - Z_q$. Orthogonality gives

$$\begin{aligned} \|Y_S - Y_T\|_F^2 = \|\widehat{P}_B(Y_S - Y_T)\|_F^2 + \|\widehat{P}_N Y_S - Z_q\|_F^2 \\ + \|\widehat{P}_\perp Y_S - R_T\|_F^2. \end{aligned} \tag{B.7}$$

The exact-span identity follows by retaining all nonzero singular directions. The spectral bound applies to the retained target Z_q when the student projection is $\widehat{P}_N Y_S$.

For the numerical component sums in Table B.3, define

$$\widehat{\mathcal{L}}_{\mathrm{br}} = \frac{\|\widehat{P}_N(Y_S - Y_T)\|_F^2}{2n}, \qquad \widehat{\mathcal{L}}_{\mathrm{fid}} = \frac{\|\widehat{P}_B(Y_S - Y_T)\|_F^2 + \|\widehat{P}_\perp Y_S\|_F^2}{2n}. \tag{B.8}$$

Expanding the last term of Equation (B.7) yields the teacher-tail correction

$$\begin{aligned} \mathcal{L}_{\mathrm{resp}} &= \widehat{\mathcal{L}}_{\mathrm{br}} + \widehat{\mathcal{L}}_{\mathrm{fid}} + \Delta_{\mathrm{tail}}, \\ \Delta_{\mathrm{tail}} &= \frac{\|R_T\|_F^2 - 2\langle \widehat{P}_\perp Y_S, R_T \rangle_F}{2n}. \end{aligned} \tag{B.9}$$

This signed correction includes the teacher-tail energy and its cross term with the student's residual response. Retaining the full teacher margin gives $R_T = 0$ and recovers the exact component sum.

Base truncation also leaves $R_B = (I - \widehat{P}_B)Y_B$, so for $\Delta = Y_T - Y_B$,

$$Z_T = (I - \widehat{P}_B)\Delta + R_B. \tag{B.10}$$

The change-energy share in Equation (4) measures $(I - \widehat{P}_B)\Delta$; the teacher norm-ratio scan measures Z_T.

Angle and magnitude resolve the response mismatch. For nonzero $A_S = (I - \widehat{P}_B)Y_S$ and Z_T, expanding the squared difference and dividing by $\|Z_T\|_F^2$ gives

$$\begin{aligned} \frac{\|A_S - Z_T\|_F^2}{\|Z_T\|_F^2} &= \rho^2 + 1 - 2\rho \operatorname{Align}_{\mathrm{margin}}(S) \\ &= (\rho - 1)^2 + 2\rho\big(1 - \operatorname{Align}_{\mathrm{margin}}(S)\big), \end{aligned} \tag{B.11}$$

where $\rho = \|A_S\|_F / \|Z_T\|_F$. Positive rescaling of A_S preserves the angle and changes its magnitude. The Frobenius inner product compares corresponding probe columns across the full base-external responses.

B.3 When the requirement also increases total rank

Target-space rank and total response rank answer different questions. The latter admits a stronger conclusion when the teacher retains the entire base span.

Proposition B.1 (Total rank under exact inclusion)**.** *Suppose* $\mathcal{U}_B \subseteq \mathrm{col}(Y_T)$*,* $P_B Y_S = P_B Y_T$*, and* $\mathcal{L}_{\mathrm{br}} = 0$*. Then*

$$\mathrm{col}(Y_T) = \mathcal{U}_B \oplus \mathcal{U}_N, \qquad \mathrm{rank}(Y_S) \geq r_B + q. \tag{B.12}$$

Proof. For $y \in \mathrm{col}(Y_T)$, exact inclusion gives $P_B y \in \mathrm{col}(Y_T)$, so $(I - P_B)y$ also belongs to that space. Both $\mathcal{U}_B$ and $\mathcal{U}_N$ therefore lie in the teacher span. Equation (B.1) gives the reverse inclusion, and orthogonality gives $\mathrm{rank}(Y_T) = r_B + q$. The matching assumptions yield $(P_B + P_N)Y_S = Y_T$; projection cannot increase rank, proving the claim. □

The inclusion premise distinguishes rank addition from rotation. For $n = 1$, $Y_B = (1, 0)^\top$ and $Y_S = Y_T = (1, 1)^\top$ give rank-one base, margin, and student. The output components are orthogonal while their row spaces coincide. Proposition B.1 supplies the total-rank conclusion when the teacher retains the base span.

B.4 Target choice, one-step reachability, and composition

The target decides whether departure is required. For any target response $Y_{T'}$, define $N_{T'} = (I - P_B)Y_{T'}$. Exact matching implies

$$(I - P_B)Y_S = N_{T'}. \tag{B.13}$$

A zero margin imposes no base-external response; a nonzero margin forces one. Replacing N_T by $N_{T'}$ in Appendix B.2 gives the same quantitative target-space rank requirement.

A finite target can be reachable without repeated computation. If the probe matrix $X = [\xi_1, \ldots, \xi_n] \in \mathbb{R}^{d \times n}$ has full column rank, then $X^\dagger X = I_n$, and

$$A^\star = Y_T X^\dagger, \qquad A^\star X = Y_T. \tag{B.14}$$

The linear map $h(s) = A^\star s + b$ has exactly these finite-difference responses on the probes, for any nonzero perturbation scale. For general X, such a matrix exists if and only if $\ker X \subseteq \ker Y_T$. Necessity follows from $AX = Y_T$, and sufficiency from $Y_T X^\dagger X = Y_T$. This constructs an explicit one-step interpolant on the common probe set.

Composition as accumulated response change. For any sequence of common-representation response matrices $Y^{(0)} = Y_B, \ldots, Y^{(L)} = Y_T$, let $\Delta Y_\ell = Y^{(\ell)} - Y^{(\ell-1)}$. Then

$$N_T = \sum_{\ell=1}^{L} (I - P_B)\Delta Y_\ell. \tag{B.15}$$

The identity expresses the margin as accumulated base-external change. Table A.5 measures its size as the number of composed calls increases.

B.5 What selective and random response corrections recover

For $\Delta = Y_T - Y_B$, the numerical corrections are $Y_{\mathrm{base}} = Y_B + \widehat{P}_B \Delta$ and $Y_{\mathrm{add}} = Y_B + \widehat{P}_N \Delta$. Applying both leaves residual $\widehat{P}_\perp \Delta$. With exact full spans, the constructions become $P_B Y_T$, $Y_B + N_T$, and Y_T, respectively. Table B.1 gives the measured errors for all six backbones.

The base-only residual fraction is $\|(I - \widehat{P}_B)\Delta\|_F^2 / \|\Delta\|_F^2$. The reported values are 99.2%, 97.4%, 95.4%, 98.2%, 90.5%, and 95.0% in table order, yielding the mean 96.0% used in the main text.

Table B.1: **Fitting only within the base leaves most of the change unlearned.** Response-construction errors $\mathcal{L}_{\text{resp}} = \|Y - Y_T\|_F^2/(2n)$ at $n = 64$. None is the unmodified base response Y_B. Margin-only and base-only corrections use $\widehat{P}_N\Delta$ and $\widehat{P}_B\Delta$; Full is the teacher target.

Backbone	None	Margin only	Base only	Both	Full
SD1.5	818116	6390	811866	141	0
FLUX.1-dev	2493	66	2427	≈ 0	0
klein-4B	10960	501	10461	1.5	0
SDXL	5553	101	5452	≈ 0	0
SD3.5-medium	2638	252	2387	≈ 0	0
SD3.5-Large	16372	815	15558	≈ 0	0

Table B.2 compares teacher-aligned reconstructions with two energy-matched random controls. Both random increments are matched to $\|Z_q\|_F$, where $Z_q = \widehat{P}_N Y_T$ uses the retained projector in Equation (A.2). For $G, G' \in \mathbb{R}^{p\times n}$ with i.i.d. standard normal entries, the full-space and margin-space controls are, respectively,

$$Y_{\text{rand}} = Y_B + \frac{\|Z_q\|_F}{\|G\|_F} G,$$
$$Y_{\text{rand},N} = Y_B + \frac{\|Z_q\|_F}{\|\widehat{P}_N G'\|_F} \widehat{P}_N G'.$$

Each scale factor matches the increment over the entire matrix, across all probe columns. Both controls add their random increment to Y_B; they differ only in whether the increment is restricted to the retained margin span. Alignment sharply reduces error in every backbone. These constructions separate occupying the target space from matching its probe-wise response values.

Table B.2: **Teacher alignment determines reconstruction accuracy.** Response errors at $n = 64$. Both random controls add increments of Frobenius norm $\|Z_q\|_F$ to Y_B, in the full response space or the retained margin span.

Backbone	Teacher-aligned	Random, matched energy	Random in margin span
SD1.5	288	1825735	1901373
FLUX.1-dev	≈ 0	4920	4927
klein-4B	≈ 1.5	21416	21226
SDXL	≈ 0	11002	11083
SD3.5-medium	≈ 0	5028	5035
SD3.5-Large	≈ 0	31908	31825

Table B.3 reports total response error and the numerical component sum of Equation (B.8) at $n = 64$. SD1.5 and FLUX use the final weights of the complete samplewise configuration with DM, expansion, alignment, and GAN specified in Appendix D.2; klein uses our trained one-step student, and SDXL and the two SD3.5 models use released one-step students. The base and student comparisons share teacher targets, measurement locations, probes, output representations, measurement seeds, and the $1/(2n)$ normalization, following Appendix A.2.

At these matched measurements, the SD1.5 student's complete response error is 892,861 versus 818,116 for the base; the FLUX student's is 5,078 versus 2,493. The FLUX student has $\widehat{\mathcal{L}}_{\text{br}} = 2455$ and $\widehat{\mathcal{L}}_{\text{fid}} = 2623$. Complete error compares corresponding response columns under the common probes and depends on response magnitude, base-component matching, and the residual term. The 12-pair response alignment in Appendix A.4 measures angular agreement; generation metrics evaluate image distributions and prompt following.

For SD1.5, the teacher-tail correction is $\Delta_{\text{tail}} = 210$, approximately 0.024% of total error. Equation (B.9) accounts for both the teacher-tail energy and its cross term.

Table B.3: **Response errors and truncated-projection component sums.** One-step students at $n = 64$. Hatted losses use Equation (B.8); Equation (B.9) gives their teacher-tail correction. Relative gaps divide the absolute difference by total error. Approximate zeros retain the reported precision.

Backbone	$\mathcal{L}_{\mathrm{resp}}$	$\widehat{\mathcal{L}}_{\mathrm{br}} + \widehat{\mathcal{L}}_{\mathrm{fid}}$	Relative gap
SD1.5	892861	892651	0.024%
FLUX.1-dev	5078	5078	≈ 0
klein-4B	12238	12237	≈ 0
SDXL	5548	5548	≈ 0
SD3.5-medium	5792	5792	≈ 0
SD3.5-Large	17461	17461	≈ 0

Residual energy and residual matching in the paired FLUX analysis. At P1-N1 under the baseline paired protocol, let

$$U_S = \frac{\|\widehat{P}_\perp Y_S\|_F^2}{2n}, \qquad K_{\mathrm{tail}} = \frac{\|\widehat{P}_\perp Y_T\|_F^2}{2n}, \qquad C_S = \frac{\langle \widehat{P}_\perp Y_S, \widehat{P}_\perp Y_T\rangle_F}{2n}. \tag{B.16}$$

Then $e_R = U_S + K_{\mathrm{tail}} - 2C_S$, $\widehat{\mathcal{L}}_{\mathrm{fid}} = e_B + U_S$, and $\Delta_{\mathrm{tail}} = K_{\mathrm{tail}} - 2C_S$. Table B.4 gives this decomposition for the three students with $H = 2436.4$, $K_{\mathrm{tail}} = 0.02073431$, $n = 64$, and retained margin rank $q_\tau = 59$. Thus U_S is student residual energy, while e_R measures student–teacher residual mismatch.

Table B.4: **Teacher-tail and cross-term accounting at P1-N1.** The same $n = 64$, $\delta = 0.05$, $\tau = 0.01$ record supplies this table, the precision spectrum, and the threshold sweep. Here $K_{\mathrm{tail}} = 0.02073431$. Every row obeys $e_R = U_S + K_{\mathrm{tail}} - 2C_S$ and $|C_S| \le \sqrt{U_S K_{\mathrm{tail}}}$.

Student	U_S	C_S	$U_S + K_{\mathrm{tail}} - 2C_S$	e_R
DM	256.62000391	0.02037993	256.59997835	256.59997835
SW	303.61958999	0.02016125	303.60000180	303.60000180
COV	666.10525721	0.01297622	666.10003909	666.10003909

The same singular spectrum gives retained rank $q_\tau = 59$, leading singular value $\sigma_1 = 126.34797$, and teacher-tail energy $K_{\mathrm{tail}} = \sum_{j=60}^{64} \sigma_j^2/128 = 0.02073431$. The sharp cutoff bound is $K_{\mathrm{tail}} \le 5(0.01\sigma_1)^2/128 = 0.06235864$. The table also satisfies $|C_S| \le \sqrt{U_S K_{\mathrm{tail}}}$ and $U_S \le \rho^2 H$. Appendix A.8 specifies the joint coordinate representation used to compute these quantities and every threshold-dependent component.

B.6 An absolute local rank requirement

A nondegenerate output distribution also imposes a local rank requirement. The following statement follows from the area formula for Lipschitz maps (Federer, 1996).

Proposition B.2 (Local full rank on a set of positive probability)**.** *Let $z \in \mathbb{R}^m$, $p_z \ll \mathcal{L}^m$, and let $f_c : \mathbb{R}^m \to \mathbb{R}^p$, $p \ge m$, be Lipschitz for a fixed condition c. Suppose $q_c = (f_c)_{\#} p_z$ has a nonzero component absolutely continuous with respect to $\mathcal{H}^m$ on a countably m-rectifiable set S_c. Then*

$$p_z\{z : \operatorname{rank} Df_c(z) = m\} > 0. \tag{B.17}$$

Proof. Differentiability holds outside a Lebesgue-null set, also null for p_z. Let E be the differentiable points with rank less than m. Their m-dimensional Jacobian $J_m f_c$ is zero. The area formula, applied on bounded pieces of E, states

$$\int_E J_m f_c(z)\, dz = \int_{\mathbb{R}^p} \#(E \cap f_c^{-1}(y))\, d\mathcal{H}^m(y). \tag{B.18}$$

Consequently $\mathcal{H}^m(f_c(E)) = 0$. If $p_z(E) = 1$, the pushforward is concentrated on this null image, and cannot possess the stated nonzero absolutely continuous component. Thus the complement of E has positive probability and consists, up to a null set, of full-rank points. □

The proposition characterizes local nondegeneracy for an absolutely continuous latent law and the stated output-support dimension. It complements the finite-response and base-relative requirements developed above.

C Feature Interfaces, Conditioning, and Distributional Support

Response geometry identifies the teacher components required for exact response matching on common probes. This geometry motivates examining how image-feature statistics encode spread, reference geometry, and conditional correspondence.

C.1 Representations, centering, and projected spectra

For features $f_i = \phi(x_i)$, let $\bar{f} = N^{-1}\sum_i f_i$, $C = N^{-1}\sum_i (f_i - \bar{f})(f_i - \bar{f})^\top$, and $M = N^{-1}\sum_i f_i f_i^\top$. They satisfy

$$M = C + \bar{f}\bar{f}^\top. \tag{C.1}$$

A dominant shared mean and a broad centered spectrum can therefore coexist, as observed in the conditional failure of Section 4. The diagnostic cm^2 measures the energy share in the dominant uncentered direction.

Variation across conditions is also different from variation within a condition. At the population level, for $m_c = \mathbb{E}[f \mid c]$,

$$\mathrm{Cov}(f) = \mathbb{E}_c[\mathrm{Cov}(f \mid c)] + \mathrm{Cov}_c(m_c). \tag{C.2}$$

The cross terms vanish because $\mathbb{E}[f - m_c \mid c] = 0$. The first term measures within-condition variation; the second measures variation between conditional means. The pooled diagnostic combines them.

To connect covariance spectra with the response decomposition, let $\widetilde{Y}_G$ be responses under a common representation and centering convention. Construct $\widetilde{P}_B, \widetilde{N}_T, \widetilde{P}_N$ from those matrices, exactly as in Section 3, and set $A = \widetilde{P}_N \widetilde{Y}_S$. Then

$$C_N(S) = AA^\top/n, \qquad \operatorname{rank} C_N(S) = \operatorname{rank} A. \tag{C.3}$$

The equality follows from $\ker(AA^\top) = \ker(A^\top)$. Exact response matching gives rank $\operatorname{rank}(\widetilde{N}_T)$, and the represented base projection is zero. In the original response representation, when the full reference response space equals $\mathcal{U}_B \oplus \mathcal{U}_N$, $P_\perp Y_S = 0$ is equivalent to $\mathrm{col}(Y_S)$ being contained in that space. The other terms of the canonical decomposition specify its amplitudes and base-projected responses.

A shared linear isometry preserves this response geometry. For a new representation or centering convention, recomputing the base and teacher projectors defines the corresponding geometry directly. DINO features provide the empirical image representation used in our training diagnostics.

The retained distribution-matching objective fits the conditional output distribution. Covariance expansion broadens global feature variation, and covariance alignment matches variation around each prompt mean. The distinction between spread and alignment also appears in alignment–uniformity (Wang & Isola, 2020) and VICReg (Bardes et al., 2022), while CKA measures agreement between representations (Kornblith et al., 2019). Our diagnostics use generated-image geometry to select complementary controls. Common-probe response measurements then identify the teacher components affected by those controls. The matched factorial in Appendix F.5 isolates their effects on generation, and the response comparisons in Appendices F.6–F.8 resolve direction, amplitude, and preservation of the teacher's base component.

Participation ratio measures the distribution of spectral energy. For positive eigenvalues $\lambda_1, \ldots, \lambda_r$, Cauchy–Schwarz gives

$$1 \le \mathrm{PR}(C) = \frac{(\sum_j \lambda_j)^2}{\sum_j \lambda_j^2} \le r, \tag{C.4}$$

with equality on the right exactly when all nonzero eigenvalues are equal. Projecting before forming the covariance isolates variation in the selected response subspace; computing c_{div} on the full feature covariance summarizes its overall spectral spread.

Samplewise expansion controls the shared feature direction. For nonzero features normalized to $u_i = f_i/\|f_i\|$, expanding the squared sum of normalized features gives

$$\mathcal{L}_{\text{exp}} = \frac{\|\sum_{i=1}^{B} u_i\|_2^2 - B}{B(B-1)}, \qquad B > 1. \tag{C.5}$$

For even B, equal numbers of u and $-u$ attain the minimum $-1/(B-1)$ with rank-one covariance. The identity shows that the samplewise expansion objective directly penalizes a shared feature direction. The retained training objective and conditional reference guide the variation during training.

An isotropic reference for base-external energy. To interpret base-external energy relative to an unstructured change, fix a rank-r_B projector P_B and let the normalized direction of a nonzero $p \times n$ change matrix Δ be uniform in ambient matrix space. Rotational symmetry gives

$$\mathbb{E}\left[\frac{\|(I - P_B)\Delta\|_F^2}{\|\Delta\|_F^2}\right] = 1 - \frac{r_B}{p}. \tag{C.6}$$

Indeed, the projected matrix subspace has dimension $(p - r_B)n$ out of pn. Appendix A.4 complements this energy reference with a Gaussian reference for angular alignment with the teacher residual.

At the two FLUX baseline pairs in Table C.1, the actual retained base rank is 64 in $p = 49152$ RGB coordinates. The isotropic expectation is therefore $1 - 64/49152 = 99.8698\%$, compared with measured shares of 97.45% and 97.25%. These proportions locate teacher–base changes relative to the base span in a high-dimensional output space. Teacher-directed response constructions in Appendix B.5 and paired student measurements in Appendix A.5 determine how well the actual responses match this target. For other RGB assays with 64 probes, $r_B \leq 64$ gives an isotropic reference of at least 99.8698%; each representation uses its own dimension, including the SD1.5 latent endpoint assay.

Table C.1: **Measured outside-base energy and its geometric reference.** FLUX baseline pairs at $\delta = 0.05$, $\tau = 0.01$, and $n = 64$. The reference uses the actual output dimension and retained base rank of each pair.

Pair	p	r_B	Measured ρ_Δ	Isotropic expectation
P1-N1	49152	64	97.45%	99.8698%
P4-N1	49152	64	97.25%	99.8698%

C.2 Reference geometry and conditional correspondence

For leading r-dimensional covariance eigenspaces U_r, U_r^{ref}, reference alignment ra equals $r^{-1}\|U_r^\top U_r^{\text{ref}}\|_F^2$, the mean squared cosine of their principal angles. Our diagnostic uses the covariance of 32 prompt-mean features with $r = 8$ in the common DINO representation.

To examine how this overlap depends on prompt assignment, let F_{ref} collect reference features as columns, $H = I - \mathbf{1}\mathbf{1}^\top/N$, and Π be a permutation matrix. Because $\Pi H \Pi^\top = H$,

$$F_{\text{ref}}\Pi H \Pi^\top F_{\text{ref}}^\top/N = F_{\text{ref}} H F_{\text{ref}}^\top/N. \tag{C.7}$$

Permuting intact prompt groups therefore preserves the reference covariance and reference alignment. A deterministic CKA objective applied to these aggregate covariances has the same invariance.

A separate cross-prompt DINO check records 200 reference permutations per backbone in Table C.2. The means remain unchanged, with numerical differences around 10^{-16} and standardized differences approximately zero, as predicted by Equation (C.7). Tables F.4 and F.3 report the diagnostics used in the training comparisons.

Table C.2: **Reference geometry is preserved under prompt reassignment.** An independent DINOv2 (Oquab et al., 2024) cross-prompt check uses $N_c = 32$ and 200 teacher-reference permutations. Permuting the prompt assignments preserves the feature set used to compute its covariance.

Backbone	Original pairing	Permuted mean	Reported z
SD1.5	0.05124	0.05124	≈ 0
FLUX.1-dev	0.72734	0.72734	≈ 0
klein-4B	0.68246	0.68246	≈ 0
SDXL	0.01691	0.01691	≈ 0
SD3.5-medium	0.37587	0.37587	≈ 0
SD3.5-Large	0.59670	0.59670	≈ 0

We therefore preserve prompt associations in the training objectives. The samplewise alignment term uses associated reference images, while covariance alignment compares corresponding prompt groups, as specified in Appendix D.3.

C.3 A CONDITIONAL OUTPUT-TO-QUALITY BOUND

Conditional pairing also lets us relate output error to scalar image-quality functionals. The following bound uses the norm of the paired representations.

Proposition C.1 (Paired output control)**.** *Let Z_S^i, Z_T^i, $i = 1, \dots, n$, be student and teacher representations paired under the same condition, with n the number of pairs rather than the probe count of Section 3. Suppose a decoder D and scalar per-image quality function Q are L_D- and L_Q-Lipschitz on the compared sets. Define $\bar{Q}_G = n^{-1} \sum_i Q(D(Z_G^i))$. Then*

$$|\bar{Q}_S - \bar{Q}_T| \le L_Q L_D \sqrt{\mathcal{E}_{\text{out}}}, \qquad \mathcal{E}_{\text{out}} = \frac{1}{n} \sum_i \|Z_S^i - Z_T^i\|_2^2. \tag{C.8}$$

Proof. Apply the triangle inequality and the two Lipschitz bounds to each pair, then average. The resulting upper bound is $L_Q L_D$ times the mean paired representation distance. Cauchy–Schwarz bounds this mean by $\sqrt{\mathcal{E}_{\text{out}}}$. □

Table C.3 gives paired pixel MSE on the $[0, 1]$ scale. SD1.5 uses a distilled student, $N = 2000$, and 256×256 images; the other rows use base one-step proxies (the undistilled base evaluated for a single step) and $N = 200$. Every comparison pairs prompts and seeds against a 50-step teacher. For an image with d pixel coordinates, these values use $\|x\|_{\text{pix}}^2 = d^{-1} \sum_j x_j^2$; the proposition applies with Lipschitz constants in this norm.

Table C.3: **Paired output measurements.** Pixel-averaged MSE against the 50-step teacher, with the one-step model identified in each row.

Backbone	One-step object	Mean	Median	Std.
SD1.5	Distilled student	0.0877	0.0853	0.0311
FLUX.1-dev	Base proxy	0.0685	0.0638	0.0276
klein-4B	Base proxy	0.0862	0.0794	0.0430
SDXL	Base proxy	0.0890	0.0872	0.0269
SD3.5-medium	Base proxy	0.0503	0.0479	0.0193
SD3.5-Large	Base proxy	0.0718	0.0670	0.0267

C.4 CONDITIONAL CORRESPONDENCE AND NON-COLLAPSE ARE COMPLEMENTARY

At the distribution level, the reference assigned to each condition determines the target. Suppose a divergence $\mathcal{D}(q, p)$ is nonnegative and zero exactly when $q = p$, and the model can choose each

conditional distribution q_c independently. If the unconditional reference p_{ref} is attainable for every condition, the minimum is zero, attained exactly when

$$\mathbb{E}_c \mathcal{D}(q_c, p_{\mathrm{ref}}) = 0 \quad \Longleftrightarrow \quad q_c = p_{\mathrm{ref}} \text{ for almost every } c. \tag{C.9}$$

Replacing the pooled reference by $p(\cdot \mid c)$ makes the zero-divergence solution the desired conditional family. The construction isolates the objective-level role of conditioning; the FeatNN comparison in Appendix F.4 tests reference selection in a trained generator.

Distributional fidelity can nevertheless exclude exact point collapse. If $p_{\mathrm{ref}}(\{x_0\}) = 0$, then δ_{x_0} is not absolutely continuous with respect to p_{ref}, giving $\mathrm{KL}(\delta_{x_0} \| p_{\mathrm{ref}}) = +\infty$. If the reference has finite mean μ and nonzero finite variance $\sigma^2_{\mathrm{ref}} = \mathbb{E}\|X - \mu\|_2^2$, its only coupling to a point mass gives

$$W_2^2(\delta_{x_0}, p_{\mathrm{ref}}) = \|x_0 - \mu\|_2^2 + \sigma^2_{\mathrm{ref}} \geq \sigma^2_{\mathrm{ref}}. \tag{C.10}$$

Thus distributional fidelity excludes point collapse, while the condition-indexed target specifies which distribution each prompt should generate. Non-collapse and correspondence impose complementary requirements.

C.5 Distributional structure and optimization

Finite-moment equivalence classes. The feature objectives summarize distributions through finitely many statistics. For k such statistics, choose $k + 2$ support points and form their vectors $(1, \phi_1(x), \ldots, \phi_k(x)) \in \mathbb{R}^{k+1}$. Linear dependence gives a nonzero signed weight vector a with zero total mass and zero weighted statistics. Perturbing strictly positive probability weights by sufficiently small $\pm\eta a$ yields two different distributions with identical specified moments.

The polynomial degree also yields a bound on the number of nondegenerate modes in a low-degree maximum-entropy model. Suppose such a model exists with positive density $p(x) = Z^{-1}\exp(g(x))$, where g is a polynomial of degree at most $d \geq 2$, $Z < \infty$, and the complex critical set of g is finite. The gradient equations have degree at most $d-1$; the standard polynomial-system bound gives at most $(d-1)^D$ isolated complex solutions in dimension D (Basu et al., 2006). Every nondegenerate local maximum of p is among these critical points. In particular,

$$\#\{\text{nondegenerate local maxima of } p\} \leq (d-1)^D \leq d(2d-1)^{D-1}. \tag{C.11}$$

A target with more modes lies outside this low-degree maximum-entropy family.

Patch marginals and joint structure. A similar distinction arises between local image statistics and joint spatial structure. Let Π_j select a patch and let a finite transformation group G permute patch indices, with $\Pi_j \circ g = \Pi_{\pi_g(j)}$. If all selected patch marginals of P_1 agree, the group average

$$P_2 = |G|^{-1} \sum_{g \in G} g_{\#} P_1 \tag{C.12}$$

has the same patch marginals. The average is G-invariant; if P_1 is not, then $P_2 \neq P_1$. For example, a uniform distribution over cyclic rotations of the four binary patches $(1, 1, 0, 0)$ has the same one-patch marginals as its average over all permutations, but different adjacency structure. Joint constraints can therefore distinguish arrangements with identical selected patch marginals.

The dynamics of a moving adversary. Adversarial refinement adds a coupled optimization problem to the feature constraints. For $\min_\theta \max_\psi V(\theta, \psi)$ with $V = \theta\psi$, gradient flow satisfies $\dot\theta = -\psi$, $\dot\psi = \theta$, and conserves $\theta^2 + \psi^2$. Except at the origin, it rotates rather than converges. Simultaneous discrete updates with step size $\eta > 0$ instead satisfy

$$\theta_{k+1}^2 + \psi_{k+1}^2 = (1 + \eta^2)(\theta_k^2 + \psi_k^2), \tag{C.13}$$

and grow geometrically from any nonzero initialization. A fixed reference removes this two-player interaction from the feature objectives. We combine these fixed-reference signals with adversarial refinement to control spread and correspondence while refining image structure.

Latent correspondence and distributional matching. Common latent inputs and probes specify the teacher–student transfer map. For standard-normal z and an orthogonal map Q, $h_T(Qz, c)$ has the same conditional output distribution as $h_T(z, c)$, while its responses rotate relative to the fixed inputs. Response measurements therefore identify which teacher computation the student transfers through that map. The base reference separates preservation of existing teacher components from acquisition of the responses added by repeated computation. The factorial measures both transfer components and generation quality, and Appendix F.8 compares their resolution with full response angle, norm, and feature diagnostics. The same response ordering against guidance-1 and guidance-3.5 teachers connects the analysis to the training and generation-evaluation targets.

D Training and Evaluation

Each experiment retains its starting objective, adds complementary feature controls, and evaluates the final student under fixed prompts and sampling budgets. The following sections specify the models, references, optimization, and evaluation protocol used for these comparisons.

D.1 Models, initialization, and references

Table D.1 gives the six backbones in the response study. They span latent-diffusion UNets (Rombach et al., 2022; Podell et al., 2024), diffusion transformers (Peebles & Xie, 2023), and flow-matching models (Lipman et al., 2023; Esser et al., 2024). Generation experiments train students on five of these, with SD3.5-medium contributing the additional response measurements.

Table D.1: **Backbones and resolutions.**

Backbone	Model family	Parameters (B)	Image side
SD1.5	UNet, latent diffusion	0.9	512
SDXL	UNet, latent diffusion	2.6	1024
SD3.5-medium	MMDiT, flow matching	2.5	1024
SD3.5-Large	MMDiT, flow matching	8	1024
FLUX.1-dev	DiT, flow matching	12	1024
klein-4B	DiT	4	1024

All students we train initialize from their respective pretrained base models. The opening SD1.5 network-output assay compares the official pretrained base, the rank-32 standard adapter specified in Appendix D.7, fused into the UNet, LCM-LoRA (Luo et al., 2023b), and the official DMD2 checkpoint (Yin et al., 2024a). Among the instances described in Appendix A.1, LCM-B is official, while LCM-A and LCM-C are reproduced. The LCM control-swap study uses the LCM consistency objective (Luo et al., 2023a).

Training prompts are sampled from a filtered subset of LAION-5B (Schuhmann et al., 2022). We use only these text captions and generate the paired images offline with the teacher, yielding 500,000 prompt–teacher-image pairs. Each prompt has four distinct teacher images, generated offline by 50-step trajectories with different random seeds; the pair count includes these repeated prompt associations. These fixed images supply the alignment objectives, adversarial discriminator, and reference-based diagnostics. Their stored prompt associations determine the references used for alignment.

FeatNN selects $K = 64$ teacher references by cosine similarity between their prompt text embeddings and the current prompt's embedding, then matches features at corresponding spatial patch positions. For generated image x_i, let $F_i \in \mathbb{R}^{H_{\mathrm{pos}} \times d}$ be its frozen-encoder patch features and $R_{i,k} \in \mathbb{R}^{H_{\mathrm{pos}} \times d}$ those of its k-th selected reference, with H_{pos} patch positions and d feature channels. Hats denote row-wise unit-norm vectors, $\widehat{F}_{i,h} = F_{i,h}/\|F_{i,h}\|_2$ and $\widehat{R}_{i,k,h} = R_{i,k,h}/\|R_{i,k,h}\|_2$.

With temperature $\tau_{\mathrm{NN}} = 0.5$ and batch size m, the soft nearest-neighbor weights and objective are

$$w_{i,h,k} = \frac{\exp(\langle \widehat{F}_{i,h}, \widehat{R}_{i,k,h} \rangle / \tau_{\mathrm{NN}})}{\sum_{\ell=1}^{K} \exp(\langle \widehat{F}_{i,h}, \widehat{R}_{i,\ell,h} \rangle / \tau_{\mathrm{NN}})},$$
$$\mathcal{L}_{\mathrm{FeatNN}} = \frac{1}{m H_{\mathrm{pos}}} \sum_{i=1}^{m} \sum_{h=1}^{H_{\mathrm{pos}}} \left\| F_{i,h} - \mathrm{sg}\left(\sum_{k=1}^{K} w_{i,h,k} R_{i,k,h} \right) \right\|_2^2 . \tag{D.1}$$

Here sg stops gradients through the entire weighted target, including its softmax weights; generator gradients flow through $F_{i,h}$ in the regression error. Positionwise matching normalizes weights over neighbors independently at each patch position h, while the reference neighborhood is selected per prompt and shared across positions. The weighted target and squared error use unnormalized patch features. Pooled-reference FeatNN changes only reference selection, drawing its K neighbors randomly across prompts while retaining the objective, target construction, and gradient path.

For klein, the teacher is the undistilled base sampled for 50 steps with classifier-free guidance 4.0 (Ho & Salimans, 2021). Evaluating both conditional and unconditional branches at each step gives 100 branch evaluations. For comparison, the official distilled checkpoint is evaluated at one, two, and four steps.

D.2 Optimization and complementary configurations

All students use full-parameter AdamW with learning rate 3×10^{-6}, warmup, global batch 32, and gradient-norm clipping at 1.0. Training uses H20 96 GB GPUs. SD1.5 and SDXL run for 15,000 generator updates; SD3.5-Large, FLUX, and klein run for 30,000 generator updates. Evaluation uses the final weights. SD1.5 uses eight GPUs with four images per GPU; FLUX and klein use 32 GPUs with one image per GPU. Features are gathered across devices before computing the feature objectives. Across all five backbones, DM configurations follow the two-time-scale update rule of SenseFlow (Ge et al., 2026). The DM auxiliary (fake) model updates every iteration, and the generator updates every five iterations, giving five fake-model updates per generator update. Thus 15,000 and 30,000 generator updates correspond to 75,000 and 150,000 outer iterations, respectively. Active discriminators take one update at each generator event, including warmup.

Table D.2: **Shared optimization, backbone-specific training duration.** All rows use global batch 32, AdamW, learning rate 3×10^{-6} with warmup, and gradient clipping at 1.0.

Backbone	Generator updates	Global batch	Initialization
SD1.5	15,000	32	Base
SDXL	15,000	32	Base
SD3.5-Large	30,000	32	Base
FLUX.1-dev	30,000	32	Base
klein-4B	30,000	32	Base

The retained objective, expansion, alignment, and adversarial terms have weights $1, 1, 100, 2$, respectively, when active. These weights are shared across backbones. The adversarial component extracts multi-layer features from generated images and fixed teacher references using a frozen visual network; a trainable discriminator head maps the features to a scalar score. Hinge adversarial training uses one discriminator update per generator update and a 200-generator-update warmup. Appendix D.7 specifies the frozen feature layers, discriminator architecture, DM sampling rule, and optimizer settings. Samplewise and covariance configurations share this component and its settings.

Table D.3: **Complementary configurations.** Each row retains the starting objective. The control-swap study exchanges samplewise expansion and alignment within LCM and FeatNN.

Starting recipe	Retained objective	Added constraints
LCM, SD1.5	Consistency	Alignment + GAN
FeatNN, SD1.5	Feature matching	Expansion + GAN
DM, SD1.5 / FLUX	Distribution matching	Expansion + Alignment + GAN
DM, covariance (Ours)	Distribution matching	Expansion + Alignment + GAN

The SD1.5 cumulative table starts with the official DMD2 checkpoint, evaluated without retraining; it is the same checkpoint as the DMD2 row in Table 4. The following three rows are trained independently from the SD1.5 base with DM plus expansion, then alignment, then GAN. These three configurations share the remaining settings. The official checkpoint supplies a model reference, while controlled component comparisons use the three from-base configurations. FLUX uses four from-base configurations, starting with DM and successively adding the same controls. The cumulative response analyses use the final weights of each samplewise configuration; Ours in the five-backbone main comparison uses the complete covariance recipe. In the control swap, GAN, data, update count, and remaining settings are fixed within each starting recipe; only the selected feature control changes. The samplewise/covariance comparison replaces the two feature controls together.

D.3 Features, losses, and inference

For the image-feature diagnostics and complementary controls, the frozen map ϕ uses DINOv2 ViT-L/14 pretrained on LVD-142M. Images are bilinearly resized to 518×518, normalized by ImageNet mean and standard deviation, and processed by the frozen encoder. We average the patch tokens to form each image feature. Extractor weights remain fixed while gradients propagate to the generated images.

Diagnostics use 32 prompts and eight samples per prompt. Spread c_{div} and concentration cm^2 use all 256 image features. For reference alignment ra, we average the eight features within each prompt and compare leading covariance eigenspaces of the 32 prompt means with $r = 8$. The raw feature statistics use the definitions in Section 4.2, with no extra feature normalization or cohort averaging. Cosine operations normalize their own arguments.

In the SD1.5 and FLUX implementation comparisons, samplewise and covariance configurations share a global batch of eight prompts with four independent initial noises per prompt. The resulting 32 student images are paired within prompts with the 32 fixed teacher references, using all four references for each prompt. Both implementations use the same sampling and pairing rules.

For this batch $\{(x_i, c_i, r_i)\}_{i=1}^{32}$, the samplewise alignment objective $\mathcal{L}_{\mathrm{align}}$ rewards squared cosine to the associated reference r_i and penalizes squared cosine across all pairs with $c_i \neq c_j$. The samplewise expansion objective $\mathcal{L}_{\mathrm{exp}}$ uses distinct generated-image pairs. Both objectives follow the definitions in Section 4.3.

Covariance training uses the same eight-prompt, four-sample grouping across all five backbones. Covariance expansion uses C_{gen}, computed from all 32 generated features centered by the global batch mean. Covariance alignment uses eight prompt groups, with generated and reference features centered separately within each group. For $a \in \{\mathrm{gen}, \mathrm{ref}\}$, let $F_{a,g} \in \mathbb{R}^{m_g \times d}$ contain group g's patch-averaged feature vectors as rows and let $\mu_{a,g}$ be their mean. With $m_g = 4$ for every group, the training estimator is

$$\widetilde{F}_{a,g} = F_{a,g} - \mathbf{1}\mu_{a,g}^{\top}, \qquad C_{a,g} = \frac{\widetilde{F}_{a,g}^{\top}\widetilde{F}_{a,g}}{m_g - 1}. \tag{D.2}$$

Each four-image group covariance has rank at most three; the global 32-image covariance has rank at most 31. The losses aggregate these estimates across the eight prompt groups and sampled training

batches. These feature-centered covariances enter the alignment score directly,

$$s_{\epsilon_{\mathrm{num}}}(C_{\mathrm{gen},g}, C_{\mathrm{ref},g}) = \frac{\langle C_{\mathrm{gen},g}, C_{\mathrm{ref},g}\rangle_F}{\sqrt{\|C_{\mathrm{gen},g}\|_F^2 + \epsilon_{\mathrm{num}}}\sqrt{\|C_{\mathrm{ref},g}\|_F^2 + \epsilon_{\mathrm{num}}}}, \qquad \epsilon_{\mathrm{num}} = 10^{-6}. \quad \text{(D.3)}$$

Here $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$. The covariance alignment loss $\mathcal{L}_{\mathrm{align}}^{\mathrm{cov}}$ is the negative arithmetic mean of $s_{\epsilon_{\mathrm{num}}}$ across the eight prompt groups. Covariance expansion $\mathcal{L}_{\mathrm{exp}}^{\mathrm{cov}}$ uses the same numerical stabilizer in the participation-ratio objective.

The one-, two-, and four-step evaluations reuse the same final student weights within each backbone. This sweep uses samplewise students for SD1.5 and FLUX and covariance students for SDXL, SD3.5-Large, and klein. The exact time and sigma grids are specified in Appendix D.7. Our students and the official klein distilled model use guidance 1 and one network evaluation per step. Published comparison budgets retain their original NFE units.

D.4 Metrics and evaluation randomness

Our evaluation follows the SenseFlow protocol (Ge et al., 2026), including caption selection, teacher-reference generation, metric implementations, and GenEval evaluation. COCO-5K uses all 5,000 images in the COCO 2017 validation split. For each image, we select the associated caption with the highest CLIP Score (ViT-B/32) against the original COCO image, yielding one evaluation prompt per image. Training and evaluation prompts do not overlap, and evaluation teacher references are generated separately from the training images. For each evaluation seed, a student generates one image per selected caption. The teacher reference set contains 5,000 images generated from the same prompts. Within each backbone and seed, we reuse this reference set across students and sampling budgets.

We use the same preprocessing and metric implementations as SenseFlow. Patch FID-T measures distributional agreement with teacher-generated reference images using 2048-dimensional InceptionV3 features (Heusel et al., 2017; Szegedy et al., 2016), following SenseFlow's patch-evaluation procedure. We compute CLIP Score with ViT-B/32 (Radford et al., 2021) and use HPSv2 version 2.1 (Wu et al., 2023), PickScore version 1 (Kirstain et al., 2023), and the official ImageReward weights (Xu et al., 2023). We evaluate compositional alignment on GenEval (Ghosh et al., 2023) using its standard 553 prompts and four samples per prompt.

For our evaluations, we use seeds 0, 42, and 999, compute each metric independently, and then report its arithmetic mean across seeds, including Patch FID-T. The evaluation-seed standard deviations are below 0.3 for Patch FID-T, 0.003 for CLIP and HPSv2, 0.15 for PickScore, 0.03 for ImageReward, and 0.005 for GenEval. Individual scores and paired differences for the samplewise–covariance comparison and the FLUX schnell comparison appear in Appendix D.5. Each main-table trained configuration uses one training run with seed 101 and its final checkpoint. The FLUX covariance factorial additionally uses independent training seeds 202 and 303, reported separately in Appendix F.5.

Baseline scores without daggers for SDXL, SD3.5-Large, and FLUX are quoted from Ge et al. (2026), Tables 1 and 7, under the same evaluation protocol used here. For these backbones, teacher-reference generation follows the corresponding SenseFlow configurations. Appendix E.1 retains the reported teacher and baseline NFE budgets, including 80 NFE for the SDXL and SD3.5 teachers and the 25- and 50-NFE FLUX teacher results. Table D.4 specifies the teacher configurations for SD1.5 and klein, the two additional backbones. Sampling steps count sampler updates; NFE counts network forward evaluations. Daggers mark our evaluations of official SD1.5 LCM/DMD2, klein, FLUX.1-schnell, and FLUX SenseFlow checkpoints under the common protocol. FLUX.1-schnell uses its official one-step inference configuration. The direct one-step FLUX SenseFlow evaluation in Appendix E.3 samples released four-step weights in one step. The published one-step SenseFlow SDXL and SD3.5 students are adapted from four-step weights for 6,000 updates. Appendix E gives the complete comparisons and checkpoint settings.

Exact patch extraction and feature statistics. Decode images as RGB floats in $[0, 1]$ at their native square resolution. Partition every image into a complete non-overlapping grid of 256×256 patches. This gives four patches for 512×512 images and 16 for 1024×1024 images. Patch origins are $(256i, 256j)$ in row-major order, with $i, j \in \{0, \ldots, L/256 - 1\}$. There is no whole-image

Table D.4: **Evaluation teachers for the additional backbones.** SDXL, SD3.5-Large, and FLUX use the corresponding SenseFlow teacher-reference configurations (Ge et al., 2026). The SD1.5 and klein teachers below use 50 sampling steps; klein uses two classifier-free-guidance branches per step, totaling 100 branch evaluations.

Backbone	Teacher checkpoint	Sampler	Steps	Guidance
SD1.5	Stable Diffusion v1.5 base	DDIM	50	1.0
klein-4B	FLUX.2-klein-base-4B	Euler	50	4.0

resize before cropping. Resize each patch once to 299×299 with PyTorch bilinear interpolation and antialiasing; map to $[-1, 1]$ and extract 2048-dimensional pool3 activations. We use the FID Inception weights from the standard PyTorch FID implementation, with internal resizing disabled and a single input normalization step. The extractor runs in FP32 without autocast.

Each seed therefore contributes 20,000 SD1.5 patches or 80,000 patches for a 1024-pixel backbone, from 5,000 images. Teacher and student use the same deterministic grid, and every patch has equal weight. Compute means and unbiased covariances in FP64 and use the symmetric positive-semidefinite covariance square-root form of FID, clipping negative eigenvalues caused by roundoff to zero. FID is computed separately per evaluation seed and then averaged. Quoted comparison rows retain their source-reported scores and provenance.

Evaluated student checkpoints. Table D.5 summarizes the evaluated student checkpoints and their use. The SD1.5 LCM evaluation fuses the specified adapter at scale one into the SD1.5 base; DMD2 replaces the full UNet. schnell and distilled klein replace the transformer. The released four-step FLUX SenseFlow checkpoint is used in the direct one-step stress test. The one-step SDXL and SD3.5 checkpoints are the dedicated adapted students used in the endpoint assay. The FLUX spectral assay also includes Hyper and Turbo-Alpha. VAE and text components follow the corresponding base identifiers in Appendix D.7.

Table D.5: **Evaluated student checkpoints.** The same weights are used for each model across its listed sampling budgets.

Checkpoint	Form and use
SD1.5 LCM	LoRA fused at scale 1
SD1.5 DMD2	Full UNet, one-step
FLUX.1-schnell	Full transformer
klein distilled	Full transformer; shared across budgets
FLUX SenseFlow	Released four-step transformer, direct one-step stress test
SDXL SenseFlow	Dedicated one-step adapted checkpoint
SD3.5-Medium SenseFlow	Dedicated one-step endpoint checkpoint
SD3.5-Large SenseFlow	Dedicated one-step adapted checkpoint

D.5 Paired score differences

Tables D.6 and D.7 report individual scores and paired differences for the fixed SW, COV, and official schnell checkpoints. For each metric, d_i is the difference between scores computed with the same evaluation seed $i \in \{0, 42, 999\}$. We report $\bar{d}$, the sample standard deviation s_d with denominator two, and the unadjusted interval $\bar{d} \pm 4.30265 s_d/\sqrt{3}$. The unit of replication is an independently generated evaluation set; these intervals condition on the fitted checkpoint.

For COV minus SW, ImageReward differences are 0.1264 on SD1.5 and 0.0686 on FLUX, with intervals $[0.1198, 0.1330]$ and $[0.0620, 0.0752]$. SD1.5 Patch FID-T improves by 0.15 with interval $[-0.2583, -0.0417]$, while the FLUX difference of -0.12 has interval $[-0.2583, 0.0183]$. The SD1.5 GenEval difference of 0.0028 has interval $[-0.0011, 0.0067]$; the FLUX difference of 0.0047 has interval $[0.0028, 0.0066]$. Against schnell, COV improves Patch FID-T by 3.62 and GenEval

by 0.0026; Table D.7 reports their paired intervals. The three-run factorial separately characterizes training variation in the contribution of each covariance term.

Table D.6: **Individual evaluation-seed scores.** SW and COV denote the complete samplewise and covariance recipes. Each row evaluates the fixed checkpoint used in Tables 3 and 4; schnell is the official one-step checkpoint. The arithmetic means reproduce those tables.

Backbone	Recipe	Seed	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
SD1.5	SW	0	16.24	0.3042	0.2464	20.9617	-0.1026	0.3788
SD1.5	SW	42	16.49	0.3057	0.2483	21.0821	-0.0779	0.3838
SD1.5	SW	999	16.53	0.3063	0.2487	21.1134	-0.0748	0.3852
SD1.5	COV	0	16.14	0.3068	0.2480	20.9493	0.0268	0.3833
SD1.5	COV	42	16.32	0.3082	0.2495	21.0397	0.0465	0.3852
SD1.5	COV	999	16.35	0.3084	0.2498	21.0710	0.0506	0.3877
FLUX	SW	0	17.00	0.3204	0.3053	22.3378	1.1339	0.6515
FLUX	SW	42	17.21	0.3216	0.3070	22.4354	1.1507	0.6559
FLUX	SW	999	17.27	0.3222	0.3075	22.4753	1.1570	0.6579
FLUX	COV	0	16.94	0.3203	0.3121	22.3603	1.2045	0.6571
FLUX	COV	42	17.08	0.3215	0.3134	22.4579	1.2203	0.6602
FLUX	COV	999	17.10	0.3218	0.3138	22.4678	1.2226	0.6621
FLUX	schnell	0	20.47	0.3169	0.2903	22.5616	1.0077	0.6570
FLUX	schnell	42	20.72	0.3183	0.2919	22.6692	1.0265	0.6599
FLUX	schnell	999	20.79	0.3188	0.2923	22.6891	1.0318	0.6547

Table D.7: **Paired differences across evaluation seeds.** Differences are COV minus the indicated comparator. SD uses denominator two; intervals are unadjusted 95% Student-t intervals with two degrees of freedom. All differences reuse the same prompts, initial-noise seeds, and teacher references within each backbone. These intervals quantify evaluation variation for fixed checkpoints.

Backbone	Contrast	Metric	Mean	SD	95% interval
SD1.5	COV–SW	P-FID-T	-0.150	0.044	[-0.258, -0.042]
SD1.5	COV–SW	CLIP	+0.0024	0.0003	[0.0017, 0.0031]
SD1.5	COV–SW	HPSv2	+0.0013	0.0003	[0.0006, 0.0020]
SD1.5	COV–SW	Pick	-0.0324	0.0173	[-0.0754, 0.0106]
SD1.5	COV–SW	IR	+0.1264	0.0026	[0.1198, 0.1330]
SD1.5	COV–SW	GenEval	+0.0028	0.0016	[-0.0011, 0.0067]
FLUX	COV–SW	P-FID-T	-0.120	0.056	[-0.258, 0.018]
FLUX	COV–SW	CLIP	-0.0002	0.0002	[-0.0006, 0.0002]
FLUX	COV–SW	HPSv2	+0.0065	0.0003	[0.0058, 0.0072]
FLUX	COV–SW	Pick	+0.0125	0.0173	[-0.0305, 0.0555]
FLUX	COV–SW	IR	+0.0686	0.0026	[0.0620, 0.0752]
FLUX	COV–SW	GenEval	+0.0047	0.0008	[0.0028, 0.0066]
FLUX	COV–schnell	P-FID-T	-3.620	0.082	[-3.823, -3.417]
FLUX	COV–schnell	CLIP	+0.0032	0.0002	[0.0027, 0.0037]
FLUX	COV–schnell	HPSv2	+0.0216	0.0002	[0.0212, 0.0220]
FLUX	COV–schnell	Pick	-0.2113	0.0100	[-0.2361, -0.1865]
FLUX	COV–schnell	IR	+0.1938	0.0030	[0.1863, 0.2013]
FLUX	COV–schnell	GenEval	+0.0026	0.0042	[-0.0077, 0.0129]

D.6 Reference generation and training cost

Table D.8 reports reference generation and end-to-end training for representative small and large backbones. The SD1.5 reference set takes 11.6 hours on eight H20 GPUs (92.8 GPU-hours); the FLUX set takes 27.4 hours on 32 GPUs (876.8 GPU-hours). Each reference set is generated once and reused across its factorial cells and training seeds. All teacher images are written to local storage before training.

SD1.5 DM+GAN training takes 9.00 hours, compared with 10.75 hours for the complete covariance recipe, at 15,000 generator events. FLUX takes 121.0 and 135.0 hours, respectively, at 30,000 events. The complete recipe therefore adds 19.4% and 11.6% training time at the same optimizer-event budgets. The reported cycle times are elapsed training time divided by generator-event count, including optimizer warmup, feature gathering, VAE/DINO forward and backward passes, data loading, and checkpoint writes. Common DINO features are shared between the discriminator and feature controls at each generator event; reference features are cached. The extra work is propagation of the two feature losses through the fixed extractor and decoder to the generator. Reference generation plus one complete training run totals 178.8 GPU-hours for SD1.5 and 5196.8 GPU-hours for FLUX. The final students retain one network evaluation per sampling step.

Table D.8: **Reference generation and training cost on H20 96 GB GPUs.** Reference cost generates 500,000 fixed teacher images. Training cost is per final checkpoint. A cycle contains five fake-model updates, one discriminator update, and one generator update; times include VAE/DINO work and communication.

Backbone	GPUs	Ref. GPU-h	DM+GAN GPU-h	Full GPU-h	Cycle s	Extra
SD1.5	8	92.8	72.0	86.0	2.16 / 2.58	19.4%
FLUX	32	876.8	3872.0	4320.0	14.52 / 16.20	11.6%

D.7 Detailed experimental settings

Implementations and update accounting. Distribution matching follows the teacher-minus-fake denoising construction of DMD2 (Yin et al., 2024a); the fake model takes five updates per generator update as in SenseFlow (Ge et al., 2026). Our added terms are the feature controls and the frozen-DINO discriminator specified below. We use a one-step generator, initialize both generator and fake model from the same pretrained base, freeze text encoders and VAE parameters, and train all generator and fake-model denoiser parameters. Generated images remain differentiable through the fixed VAE and visual extractor.

Generator and fake optimizers are AdamW with learning rate 3×10^{-6}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, and gradient clipping at 1.0. Learning rates increase linearly over 500 generator events, equivalently 2,500 fake-model updates, and then stay constant. The discriminator uses AdamW at 10^{-4} with $(0, 0.99)$, $\epsilon = 10^{-8}$, zero weight decay, and clipping at 1.0. Global batch size is 32 for each optimizer event. SD1.5 uses 8 GPUs with local batch 4; SDXL uses 16 with local batch 2; SD3.5-Large, FLUX, and klein use 32 with local batch 1. Training uses BF16 forwards and FP32 optimizer states and feature-statistic reductions. Reported checkpoints are final weights without EMA or validation-based selection.

Table D.9: **Optimizer events in the matched training comparisons.** Outer iterations update the fake model. At every fifth outer iteration, an active discriminator updates once and the generator updates once. Discriminator counts include the 200-event adversarial warmup.

Backbone	Generator	Fake / outer iterations	Active discriminator
SD1.5, SDXL	15,000	75,000	15,000
SD3.5-Large, FLUX, klein	30,000	150,000	30,000

LCM and FeatNN interventions take 15,000 generator and 15,000 discriminator updates, with no fake model. Each fake event samples a fresh generated batch and detaches the generator output. Each generator event uses one shared batch for DM, active feature controls, and GAN. The discriminator first updates on detached generated images and fixed teacher references, then its parameters are frozen for the generator update. The adversarial generator weight increases linearly from zero to 2 during its first 200 events; the discriminator is active throughout. Feature weights are 1 and 100 from the first generator event. Inactive terms have weight zero and no optimizer events.

Distribution-matching noise and targets. For a generated clean latent x, draw fresh standard Gaussian ϵ and $u \sim \mathrm{Uniform}[0.02, 0.98]$ independently for each image. SD1.5 and SDXL set $t =$

$\lfloor 1000u \rfloor$ and use the native scaled-linear schedule $\beta_t = (\sqrt{0.00085} + t(\sqrt{0.012} - \sqrt{0.00085})/999)^2$. With $\alpha_t = \sqrt{\prod_{j=0}^{t}(1-\beta_j)}$ and $\sigma_t = \sqrt{1-\alpha_t^2}$, the noised latent is $x_t = \alpha_t x + \sigma_t \epsilon$ and each model's denoised estimate is $(x_t - \sigma_t \widehat{\epsilon})/\alpha_t$. Flow backbones use $x_\sigma = (1-\sigma)x + \sigma\epsilon$, velocity target $\epsilon - x$, and denoised estimate $x_\sigma - \sigma\widehat{v}$. SD3.5 uses $\sigma = 3u/(1+2u)$; FLUX and klein use $\sigma = u$.

Let $\widehat{x}_T$ and $\widehat{x}_F$ be the teacher and fake-model denoised estimates at the same noised latent, prompt, and time. Define the stopped-gradient direction and generator surrogate by

$$g = \frac{\widehat{x}_F - \widehat{x}_T}{\mathrm{mean}_{\mathrm{coords}}\,|x - \widehat{x}_T| + 10^{-6}}, \qquad \mathcal{L}_{\mathrm{DM}} = \frac{1}{2m}\sum_{i=1}^{m} \|x_i - \mathrm{sg}(x_i - g_i)\|_2^2 . \tag{D.4}$$

The numerator and denominator are detached in g. The norm sums latent coordinates and the outer average is over images. The fake model minimizes mean coordinatewise squared noise-prediction error for diffusion backbones or velocity-prediction error for flow backbones on detached generated latents. No additional time-dependent loss weighting or time truncation is applied. Text conditioning is identical for teacher and fake models. Teacher and fake denoising use guidance 1, except klein's fixed teacher which uses guidance 4; FLUX guidance embeddings are set to 1. These choices, the displayed noising distributions, and the additional feature losses completely specify the adopted DM configuration.

Discriminator architecture. The frozen visual network is the same DINOv2 ViT-L/14 as the feature controls. We extract patch tokens after transformer blocks 6, 12, 18, and 24 (one-based indices), apply a separate trainable LayerNorm to each 1024-channel block output, and spatially average each block. Their concatenation gives a 4096-dimensional vector. The discriminator head is a spectrally normalized linear map $4096 \to 512$, GELU, and a spectrally normalized linear map $512 \to 1$, including biases and without dropout. LayerNorm and head parameters are trainable; DINO weights are frozen. The losses are $\mathbb{E}[\max(0, 1 - D(r))] + \mathbb{E}[\max(0, 1 + D(x))]$ for the discriminator and $-\mathbb{E}[D(x)]$ for the generator. There is no extra discriminator regularizer or image augmentation. Resizing and normalization match the feature-control pipeline. Every factorial cell uses this identical architecture and update rule.

Caption selection and reference seeds. We normalize LAION-5B captions with Unicode NFKC, trim and collapse whitespace, retain English captions with fastText language-identification probability at least 0.9, and require 5–60 whitespace-delimited tokens. We discard captions containing URLs, email addresses, or nonprinting characters. We deduplicate case-folded normalized text and remove captions matching any COCO-5K or GenEval evaluation prompt under that same normalization. Captions with token-set Jaccard similarity at least 0.8 to an evaluation prompt are also removed. After also excluding the eight response prompts, eligible captions are placed in a fixed deterministic order based on their normalized text. The first 32 are reserved for diagnostic and gradient calibration, and the following 125,000 are selected for training. Selection uses captions alone. For prompt index $i \in \{0, \ldots, 124999\}$, four offline teacher trajectories use seeds $1000000 + 4i + j$, $j = 0, 1, 2, 3$, giving 500,000 pairs. A training batch samples eight distinct prompt indices uniformly and uses all four references for each prompt; generated noise is independent of reference noise. These rules also define the caption pool for the LCM and standard-adapter controls.

Standard-adapter construction. Table D.10 lists the adapters and checkpoints used in the geometric comparisons. The 12-pair FLUX comparison uses DM, SW, and COV students trained for 30,000 generator updates with seed 101. The SD1.5 samplewise student uses 15,000 updates with seed 101. The official distilled klein checkpoint is distinct from its undistilled base teacher.

We train rank-32 standard adapters on the query, key, value, and output projections of all UNet attention blocks, freezing the remaining weights. It uses the standard SD1.5 noise-prediction MSE on teacher-generated training images, uniformly sampled integer $t \in \{0, \ldots, 999\}$, batch 32, AdamW at 10^{-4} with $(0.9, 0.999)$ and zero weight decay, and 2,000 updates with seed 17. The adapter is fused at scale 1 for both spectral assays. LCM-A and LCM-C use the same adapter targets and rank, the LCM consistency objective with a 50-point DDIM teacher grid, guidance 1, Huber constant 0.001, target EMA decay 0.95, AdamW at 10^{-4}, batch 32, and 15,000 updates; their seeds are 101 and 303. LCM-B uses the public adapter unchanged. The LCM-only and control-swap students

Table D.10: **Adapter and checkpoint settings.** Configurations for the trained adapters and the released checkpoints used as references.

Role	Configuration
Standard adapter	Rank 32, scaling 32, zero dropout
LCM-B	Public LCM-LoRA adapter, fused at scale 1
LCM-A / LCM-C	Rank-32 adapters trained for 15,000 updates with seeds 101 and 303
SD1.5 DMD2	Full one-step UNet
Released endpoint students	One-step SenseFlow students for SDXL, SD3.5-Medium, and SD3.5-Large at native resolutions
FLUX adapters	Hyper-FLUX 4-step LoRA, FLUX-Turbo-Alpha 8-step LoRA, official FLUX.1-schnell, and SenseFlow-FLUX 4-step student; each evaluated at the common assay input

instead train the full base UNet with the stated LCM objective and the common 3×10^{-6} optimizer, so the from-base diagnostic and interventions share their training parameterization.

Inference and endpoint sampler grids. For our SD1.5 and SDXL students, deterministic DDIM uses $\eta = 0$, native epsilon prediction, no clipping or thresholding, and descending grid $t_j = \mathrm{round}(999(1 - j/K))$, $j = 0, \ldots, K - 1$, followed by clean endpoint $t = -1$ with $\alpha = 1$. Thus one, two, and four steps use $[999]$, $[999, 500]$, and $[999, 749, 500, 250]$. Flow students use Euler updates from $\sigma_j = s(1 - j/K)$ to σ_{j+1}, $j = 0, \ldots, K - 1$, ending at zero. For SD3.5, $s(u) = 3u/(1 + 2u)$, giving four-step grid $[1, 0.9, 0.75, 0.5, 0]$; for FLUX and klein, $s(u) = u$, giving $[1, 0.75, 0.5, 0.25, 0]$. Dynamic shifting and ancestral noise are disabled. Our student guidance is 1 throughout.

The 50-step response teachers and offline training-reference teachers use these same DDIM or Euler grid formulas with $K = 50$, including the additional SD3.5-medium response teacher. Their guidance is 1 except klein's two-branch guidance 4. The six-backbone base map uses the first clean-sample estimate at the same initial state and schedule value; the teacher completes the trajectory from that state. The SD1.5 response representation is the clean latent. Other endpoint responses decode with the respective pretrained VAE, convert to floating RGB, and bilinearly resize to 128×128 with antialiasing. Generation images retain the native resolution. Published baseline scores retain their source sampler configurations; author-evaluated official SD1.5 LCM uses its native 50-origin-step LCM schedule, DMD2 uses the one-step DDIM grid above, and schnell uses Euler $[1, 0]$ with guidance 0. Official distilled klein uses the stated Euler grids at guidance 1. Direct one-step SenseFlow uses Euler $[1, 0]$ and guidance 1.

For evaluation teacher references, SD1.5 uses 50-step DDIM at guidance 1 and klein uses 50-step Euler at guidance 4. SDXL uses 40-step DDIM with $\eta = 0$ and guidance 5, and SD3.5-Large uses 40-step shifted Euler at guidance 4.5, each requiring 80 conditional/unconditional branch evaluations. FLUX uses 50-step Euler with guidance embedding 3.5 and 50 network evaluations. These evaluation-teacher settings follow the reported comparison protocol; the offline training and response teachers use the configurations specified above. All author-evaluated models use the same evaluation teacher reference within each backbone and seed.

Response prompts, states, and probes. Prompt P1 is "A red fox sitting on a moss-covered rock in a forest." Prompt P2 is "A blue ceramic teapot beside two yellow lemons on a wooden table." Prompt P3 is "A small sailboat on a calm lake beneath snow-covered mountains." Prompt P4 is "A black bicycle leaning against a white brick wall with a red door." Initial-noise seeds N1, N2, and N3 are 11, 23, and 47. Each seed draws a standard-normal tensor at the backbone's native latent shape using the CPU PyTorch generator, then transfers it to the device. The six-backbone probe-budget, teacher-step, random-correction, and base-span tests use P1 with state seed 1001 and probe seed 18001. The paired endpoint robustness locations are P1-N1 and P4-N1.

The replica cohort adds four prompts. Prompt P5 is "A silver kettle steaming beside a green cup on a kitchen counter." Prompt P6 is "Three orange umbrellas reflected in a rainy city street." Prompt P7 is "A spotted dog jumping over a fallen tree in a meadow." Prompt P8 is "A glass bowl of purple

grapes beneath a brass desk lamp." It uses the same initial-noise seeds 11, 23, and 47 and probe-seed formula as P1–P4. These eight response prompts and the calibration prompts are removed from the training-caption pool before selecting its 125,000 entries. The four covariance factorial configurations use neither control, expansion alone, alignment alone, or both controls, each trained for 30,000 generator updates with seeds 101, 202, and 303. The both-controls student with seed 101 is the COV student in the paired comparison.

At pair Pp-Nq, probe seed $17000 + 100p + q$ draws independent standard Gaussian directions, normalizes each to unit Euclidean norm in FP64, and stores the result in FP32. The 32- and 64-probe assays use prefixes of the same 128-probe sequence. Baseline steps are $\delta = 0.05$; the paired checks also use 0.025 and 0.10. All configurations at a pair share exactly these probes. In BF16 assays, addition is performed in FP32 before the network-input cast; full-FP32 checks preserve the input through the forward pass. Every map reuses its unperturbed output across all response columns.

The single-call six-dimensional family uses P1, state seed 11, six QR-orthonormalized Gaussian directions from seed 6101, and 80 coefficient vectors drawn uniformly from $[-1, 1]^6$ with seed 6102. Coordinates perturb the same state at scales 0.05, 0.1, and 0.2. SD1.5 uses $t = 500$; FLUX uses $\sigma = 0.5$. The separate SD1.5 full-output assay uses P1, state seed 11, probe seed 17101, $t = 999$, and 128 probes. Gaussian-reference controls use seed 6201. Permutation checks use 999 uniform derangements of 64 columns from seed 7001, drawing once per pair–scale setting and sharing the resulting permutations across students.

The feature-diagnostic cohort consists of the 32 reserved LAION captions selected before the training set. Generated diagnostic images use seeds 3100–3107 per caption. Teacher references use seeds 4100–4115; the first eight form the fixed reference half, and all 16 supply the split-half calibration. This same cohort supplies the starting profiles, pooled-reference comparison, cumulative feature diagnostics, samplewise/covariance diagnostics, and covariance factorial diagnostics. Table D.11 identifies the associated tables and reference sets. Response assays use P1–P8 and the state/probe seeds specified above. COCO-5K and GenEval supply the generation metrics, with seeds 0, 42, and 999 traversing prompts in ascending image-ID order or the supplied GenEval order. Teacher references use a separate generator with the same evaluation seed and prompt order.

Table D.11: **Prompt cohorts and reference assignments.** All control selections use diagnostics and gradient calibration on the reserved cohort. The generation columns report evaluation after those choices are fixed. A table with both diagnostic and generation columns uses the respective cohort for each column.

Quantities	Locations	Cohort and references
Starting and control-swap diagnostics	Tables 1, F.4, F.5	Calibration cohort; teacher seeds 4100–4107
Factorial and cumulative feature diagnostics	Tables F.10, F.3	Calibration cohort; same reference half
Threshold and gradient calibration	Equation (F.1)	Calibration cohort; split halves of 16 teacher samples
Generation columns and weight neighborhood	Tables 1, F.5, F.6	COCO-5K / GenEval; evaluation seeds 0, 42, 999

E Complete Generation Results

We report comparisons with existing students, fixed-weight step sweeps, and direct one-step inference.

E.1 Teachers and comparison models

All generation comparisons follow the SenseFlow evaluation protocol specified in Appendix D.4. Tables E.1, E.2, and E.3 collect results reported by Ge et al. (2026), Tables 1 and 7, retaining their original NFE budgets. Table E.3 also includes our evaluation of official FLUX.1-schnell at one

sampling step. Tables E.4 and E.5 report our evaluations of official checkpoints under the same protocol, with sampling-step budgets matching Table 4. A dash marks an unavailable score.

The SDXL comparison includes Phased Consistency Models (PCM) (Wang et al., 2024), Flash Diffusion (Chadebec et al., 2025), and Hyper-SD (Ren et al., 2024), alongside LCM (Luo et al., 2023a), DMD2 (Yin et al., 2024a), and SenseFlow (Ge et al., 2026).

Table E.1: **SDXL teachers and baselines.** Results reported by Ge et al. (2026) under the shared evaluation protocol; P-FID-T denotes Patch FID-T. One-step SenseFlow uses the adapted student described in the source paper.

Method	NFE	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
SDXL teacher	80	–	0.3293	0.2930	22.67	0.8719	0.5461
LCM-SDXL	4	30.63	0.3230	0.2824	22.22	0.5693	0.5036
PCM-SDXL	4	17.77	0.3242	0.2920	22.54	0.6926	0.4944
Flash-SDXL	4	23.24	0.3216	0.2830	22.17	0.4295	0.4715
SDXL-Lightning	4	16.57	0.3214	0.2931	22.80	0.7799	0.5332
Hyper-SDXL	4	17.49	0.3254	0.3000	22.98	0.9777	0.5398
DMD2-SDXL	4	18.72	0.3277	0.2963	22.98	0.9324	0.5779
SenseFlow-SDXL	4	21.01	0.3248	0.3010	23.17	0.9951	0.5784
SenseFlow-SDXL	1	–	0.3298	0.2818	22.24	0.8570	–
SenseFlow-SDXL	2	–	0.3263	0.2827	22.89	0.9192	–

Table E.2: **SD3.5-Large teachers and baselines.** Results reported by Ge et al. (2026) under the shared evaluation protocol, with the original NFE budgets; P-FID-T denotes Patch FID-T.

Method	NFE	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
SD3.5 teacher	80	–	0.3310	0.2993	22.98	1.1629	0.7140
SD3.5-Turbo	4	22.88	0.3262	0.2909	22.89	1.0116	0.6877
SenseFlow-SD3.5	4	17.48	0.3286	0.3016	23.01	1.1713	0.6955
SenseFlow (Euler)	4	20.26	0.3287	0.3008	22.90	1.2062	0.7098
SenseFlow-SD3.5	1	–	0.3332	0.2803	22.32	1.0651	–
SenseFlow-SD3.5	2	–	0.3326	0.2889	22.88	1.2022	–

Under our shared evaluation protocol, the one-step SD1.5 student improves all six generation metrics over official one-step DMD2 (Yin et al., 2024a) and reaches Patch FID-T 16.27.

Table E.3: **FLUX.1-dev teachers and baselines.** The dagger marks our evaluation of official FLUX.1-schnell at one sampling step; other results are reported by Ge et al. (2026), retaining their original NFE budgets. All follow the shared evaluation protocol; P-FID-T denotes Patch FID-T. Table E.10 reports direct one-step use of four-step SenseFlow weights.

Method	Budget	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
FLUX.1-dev teacher	50	–	0.3202	0.3000	23.18	1.1170	0.6699
FLUX.1-dev teacher	25	–	0.3207	0.2986	23.14	1.1063	0.6733
FLUX.1-schnell	4	–	0.3264	0.2962	22.77	1.0755	0.6807
FLUX.1-schnell†	1	20.66	0.3180	0.2915	22.64	1.0220	0.6572
Hyper-FLUX	4	23.47	0.3238	0.2963	23.09	1.0983	0.6193
FLUX-Turbo-Alpha	4	24.52	0.3218	0.2907	22.89	1.0106	0.4724
SenseFlow-FLUX	4	19.60	0.3167	0.2997	23.13	1.0921	0.6471
SenseFlow (Euler)	4	20.29	0.3171	0.3008	23.26	1.1424	0.6420
SenseFlow-FLUX	2	–	0.3207	0.2866	22.38	0.9296	–

Table E.4: **SD1.5 official checkpoints.** Daggers mark our three-seed evaluations. Our final covariance student appears in Table 4; the samplewise student supplies the SD1.5 fixed-weight sweep in Table E.6.

Method	Steps	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
LCM (official)†	1	90.90	0.2760	0.1486	19.18	-1.5904	0.2141
DMD2 (official)†	1	18.75	0.2997	0.2369	20.87	-0.1712	0.3569

Klein offers a comparison using the same official distilled weights (Labs, 2025) at three budgets. Against this checkpoint, our one-step student achieves Patch FID-T 13.67, compared with 54.84, 16.14, and 17.26 at one, two, and four steps. When given four steps with unchanged weights, our student also exceeds the 50-step teacher on all five reported non-FID metrics, including HPSv2 of 0.3014 versus 0.2883 and GenEval of 0.7884 versus 0.7817. Tables E.5 and E.6 provide the complete scores.

Table E.5: **Klein teacher and official distilled checkpoint.** Daggers mark our three-seed evaluations of the official models. The teacher uses 50 steps, CFG 4.0, and 100 branch evaluations; the distilled weights use guidance 1 and one NFE per step. A dash marks an unavailable metric.

Method	Steps	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
klein-base teacher†	50	–	0.3152	0.2883	22.59	0.9671	0.7817
klein official†	1	54.84	0.3061	0.2147	20.96	0.1536	0.5746
klein official†	2	16.14	0.3139	0.2920	22.82	1.0619	0.8124
klein official†	4	17.26	0.3122	0.2999	23.05	1.1226	0.8282

E.2 One-, two-, and four-step inference with fixed weights

To examine how trained students use additional sampling steps, we evaluate each backbone at all three budgets with the same final weights. SD1.5 and FLUX use the complete samplewise recipes, while SDXL, SD3.5-Large, and klein use covariance recipes. Table E.6 reports every metric under the three-seed averaging protocol, and Figure E.1 expresses the changes relative to one step. In these sweeps, one-step Patch FID-T is close to its best measured value, while additional evaluations increase HPSv2 on four of the five backbones.

Table E.6: **Complete fixed-weight step sweeps.** S denotes samplewise students on SD1.5 and FLUX, and C denotes covariance students on SDXL, SD3.5-Large, and klein. Each backbone reuses the same weights at one, two, and four steps, with one NFE per step. Values average metrics computed separately for three evaluation seeds. Lower Patch FID-T is better, while higher values are better for all other metrics.

Backbone	Steps	P-FID-T	HPSv2	Pick	CLIP	IR	GenEval
SD1.5 (S)	1	16.42	0.2478	21.0524	0.3054	-0.0851	0.3826
SD1.5 (S)	2	15.87	0.2392	20.7948	0.2945	-0.3456	0.3781
SD1.5 (S)	4	17.24	0.2400	20.6405	0.2899	-0.4640	0.3642
SDXL (C)	1	19.77	0.3034	22.6101	0.3289	1.0584	0.5906
SDXL (C)	2	19.46	0.3108	23.0633	0.3306	1.0784	0.5871
SDXL (C)	4	19.33	0.3146	22.9829	0.3253	1.0771	0.5813
SD3.5-Large (C)	1	15.94	0.3056	22.5304	0.3363	1.1938	0.7112
SD3.5-Large (C)	2	15.61	0.3128	22.8132	0.3352	1.2071	0.7126
SD3.5-Large (C)	4	15.86	0.3103	22.8843	0.3293	1.2156	0.7063
FLUX.1-dev (S)	1	17.16	0.3066	22.4162	0.3214	1.1472	0.6551
FLUX.1-dev (S)	2	17.00	0.3152	22.7342	0.3210	1.1636	0.6511
FLUX.1-dev (S)	4	17.33	0.3183	22.8738	0.3138	1.1629	0.6428
klein-4B (C)	1	13.67	0.2837	22.3547	0.3223	1.0402	0.7882
klein-4B (C)	2	13.62	0.2943	22.9134	0.3218	1.0687	0.7896
klein-4B (C)	4	14.30	0.3014	23.0817	0.3194	1.1147	0.7884

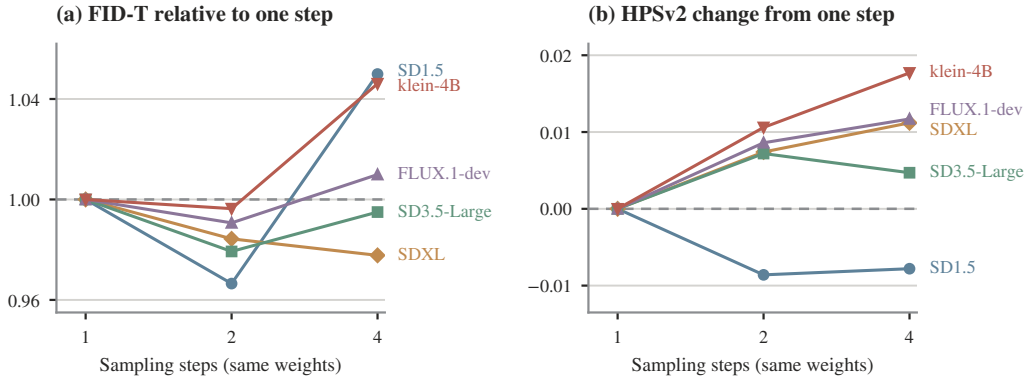


Figure E.1: **More steps for the same student.** The left panel shows Patch FID-T divided by the one-step value within each backbone, and the right panel shows the HPSv2 change from the one-step value. SD1.5 and FLUX use samplewise recipes, while the other three backbones use covariance recipes. Each backbone reuses the same weights across steps. Scores average three evaluation seeds.

To quantify the remaining Patch FID-T gain from additional steps, let $F(k)$ denote the score at $k \in \{1, 2, 4\}$ steps and define the gap as $100[F(1) - \min_k F(k)]/\min_k F(k)$. Table E.7 reports gaps from 0.37% to 3.47% across the five backbones. The minimizing budget is two steps for four backbones and four steps for SDXL.

Table E.7: **One-step Patch FID-T is close to the best measured budget.** The final column gives the relative gap to the minimum across one, two, and four steps using fixed weights. S denotes samplewise students on SD1.5 and FLUX; C denotes covariance students on the other three backbones.

Backbone	1 step	2 steps	4 steps	One-step gap
SD1.5 (S)	16.42	15.87	17.24	3.47%
SDXL (C)	19.77	19.46	19.33	2.28%
SD3.5-Large (C)	15.94	15.61	15.86	2.11%
FLUX.1-dev (S)	17.16	17.00	17.33	0.94%
klein-4B (C)	13.67	13.62	14.30	0.37%

E.3 Matched budgets and direct one-step inference

Tables E.8 and E.9 compare our students with SenseFlow (Ge et al., 2026) at each reported budget. Our FLUX rows use the samplewise student, while our SDXL and SD3.5-Large rows use covariance students, all with fixed weights across budgets. Published one-step SenseFlow students on SDXL and SD3.5 use dedicated adaptation from four-step weights. For FLUX, the one-step SenseFlow row tests whether the released four-step checkpoint can provide a one-step alternative through direct inference. In the two-step comparisons, our three backbones improve CLIP, HPSv2, and ImageReward over the corresponding SenseFlow entries.

Table E.8: **Same-backbone, matched-budget comparisons of fidelity and alignment.** SF denotes the published SenseFlow students (Ge et al., 2026). S denotes our samplewise FLUX student, and C denotes our covariance students on SDXL and SD3.5-Large. Budget is steps for ours and reported NFE for published SF rows. Daggers mark baselines evaluated by us. The one-step FLUX baseline uses four-step weights directly.

Backbone	Budget	FID ours	FID SF	CLIP ours	CLIP SF	HPS ours	HPS SF
SDXL (C)	4	19.33	21.01	0.3253	0.3248	0.3146	0.3010
SDXL (C)	2	19.46	–	0.3306	0.3263	0.3108	0.2827
SDXL (C)	1	19.77	–	0.3289	0.3298	0.3034	0.2818
SD3.5-L (C)	4	15.86	17.48	0.3293	0.3286	0.3103	0.3016
SD3.5-L (C)	2	15.61	–	0.3352	0.3326	0.3128	0.2889
SD3.5-L (C)	1	15.94	–	0.3363	0.3332	0.3056	0.2803
FLUX (S)	4	17.33	19.60	0.3138	0.3167	0.3183	0.2997
FLUX (S)	2	17.00	–	0.3210	0.3207	0.3152	0.2866
FLUX (S)	1	17.16	111.72†	0.3214	0.2878†	0.3066	0.1927†

Table E.9: **Same-backbone, matched-budget comparisons of reward and composition.** SF denotes the published SenseFlow students (Ge et al., 2026). S denotes our samplewise FLUX student, and C denotes our covariance students on SDXL and SD3.5-Large. Budget is steps for ours and published NFE for SF. Daggers mark baselines evaluated by us. The one-step FLUX baseline uses four-step weights directly.

Backbone	Budget	IR ours	IR SF	GenEval ours	GenEval SF
SDXL (C)	4	1.0771	0.9951	0.5813	0.5784
SDXL (C)	2	1.0784	0.9192	0.5871	–
SDXL (C)	1	1.0584	0.8570	0.5906	–
SD3.5-L (C)	4	1.2156	1.1713	0.7063	0.6955
SD3.5-L (C)	2	1.2071	1.2022	0.7126	–
SD3.5-L (C)	1	1.1938	1.0651	0.7112	–
FLUX (S)	4	1.1629	1.0921	0.6428	0.6471
FLUX (S)	2	1.1636	0.9296	0.6511	–
FLUX (S)	1	1.1472	-0.4822†	0.6551	0.4613†

For the direct one-step tests in Table E.10, the four-step FLUX SenseFlow weights yield Patch FID-T 111.72 and the undistilled klein base yields 250.28. Our trained one-step students achieve 17.04 and 13.67, respectively. These gaps illustrate the value of learning a mapping for the one-step budget.

Table E.10: **Direct one-step sampling.** Daggers mark our evaluations using a one-step inference budget.

Method	Steps	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
FLUX SenseFlow four-step weights†	1	111.72	0.2878	0.1927	20.53	-0.4822	0.4613
klein undistilled base†	1	250.28	0.2022	0.0956	18.13	-2.27	0.0258

F Training Comparisons and Conditional Failure

The cumulative results and control swaps relate the diagnostic profiles to generation quality. A matched factorial then isolates the covariance controls with DM and GAN fixed. Measurements across training seeds, prompt cohorts, and teacher configurations identify their effects on response direction, amplitude, and base-component error.

F.1 Teacher-margin alignment and generation across cumulative recipes

Tables F.1 and F.2 give all six generation metrics for the samplewise studies. SD1.5 includes an official DMD2 reference and three controlled configurations trained from the pretrained base. FLUX includes four from-base configurations retaining distribution matching and adding the indicated controls. Table F.3 pairs their feature diagnostics and mean common-probe response alignment with Patch FID-T and GenEval. Response alignment averages four prompts and three initial noises per prompt, giving 12 state-prompt pairs per backbone.

Table F.1: **SD1.5 cumulative samplewise configurations.** The official DMD2 checkpoint is the reference. The three trained configurations start from the pretrained base and use DM with the listed controls. Exp and Align denote samplewise controls. Values average three evaluation seeds. Lower Patch FID-T is better; the other metrics are higher-is-better.

Configuration	P-FID-T	HPSv2	Pick	CLIP	IR	GenEval
DMD2 official reference	18.75	0.2369	20.87	0.2997	-0.17	0.3569
DM + Exp	17.86	0.2412	20.93	0.3016	-0.14	0.3612
DM + Exp + Align	16.98	0.2456	21.00	0.3038	-0.11	0.3585
DM + Exp + Align + GAN	16.42	0.2478	21.05	0.3054	-0.09	0.3826

Table F.2: **FLUX cumulative samplewise configurations.** Each configuration is trained from the pretrained FLUX base and retains the distribution-matching (DM) objective. GAN is added only in the fourth configuration. Values average three evaluation seeds.

Configuration	P-FID-T	HPSv2	Pick	CLIP	IR	GenEval
DM	42.83	0.2757	21.90	0.3123	0.70	0.5316
+ Expansion	31.24	0.2896	22.14	0.3115	0.90	0.5873
+ Expansion + Alignment	18.72	0.3047	22.39	0.3208	1.12	0.6476
+ Expansion + Alignment + GAN	17.16	0.3066	22.42	0.3214	1.15	0.6551

On FLUX, adding samplewise expansion and alignment to DM increases response alignment from 0.0037 to 0.0088 and lowers Patch FID-T from 42.83 to 18.72. Adding GAN then brings these values to 0.0089 and 17.16. Across the three controlled SD1.5 configurations, response alignment rises from 0.0043 to 0.0062 and Patch FID-T falls from 17.86 to 16.42. The official DMD2 reference has response alignment 0.0026 and Patch FID-T 18.75. Appendix A.4 gives the shared-probe protocol and Gaussian reference for these comparisons. In Figure 2, the open SD1.5 marker denotes official DMD2. The segment connecting it to the first trained configuration compares independently initialized checkpoints. The three filled SD1.5 markers represent the controlled configurations trained from the pretrained base.

Table F.3: **Feature diagnostics, response alignment, and generation from the same students.** SD1.5 uses an official DMD2 reference and three from-base configurations; FLUX uses four from-base configurations. Trained configurations retain DM. Exp and Align denote samplewise controls. The DINOv2 diagnostics are spread c_{div}, reference alignment ra, and concentration cm^2. Response alignment reports the mean of $\mathrm{Align}_{\mathrm{margin}} \times 10^3$ over 12 state-prompt pairs per backbone (four prompts, three initial noises each), using the common-probe protocol in Appendix A.4.

Recipe	Spread	Ref. align.	Conc.	Response align.	P-FID-T↓	GenEval↑
SD1.5 / DMD2 ref.	18.99	0.8960	0.3332	2.6	18.75	0.3569
DM + Exp	20.38	0.8942	0.3651	4.3	17.86	0.3612
DM + Exp + Align	21.06	0.9018	0.3538	5.9	16.98	0.3585
DM + Exp + Align + GAN	21.55	0.9063	0.3470	6.2	16.42	0.3826
FLUX / DM	6.87	0.4823	0.7842	3.7	42.83	0.5316
DM + Exp	14.52	0.6591	0.6318	6.1	31.24	0.5873
DM + Exp + Align	20.34	0.8218	0.4267	8.8	18.72	0.6476
DM + Exp + Align + GAN	22.87	0.8724	0.3184	8.9	17.16	0.6551

Adding expansion and alignment also increases FLUX feature spread from 6.87 to 20.34 and reference alignment from 0.4823 to 0.8218, while reducing concentration from 0.7842 to 0.4267. Adversarial refinement further lowers concentration to 0.3184. Figure F.1 places these feature changes alongside the generation scores.

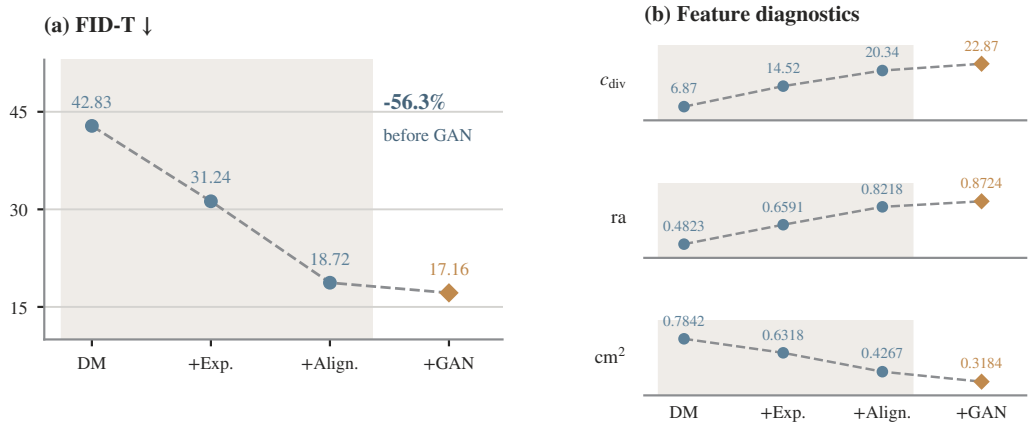


Figure F.1: **Generation and feature diagnostics across cumulative FLUX recipes.** The four samplewise configurations retain DM and successively add expansion, alignment, and GAN. Panel a reports Patch FID-T; panel b tracks spread, reference alignment, and concentration across the same recipe axis.

F.2 Control selection across starting recipes

Table F.4 gives the SD1.5 profiles that motivate the control swap. The from-base LCM-only recipe uses the same consistency loss, 15,000 generator updates, data, and training seed as the LCM interventions. Its spread of 23.41 and reference alignment of 0.8852 preserve the ordering observed with the official LCM checkpoint. LCM has broader spread and FeatNN stronger reference alignment. Its Patch FID-T / GenEval are 92.16 / 0.2118. Thus the diagnostic comparison also holds for the actual from-base starting recipe.

Table F.4: **SD1.5 starting profiles.** FeatNN is our feature-matching baseline. Diagnostics use 32 prompts with eight samples each; reference alignment compares 32 prompt means with $r = 8$.

Method	c_{div}	cm^2	reference alignment
LCM (official)	23.76	0.4416	0.8897
LCM (from-base, consistency only)	23.41	0.4498	0.8852
DMD2 (official)	18.99	0.3332	0.8960
FeatNN	19.32	0.4127	0.9184

To test the complements suggested by these profiles, we exchange expansion and alignment within each starting recipe while holding GAN, data, updates, and remaining settings fixed, as reported in Table F.5. LCM benefits more from alignment, reaching Patch FID-T 43.28 and GenEval 0.2874, compared with 68.34 and 0.2358 for expansion. FeatNN benefits more from expansion, reaching Patch FID-T 41.75 and GenEval 0.2831, compared with 65.92 and 0.2331 for alignment.

Table F.5: **Control selection across starting recipes.** FeatNN is our feature-matching baseline. Complete diagnostics and generation results for Figure F.2a. The official LCM checkpoint is identified as a reference. The from-base LCM-only row gives the actual starting recipe; its interventions share data, update count, and parameterization. GAN and all remaining settings are fixed when expansion and alignment are exchanged. All results use one-step SD1.5 generation and report metric means over seeds 0, 42, and 999.

Recipe	c_{div}	cm^2	ra	P-FID-T	GenEval
LCM (official reference)	23.76	0.4416	0.8897	90.90	0.2141
LCM (from-base starting recipe)	23.41	0.4498	0.8852	92.16	0.2118
LCM + Alignment + GAN	23.14	0.3852	0.8987	43.28	0.2874
LCM + Expansion + GAN	24.61	0.3971	0.8901	68.34	0.2358
FeatNN	19.32	0.4127	0.9184	83.47	0.2076
FeatNN + Expansion + GAN	21.63	0.3728	0.9106	41.75	0.2831
FeatNN + Alignment + GAN	19.87	0.3859	0.9241	65.92	0.2331

Although expansion gives LCM the greatest spread and alignment gives FeatNN the greatest reference alignment, the complementary control produces better generation in both cases. The preferred intervention therefore corrects each recipe's diagnosed deficiency.

Calibration and control-selection protocol. The LCM/FeatNN swap uses a fixed diagnostic cohort of 32 eligible captions, reserved before training-caption selection, with response and generation-evaluation prompts removed. Each prompt has eight generated samples. A separate set of 16 teacher samples per prompt supplies two disjoint eight-image reference sets. Across 200 seeded random equal splits, the reference spread has median 25.08 and standard deviation 1.34, and reference-subspace agreement between the halves has median 0.9581 and standard deviation 0.0152. The raw diagnostic profile uses the first reference half; the split distribution supplies a common scale for comparing deficiencies.

Define $z_{\mathrm{spread}} = \max(0, (25.08 - c_{\mathrm{div}})/1.34)$ and $z_{\mathrm{align}} = \max(0, (0.9581 - \mathrm{ra})/0.0152)$. Values above two flag a deficiency relative to teacher sampling variation. If either exceeds two, test the control associated with the larger standardized deficit first; otherwise retain the starting recipe. From-base LCM has $(z_{\mathrm{spread}}, z_{\mathrm{align}}) = (1.25, 4.80)$ and selects alignment. FeatNN has $(4.31, 2.61)$ and selects expansion. The larger standardized deficit selects the complementary control. Concentration is reported separately, and GAN is fixed in both swaps.

We set the samplewise weights by generator-parameter gradient calibration on 64 batches of eight prompts with four images each, using calibration seed 2718 and the 1,000-update starting-recipe checkpoints. Let θ_G contain all trainable denoiser parameters and let $g_j = \nabla_{\theta_G} \mathcal{L}_j$ for each active objective. We compare the Euclidean norms of the concatenated parameter gradients, accumulated in FP64 before loss weighting or gradient clipping, through the ratio

$$r_j = \mathrm{median}_{b=1}^{64} \frac{\|\nabla_{\theta_G} \mathcal{L}_j^{(b)}\|_2}{\|\nabla_{\theta_G} \mathcal{L}_{\mathrm{distill}}^{(b)}\|_2 + 10^{-12}}. \tag{F.1}$$

For LCM, the retained gradient follows the native consistency branch from the denoiser's clean-latent prediction to its stopped-gradient EMA target. For FeatNN it follows generated latents through the fixed VAE and DINO extractor to the feature-matching loss. The two auxiliary gradients follow their corresponding feature losses through DINO and the VAE to the same θ_G. LCM's consistency branch and the image-generation branch share denoiser parameters; each branch uses its prescribed time and noise sampling. Teacher and EMA targets, text encoders, VAE weights, DINO weights, and the reference images remain fixed during calibration. Unused parameter coordinates contribute zero. The generator is held at the 1,000-update checkpoint while the 64 gradient ratios are collected.

Pooling the LCM and FeatNN batch ratios gives medians 0.273 for expansion and 0.00261 for alignment. Choose each weight from $\{10^j : j = -2, \ldots, 3\}$ to make its median weighted ratio closest to 0.25 in log space; this gives 1 and 100. The same weights are carried into covariance

training and the five-backbone comparison. The diagnostic profiles and parameter-gradient ratios determine the choices; COCO and GenEval scoring follows the frozen choices.

The reference-subspace diagnostic selects the samplewise control from cross-prompt means. The covariance factorial also tracks within-prompt agreement, which measures the conditional variation targeted by its alignment objective. Spread and concentration jointly track the effect of samplewise expansion; the covariance formulation directly targets the participation ratio. Table F.6 checks a factor-of-two neighborhood around each selected weight, keeping GAN and the other settings fixed. The LCM preference for alignment and FeatNN preference for expansion both persist at all three multipliers.

Table F.6: **Control preference around the calibrated weights.** The multiplier scales $\lambda_{\text{exp}} = 1$ or $\lambda_{\text{align}} = 100$ for the named control. Every run starts from the SD1.5 base with training seed 101 and fixed GAN, data, and update counts; scores average evaluation seeds 0, 42, and 999.

Recipe	Control	Multiplier	P-FID-T	GenEval
LCM	Alignment	0.5	46.91	0.2796
LCM	Alignment	1	43.28	0.2874
LCM	Alignment	2	45.16	0.2832
LCM	Expansion	0.5	72.08	0.2281
LCM	Expansion	1	68.34	0.2358
LCM	Expansion	2	64.72	0.2417
FeatNN	Expansion	0.5	45.38	0.2740
FeatNN	Expansion	1	41.75	0.2831
FeatNN	Expansion	2	42.86	0.2813
FeatNN	Alignment	0.5	69.21	0.2274
FeatNN	Alignment	1	65.92	0.2331
FeatNN	Alignment	2	61.48	0.2406

F.3 Two implementations of complementary controls

To compare two implementations of the same complementary roles, we replace the samplewise expansion and alignment controls with their covariance formulations while retaining DM, GAN, and prompt associations. In Table F.7, covariance achieves better mean scores on five of six metrics on each backbone. Patch FID-T falls from 16.42 to 16.27 on SD1.5 and from 17.16 to 17.04 on FLUX. GenEval also rises from 0.3826 to 0.3854 and from 0.6551 to 0.6598, respectively. The covariance objectives express spread and within-prompt reference agreement directly through feature statistics. The complete covariance recipe supplies the Ours entries in Table 4; Appendix D.5 gives the paired confidence intervals for the score differences.

Table F.7: **Full metrics under auxiliary-loss replacement on SD1.5 and FLUX.** A and B use samplewise and covariance implementations of expansion and alignment, respectively. Both retain DM (Yin et al., 2024a), GAN, data, training budget, evaluation protocol, and loss weights. Delta is B minus A. Covariance improves five of six metrics on each backbone and supplies the Ours rows of Table 4.

(a) SD1.5.

Metric	Recipe A	Recipe B	Delta
P-FID-T (lower)	16.42	16.27	-0.15
CLIP (higher)	0.3054	0.3078	+0.0024
HPSv2 (higher)	0.2478	0.2491	+0.0013
Pick (higher)	21.0524	21.0200	-0.0324
IR (higher)	-0.0851	0.0413	+0.1264
GenEval (higher)	0.3826	0.3854	+0.0028

(b) FLUX.1-dev.

Metric	Recipe A	Recipe B	Delta
P-FID-T (lower)	17.16	17.04	-0.12
CLIP (higher)	0.3214	0.3212	-0.0002
HPSv2 (higher)	0.3066	0.3131	+0.0065
Pick (higher)	22.4162	22.4287	+0.0125
IR (higher)	1.1472	1.2158	+0.0686
GenEval (higher)	0.6551	0.6598	+0.0047

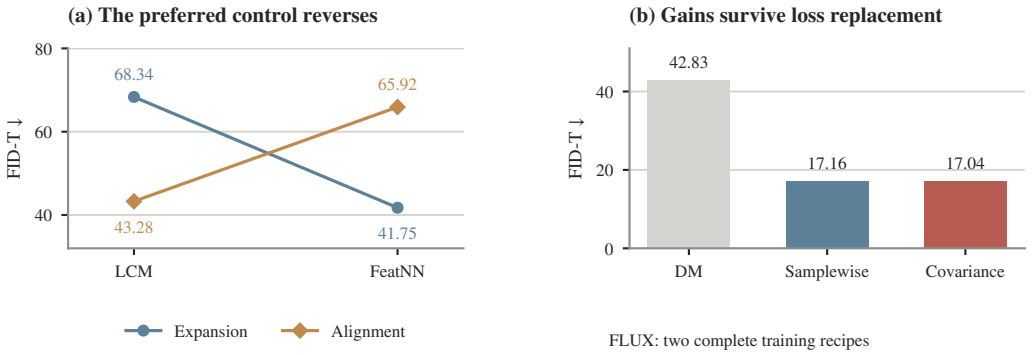


Figure F.2: **Control selection and loss formulations.** Panel a compares samplewise expansion and alignment for LCM (Luo et al., 2023a) and FeatNN on SD1.5. GAN and the remaining settings are fixed within each starting recipe. Lines connect the two recipes. Panel b compares DM with the complete samplewise and covariance recipes on FLUX. Both complete recipes retain DM and GAN. Scores average three evaluation seeds. Appendices F.2 and F.3 provide the full results.

F.4 A BROAD SPECTRUM CAN CONCEAL CONDITIONAL FAILURE

To isolate the role of conditional references, pooled-reference FeatNN replaces conditional reference selection with random selection from a pool spanning prompts while keeping the remaining training settings fixed. With this change, spread remains broad at 18.16, close to the conditional recipe's 19.32, while concentration rises to 0.9983, Patch FID-T to 449.25, and GenEval falls to 0.0042, as reported in Table F.8.

Table F.8: **Broad centered variation can coexist with conditional failure.** SD1.5, 32 prompts with eight samples each. Spread and concentration use all 256 features; reference alignment uses 32 prompt means with $r = 8$. The pooled variant randomizes reference selection across conditions, with other settings fixed.

Reference configuration	c_{div}	cm^2	ra	P-FID-T	GenEval
FeatNN	19.32	0.4127	0.9184	83.47	0.2076
Pooled-reference FeatNN	18.16	0.9983	0.4676	449.25	0.0042

The identity $M = C + \bar{f}\bar{f}^\top$ accounts for this coexistence of centered variation and a dominant shared component. The conditional reference supplies the prompt association needed to direct that variation toward the requested content.

F.5 A MATCHED FACTORIAL OF COVARIANCE CONTROLS

We fix the FLUX DM-plus-GAN configuration and independently activate covariance expansion and alignment with weights 1 and 100. All four cells use the same pretrained base, 500,000 training pairs, batches of eight prompts with four samples each, 30,000 generator updates, 150,000 fake-model updates, and 30,000 discriminator updates. The visual extractor, discriminator, adversarial warmup, and evaluation are identical. Table F.9 reports all six metrics for training seed 101, whose complete COV checkpoint also appears in the main comparison.

Table F.9: **All metrics for the matched FLUX covariance factorial.** All rows retain DM and GAN. Training seed 101 and evaluation seeds 0, 42, and 999 are shared; no checkpoint selection is performed.

Controls	P-FID-T	CLIP	HPSv2	Pick	IR	GenEval
Neither	28.63	0.3156	0.2942	22.1738	0.9724	0.6035
Expansion	22.08	0.3171	0.3008	22.2914	1.0637	0.6259
Alignment	23.74	0.3195	0.3052	22.3408	1.1026	0.6407
Both	17.04	0.3212	0.3131	22.4287	1.2158	0.6598

Table F.10: **Distinct effects of covariance expansion and alignment.** Diagnostics use the same 32-prompt, eight-image cohort. The within-prompt score is Equation (D.3) evaluated on eight images per prompt and averaged across prompts. Cross-prompt reference alignment and within-prompt covariance alignment measure different aspects of reference geometry.

Controls	c_{div}	ra	cm^2	Within-prompt score
Neither	12.41	0.7028	0.4652	0.4386
Expansion	22.15	0.7561	0.3397	0.4912
Alignment	16.08	0.8587	0.4018	0.6834
Both	23.16	0.8849	0.3072	0.7286

Expansion increases global spread by 9.74 from the matched baseline; alignment increases the within-prompt covariance score by 0.2448. Their combined configuration has the highest spread, reference alignment, and within-prompt score, with the lowest concentration. Adding expansion to alignment improves Patch FID-T by 6.70 and GenEval by 0.0191; adding alignment to expansion improves them by 5.04 and 0.0339. These contrasts isolate each term with the adversarial complement present.

Training seeds 101, 202, and 303 independently change data order and optimization randomness. Every run is evaluated with seeds 0, 42, and 999; Table F.11 reports the resulting run means. Both controls together rank first on Patch FID-T and GenEval in every training replicate. Across runs, the complete recipe reaches 17.06 ± 0.15 and 0.6596 ± 0.0016, compared with 28.67 ± 0.22 and 0.6035 ± 0.0017 for DM + GAN.

Table F.11: **Independent training replicates of the FLUX factorial.** Each seed column gives P-FID-T / GenEval averaged over the same three evaluation seeds. The last columns report the mean and sample SD across three independent training runs. Training seeds govern initialization of trainable heads, data order, and sampled noise; every generator starts from the same pretrained base.

Controls	Train 101	Train 202	Train 303	P-FID-T	GenEval
Neither	28.63 / 0.6035	28.91 / 0.6018	28.48 / 0.6052	28.67±0.22	0.6035±0.0017
Expansion	22.08 / 0.6259	22.35 / 0.6237	21.94 / 0.6281	22.12±0.21	0.6259±0.0022
Alignment	23.74 / 0.6407	23.51 / 0.6422	23.96 / 0.6389	23.74±0.23	0.6406±0.0017
Both	17.04 / 0.6598	17.22 / 0.6579	16.93 / 0.6611	17.06±0.15	0.6596±0.0016

The response comparisons below quantify the angular, amplitude, and base-component effects of these same controls. Replication across prompt cohorts and comparison with full-response summaries establish which components retain their ordering across training runs and teacher configurations.

F.6 Response behavior of the fixed-GAN factorial

We evaluate the four training-seed-101 covariance factorial checkpoints on the same four prompts and three initial noises used in the paired FLUX analysis. Each pair shares the base, 50-step teacher, 64 probe columns, $\delta = 0.05$, $\tau = 0.01$, and BF16-forward / FP32-difference / FP64-projection protocol. The both-controls cell is exactly the COV checkpoint and response record in Appendix A.5. Thus every intervention comparison keeps DM, GAN, training budget, and measurement coordinates fixed.

Table F.12: **Common-probe responses of the fixed-GAN covariance factorial.** All four training-seed-101 checkpoints use the same 12 pairs, base and teacher responses, 64 probes, $\delta = 0.05$, and $\tau = 0.01$. Entries are means with descriptive standard deviations in parentheses. Both denotes the same COV checkpoint as the paired comparison.

Controls	a	ρ	e_B	e_N	e_R	$\mathcal{L}_{\text{resp}}$
Neither	0.0057(0.0017)	0.7116(0.0322)	1214.8(21.1)	3237.7(143.6)	407.9(46.0)	4860.3(120.0)
Expansion	0.0072(0.0022)	0.8964(0.0295)	1122.4(41.7)	3889.7(164.3)	466.2(41.7)	5478.3(144.9)
Alignment	0.0083(0.0024)	0.7462(0.0312)	984.6(45.9)	3411.3(141.9)	346.3(44.7)	4742.3(122.0)
Both	0.0093(0.0027)	0.8700(0.0283)	887.0(41.0)	3726.5(191.2)	507.7(146.8)	5121.1(115.3)

Table F.13: **Paired effects with the adversarial component fixed.** Differences compare the four factorial cells at identical probes. Each contrast increases angular alignment at all 12 pairs. Error differences show which response components accompany each intervention.

Added control	Δa	$\Delta \rho$	Δe_B	Δe_N	Δe_R	$\Delta \mathcal{L}_{\text{resp}}$
Expansion without alignment	+0.0015	+0.1848	-92.4	+652.0	+58.3	+617.9
Expansion with alignment	+0.0010	+0.1238	-97.6	+315.1	+161.4	+378.8
Alignment without expansion	+0.0026	+0.0346	-230.2	+173.7	-61.6	-118.1
Alignment with expansion	+0.0021	-0.0264	-235.4	-163.2	+41.5	-357.1

With neither covariance term, mean response alignment is 0.0057 and the base-external norm ratio is 0.7116. Expansion alone reaches 0.0072 and 0.8964; alignment alone reaches 0.0083 and 0.7462; both reach 0.0093 and 0.8700. All four contrasts increase alignment at every primary-cohort pair for training seed 101. Expansion produces the larger amplitude change; alignment produces the larger angular gain and base-component error reduction.

Adding alignment to expansion decreases mean base-component error by 235.4 and retained-margin error by 163.2, with a complete-error decrease of 357.1. Adding expansion to alignment reduces base-component error by 97.6 while increasing complete error by 378.8. The following tables retain the distinct per-pair effects.

Table F.14: **Individual responses for the matched covariance factorial, prompts P1–P2.** All four cells retain the identical DM and GAN configuration. Norm ratio and all errors use the shared teacher and base at each pair.

Pair	Controls	a	ρ	e_B	e_N	e_R	$\mathcal{L}_{\text{resp}}$
P1-N1	Neither	0.0034	0.7257	1194.7	3254.5	453.0	4902.2
P1-N1	Expansion	0.0044	0.8907	1063.6	3846.0	504.6	5414.2
P1-N1	Alignment	0.0048	0.7477	891.4	3396.9	383.9	4672.2
P1-N1	Both	0.0055	0.8945	822.6	3695.6	666.1	5184.3
P1-N2	Neither	0.0038	0.6966	1192.1	3177.2	420.7	4790.0
P1-N2	Expansion	0.0048	0.8753	1087.0	3778.0	495.1	5360.1
P1-N2	Alignment	0.0054	0.7409	926.6	3359.6	386.4	4672.6
P1-N2	Both	0.0061	0.8403	834.7	3498.6	624.0	4957.3
P1-N3	Neither	0.0030	0.7361	1236.6	3387.2	356.4	4980.1
P1-N3	Expansion	0.0041	0.9251	1161.4	4072.2	428.5	5662.0
P1-N3	Alignment	0.0045	0.7675	1029.6	3556.4	296.2	4882.2
P1-N3	Both	0.0052	0.8462	948.5	3835.2	322.1	5105.8
P2-N1	Neither	0.0050	0.7626	1239.1	3409.6	410.5	5059.2
P2-N1	Expansion	0.0055	0.9329	1220.0	4063.0	451.8	5734.9
P2-N1	Alignment	0.0070	0.7971	996.5	3595.6	346.7	4938.8
P2-N1	Both	0.0077	0.8270	901.6	3575.1	481.1	4957.8
P2-N2	Neither	0.0048	0.7257	1198.3	3307.2	366.6	4872.1
P2-N2	Expansion	0.0059	0.9246	1073.4	4044.1	414.1	5531.6
P2-N2	Alignment	0.0073	0.7673	955.2	3520.9	292.9	4769.0
P2-N2	Both	0.0082	0.9049	847.5	4023.6	337.8	5208.9
P2-N3	Neither	0.0050	0.6717	1206.3	3063.8	448.2	4718.3
P2-N3	Expansion	0.0064	0.8598	1116.4	3681.2	520.3	5317.9
P2-N3	Alignment	0.0077	0.7117	1032.3	3271.7	364.3	4668.4
P2-N3	Both	0.0086	0.8928	908.7	3724.8	607.1	5240.6

Table F.15: **Individual responses for the matched covariance factorial, prompts P3–P4.** All four cells retain the identical DM and GAN configuration. Norm ratio and all errors use the shared teacher and base at each pair.

Pair	Controls	a	ρ	e_B	e_N	e_R	$\mathcal{L}_{\text{resp}}$
P3-N1	Neither	0.0063	0.6960	1172.7	3188.0	411.3	4772.1
P3-N1	Expansion	0.0081	0.8830	1098.1	3848.8	457.5	5404.4
P3-N1	Alignment	0.0096	0.7114	936.5	3309.0	331.4	4576.9
P3-N1	Both	0.0105	0.8967	845.9	3841.8	513.0	5200.7
P3-N2	Neither	0.0063	0.7557	1229.5	3440.8	363.8	5034.1
P3-N2	Expansion	0.0081	0.9435	1120.3	4150.7	417.7	5688.8
P3-N2	Alignment	0.0094	0.7985	1002.9	3642.6	310.7	4956.2
P3-N2	Both	0.0103	0.9043	943.8	3979.0	404.6	5327.4
P3-N3	Neither	0.0069	0.7084	1242.1	3253.1	360.6	4855.8
P3-N3	Expansion	0.0087	0.8839	1129.9	3854.2	422.9	5407.1
P3-N3	Alignment	0.0096	0.7451	1017.8	3441.4	290.8	4750.0
P3-N3	Both	0.0111	0.8722	898.8	3866.7	350.8	5116.3
P4-N1	Neither	0.0078	0.6531	1218.5	2981.8	450.4	4650.7
P4-N1	Expansion	0.0098	0.8580	1118.2	3648.5	517.3	5284.1
P4-N1	Alignment	0.0110	0.7003	993.9	3181.6	392.4	4568.0
P4-N1	Both	0.0123	0.8831	887.1	3612.0	648.3	5147.4
P4-N2	Neither	0.0080	0.7290	1217.2	3341.9	354.0	4913.0
P4-N2	Expansion	0.0101	0.9174	1114.1	3995.6	438.1	5547.8
P4-N2	Alignment	0.0119	0.7505	984.7	3433.9	324.0	4742.6
P4-N2	Both	0.0134	0.8402	872.8	3730.1	364.0	4966.9
P4-N3	Neither	0.0076	0.6789	1230.3	3047.0	499.1	4776.4
P4-N3	Expansion	0.0100	0.8622	1166.4	3693.5	526.3	5386.2
P4-N3	Alignment	0.0110	0.7169	1048.2	3226.4	435.9	4710.6
P4-N3	Both	0.0126	0.8377	931.9	3335.3	773.1	5040.3

F.7 Response replication and amplitude calibration

We apply the same common-probe protocol to the factorial checkpoints from training seeds 202 and 303. Each seed is assessed on the original P1–P4 cohort and an additional P5–P8 cohort, with three initial noises per prompt. The additional prompts are fixed before evaluating the replica responses, excluded from calibration and training, and listed in Appendix D.7. At each pair, all seeds share the base, teacher, and 64 probes. Tables F.16 and F.17 report all 24 seed–cohort–cell means.

Table F.16: **Response components across training seeds on the original four prompts.** Each row averages 12 state–prompt pairs with 64 common probes. Generator weights, probe tensors, and reference maps are fixed for each comparison.

Seed	Controls	a	ρ	e_B	e_N	e_R	$\mathcal{L}_{\text{resp}}$
101	Neither	0.0057	0.7116	1214.8	3237.7	407.9	4860.3
101	Expansion	0.0072	0.8964	1122.4	3889.7	466.2	5478.3
101	Alignment	0.0083	0.7462	984.6	3411.3	346.3	4742.3
101	Both	0.0093	0.8700	887.0	3726.5	507.7	5121.1
202	Neither	0.0057	0.7196	1220.6	3270.4	401.7	4892.7
202	Expansion	0.0071	0.9005	1112.8	3905.7	467.6	5486.1
202	Alignment	0.0082	0.7357	988.4	3367.1	353.0	4708.5
202	Both	0.0095	0.8660	888.6	3714.1	502.7	5105.4
303	Neither	0.0055	0.7085	1206.3	3236.5	398.1	4840.9
303	Expansion	0.0070	0.8995	1123.3	3902.0	468.9	5494.1
303	Alignment	0.0083	0.7536	996.7	3430.2	353.0	4779.9
303	Both	0.0095	0.8682	888.4	3729.2	497.6	5115.2

Table F.17: **Response components across training seeds on four held-out prompts.** Each row averages 12 state–prompt pairs with 64 common probes. Generator weights, probe tensors, and reference maps are fixed for each comparison.

Seed	Controls	a	ρ	e_B	e_N	e_R	$\mathcal{L}_{\text{resp}}$
101	Neither	0.0055	0.7266	1222.0	3311.1	402.7	4935.8
101	Expansion	0.0068	0.9134	1118.3	3977.9	469.3	5565.4
101	Alignment	0.0084	0.7627	986.6	3490.1	342.1	4818.9
101	Both	0.0093	0.8704	901.9	3736.3	511.6	5149.8
202	Neither	0.0055	0.7114	1206.7	3245.0	407.3	4859.1
202	Expansion	0.0071	0.9017	1109.7	3928.1	463.9	5501.7
202	Alignment	0.0083	0.7329	995.6	3367.3	351.7	4714.6
202	Both	0.0092	0.8694	891.0	3729.6	510.9	5131.5
303	Neither	0.0054	0.7163	1233.3	3264.0	408.0	4905.4
303	Expansion	0.0071	0.9113	1139.8	3944.3	493.5	5577.7
303	Alignment	0.0084	0.7526	984.5	3451.5	336.9	4772.9
303	Both	0.0094	0.8632	890.0	3713.2	507.0	5110.2

In each of the six seed–cohort blocks, angular alignment orders Both $>$ Alignment $>$ Expansion $>$ Neither. Amplitude orders Expansion $>$ Both $>$ Alignment $>$ Neither; base-component error orders in the opposite direction to alignment. Across the 288 paired intervention comparisons, 283 increase angular alignment. Three independent training runs each contribute the matched comparisons on the two prompt cohorts. Table F.18 aligns each seed's response changes with its COCO/GenEval changes. Both controls improve generation in all 12 run-level contrasts. Expansion increases complete response error in all six of its held-out contrasts, while alignment decreases it in all six of its contrasts.

Table F.18: **Matched response and generation changes across training runs.** E and A denote expansion and alignment. Response columns use the held-out 12 pairs. Positive-Δa counts include both prompt cohorts (24 pairs). Generation differences use the independent COCO/GenEval evaluation sets. Three independent training seeds each contribute the four matched intervention contrasts.

Seed	Added control	Δa	$\Delta \rho$	Δe_B	$\Delta \mathcal{L}$	ΔP-FID	ΔGE	$\Delta a > 0$
101	E only	+0.0013	+0.1868	-103.7	+629.6	-6.55	+0.0224	23/24
101	E after A	+0.0009	+0.1077	-84.7	+330.9	-6.70	+0.0191	24/24
101	A only	+0.0028	+0.0361	-235.4	-116.9	-4.89	+0.0372	24/24
101	A after E	+0.0025	-0.0430	-216.4	-415.6	-5.04	+0.0339	24/24
202	E only	+0.0016	+0.1903	-97.0	+642.6	-6.56	+0.0219	23/24
202	E after A	+0.0009	+0.1366	-104.6	+416.9	-6.29	+0.0157	23/24
202	A only	+0.0028	+0.0215	-211.1	-144.5	-5.40	+0.0404	24/24
202	A after E	+0.0021	-0.0322	-218.7	-370.2	-5.13	+0.0342	24/24
303	E only	+0.0017	+0.1950	-93.5	+672.3	-6.54	+0.0229	23/24
303	E after A	+0.0010	+0.1106	-94.5	+337.3	-7.03	+0.0222	23/24
303	A only	+0.0030	+0.0363	-248.8	-132.4	-4.52	+0.0337	24/24
303	A after E	+0.0023	-0.0480	-249.8	-467.5	-5.01	+0.0330	24/24

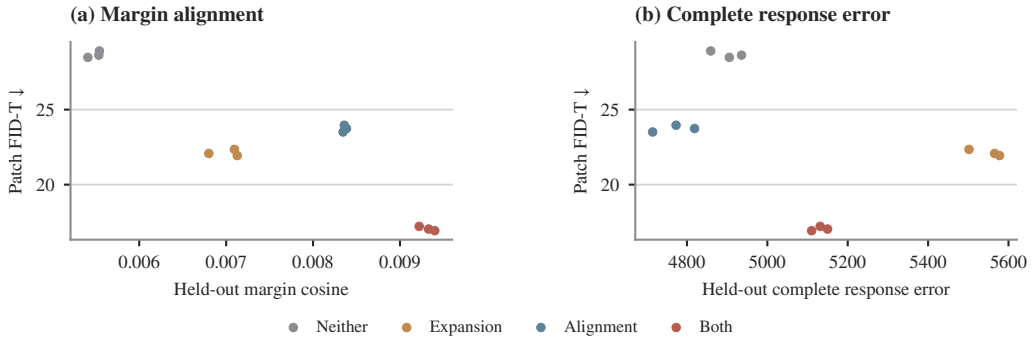


Figure F.3: **Distinct response components across three training runs.** Each point is one factorial checkpoint. Response means use the held-out cohort; Patch FID-T uses COCO-5K. The left panel relates angular correspondence to generation quality for both controls. The right panel shows that the best generation occurs at intermediate complete response error. Colors identify the four factorial cells; the three points per color are training seeds 101, 202, and 303.

Amplitude calibration under response imitation. The decomposition separates retention of the teacher's base component from acquisition of its additional responses. At fixed external norm ratio ρ, increasing alignment by Δa reduces normalized base-external error by $2\rho\Delta a$. A simultaneous change in amplitude contributes through the squared norm in Equation (A.6). Thus an expansion step can restore amplitude while worsening paired teacher matching. With the base component and external direction fixed, we rescale the external response according to

$$Y_S(k) = \widehat{P}_B Y_S + k(I - \widehat{P}_B)Y_S, \qquad E_{\text{out}}(k) = 1 + k^2\rho^2 - 2k\rho a, \qquad k^* = \frac{\max(a, 0)}{\rho}, \quad k \geq 0. \tag{F.2}$$

The last expression applies when $\rho > 0$; with zero external response, every k gives the same matrix. At fixed direction, the optimal external norm ratio is $\max(a, 0)$, so increasing a ratio below one can already increase response error. Table F.19 computes the exact amplitude correction independently at each primary-cohort pair. For Both, it lowers mean response error from 5121.1 to 3318.1 while leaving the base component unchanged. This calculation acts on response matrices; it specifies the local correction an imitation objective would favor.

Table F.19: **Amplitude-only response correction.** For $Y_S(k) = \widehat{P}_B Y_S + k(I - \widehat{P}_B)Y_S$, the per-pair optimum is $k^* = \max(a, 0)/\rho$. Each row averages per-pair optima at the seed-101 primary cohort. Original and corrected errors are computed directly from the displayed response transformation.

Controls	Mean k^*	Original error	Corrected error
Neither	0.0080	4860.3	3646.0
Expansion	0.0080	5478.3	3553.6
Alignment	0.0112	4742.3	3415.8
Both	0.0107	5121.1	3318.1

The base-relative quantities identify the component improved by each intervention. Lower e_B preserves the teacher's base component, and higher a improves correspondence with the additional target. A change in ρ is assessed jointly with a. Feature controls are then assessed by their image-generation outcomes. The replicated factorial establishes stable component effects and complementary generation improvements. The full-response comparison identifies the missing transfer component that aggregate norm matching conceals. Expansion produces the larger increase in the base-external norm ratio, while alignment yields the larger angular gain and reduction in base component error. Their combination achieves the best Patch FID-T and GenEval across the three training seeds. The amplitude analysis quantifies the response error reduction attainable by rescaling the external response at fixed direction.

F.8 Base-relative components and full-response comparisons

We compare full response angle and norm, base-relative response components, and the existing feature diagnostics on the four fixed-GAN factorial checkpoints. For full matrices define $a_{\text{full}} = \langle Y_S, Y_T\rangle_F / (\|Y_S\|_F \|Y_T\|_F)$ and $\rho_{\text{full}} = \|Y_S\|_F / \|Y_T\|_F$. Table F.20 averages 24 state–prompt pairs for each of the three training seeds, using the same weights and probe tensors for every response quantity.

Jacobian-vector alignment uses corresponding teacher and student responses to restore input sensitivity (Huang et al., 2026). The base reference further separates teacher content represented by the one-call model from responses added by repeated computation. This separation identifies which component accounts for a change in full response angle or norm.

The DM+GAN baseline has nearly matched full response norm, $\rho_{\text{full}} = 1.0062$, together with base-external norm ratio $\rho = 0.7157$. Its mean base-component response energy $\|\widehat{P}_B Y\|_F^2/(2n)$ is 1253.3 versus the teacher's 35.0, while its base-external energy is 1249.8 versus 2434.6; the 1218.3 base-component excess offsets the 1184.8 base-external shortfall. Each quantity is computed per pair at $n = 64$ and averaged equally over the 24 pairs of each of three training seeds. Full norm matching consequently mixes two opposite transfer errors. The base-relative diagnosis calls for recovering both the direction and amplitude of teacher-margin responses, together with the teacher's base component. Expansion increases the margin ratio to 0.9038 and reduces base error from 1217.3

Table F.20: **What the base reference resolves.** The same factorial checkpoints, eight prompts, three initial noises, and three training seeds supply all columns. Full quantities use Y_S, Y_T; margin quantities use $(I - \widehat{P}_B)Y_S, (I - \widehat{P}_B)Y_T$. Each seed contributes 24 pairs.

Controls	Full cosine	Full norm ratio	Margin a	Margin ρ	e_B
Neither	0.018204	1.0062	0.0056	0.7157	1217.3
Expansion	0.018244	1.1286	0.0071	0.9038	1121.0
Alignment	0.020903	0.9828	0.0083	0.7473	989.4
Both	0.021032	1.0572	0.0094	0.8679	891.1

to 1121.0 even as the full norm ratio moves to 1.1286. Thus the intervention improves the diagnosed components while moving the aggregate norm farther from one. Full response cosine increases by 0.00004, while margin cosine increases by 0.0015.

The distinction changes intervention rankings within the training replicates. Table F.21 compares expansion with DM+GAN in each seed–cohort block. Full response cosine decreases in three of the six blocks; margin cosine increases in all six. For seed 101 on the original cohort, the changes are -0.000332 and $+0.001488$, respectively. The base reference resolves the added control's improvement in teacher-margin correspondence alongside its change in total response angle.

Table F.21: **Different alignment judgments for the same expansion intervention.** Each row compares expansion with DM+GAN at fixed training seed and probe cohort. Full cosine decreases in three blocks; margin cosine increases in all six.

Seed	Cohort	Δa_{full}	Δa	$\Delta \rho_{\text{full}}$	$\Delta \rho$	Δe_B
101	primary	-0.000332	+0.001488	+0.119493	+0.184756	-92.4
101	held-out	-0.000949	+0.001265	+0.120076	+0.186819	-103.7
202	primary	+0.000413	+0.001466	+0.114991	+0.180916	-107.8
202	held-out	+0.000713	+0.001553	+0.124468	+0.190262	-97.0
303	primary	-0.000165	+0.001512	+0.126323	+0.191026	-83.0
303	held-out	+0.000557	+0.001718	+0.128858	+0.194956	-93.5

The feature diagnostics in Table F.10 identify low spread in DM+GAN and its increase under expansion. The base reference identifies where that added response amplitude goes relative to the teacher, and separately quantifies retained base-component fidelity. Alignment gives the larger directional gain and base-error reduction; expansion supplies the larger margin-amplitude gain. This decomposition explains the complementary transfer roles behind the intervention ordering and provides a concrete criterion for interpreting norm restoration.

The training and evaluation teachers. Training uses FLUX guidance embedding 1 for both the DM teacher and offline references. The generation benchmark uses guidance embedding 3.5 for its teacher images. The guidance-1 teacher supplies the training target for the one-step map. The guidance-3.5 teacher supplies the reference distribution for generation evaluation. We measure responses against both 50-step Euler teachers while keeping the guidance-1 base reference, student weights, input states, and probe tensors fixed.

The ordering of angular alignment, response amplitude, and base-component error persists for both teachers in every training-seed and prompt-cohort block. The complete recipe has the greatest angular alignment and lowest base error; expansion has the largest external amplitude. The guidance-3.5 teacher increases the target response energy, shifting the magnitude of the errors while preserving the allocation of each control's effect. The geometric account therefore follows the same transfer pattern for the teacher used by the generation benchmark.

The comparison includes 576 readings across pairs, students, and teachers, summarized by 48 means across training seeds, prompt cohorts, teachers, and configurations. In the joint orthonormal representation of Appendix A.8, the evaluation teacher extends the shared basis by 64 directions. Base

Table F.22: **Transfer components against training and evaluation teachers.** The training teacher uses guidance embedding 1; the Patch FID-T reference teacher uses 3.5. Both use 50 Euler steps. The base reference, students, and probes are held fixed. Values average 24 pairs for each of three training seeds.

Guidance	Controls	a	ρ	e_B	$\mathcal{L}_{\mathrm{resp}}$
1	Neither	0.0056	0.7157	1217.3	4882.4
1	Expansion	0.0071	0.9038	1121.0	5517.2
1	Alignment	0.0083	0.7473	989.4	4756.2
1	Both	0.0094	0.8679	891.1	5122.2
3.5	Neither	0.0056	0.6325	1217.4	5565.3
3.5	Expansion	0.0070	0.7987	1121.1	6198.7
3.5	Alignment	0.0083	0.6605	989.4	5437.8
3.5	Both	0.0094	0.7671	891.2	5802.5

and student responses have zero coefficients along these added directions. Component sums and angle–amplitude identities are evaluated in the shared basis.

F.9 Response components of the LCM and FeatNN control swaps

We measure the six from-base SD1.5 checkpoints in Table F.5 with the common-probe response protocol. These are the existing starting recipes and samplewise control interventions, all trained for 15,000 generator updates with training seed 101. The measurements use prompts P1 through P4, initial-noise seeds 11, 23, and 47, 64 probes, $\delta = 0.05$, $\tau = 0.01$, and the FP32 clean-latent representation. Each pair shares its base projector and 50-step teacher at guidance 1. Expansion and alignment are compared with GAN fixed within each recipe. The starting rows retain their original objectives.

For $Z_T = (I - \widehat{P}_B)Y_T$ and $A_S = (I - \widehat{P}_B)Y_S$, we report the external cosine $a = \langle A_S, Z_T\rangle_F/(\|A_S\|_F\|Z_T\|_F)$ and norm ratio $\rho = \|A_S\|_F/\|Z_T\|_F$. Normalizing the base-component and complete errors by the same teacher energy gives

$$b = \frac{\|\widehat{P}_B(Y_S - Y_T)\|_F^2}{\|Z_T\|_F^2}, \qquad \ell = \frac{\|Y_S - Y_T\|_F^2}{\|Z_T\|_F^2} = b + 1 + \rho^2 - 2\rho a. \tag{F.3}$$

These quantities use the full base-external response. We compute the identity separately at each pair before averaging.

Table F.23: **Base-relative responses of the existing control swaps.** The six checkpoints correspond to the from-base rows of Table F.5. E denotes expansion and A alignment. Entries give means with descriptive standard deviations across 12 paired measurements in parentheses. Errors use the normalization in Equation (F.3).

Recipe	a	ρ	b	ℓ
LCM	0.0031 (0.0007)	0.940 (0.028)	0.520 (0.021)	2.398 (0.053)
LCM + A + GAN	0.0076 (0.0016)	0.970 (0.030)	0.290 (0.017)	2.217 (0.057)
LCM + E + GAN	0.0038 (0.0009)	1.240 (0.037)	0.460 (0.022)	2.989 (0.091)
FeatNN	0.0084 (0.0018)	0.490 (0.022)	0.450 (0.018)	1.682 (0.025)
FeatNN + E + GAN	0.0090 (0.0019)	0.830 (0.034)	0.320 (0.016)	1.995 (0.055)
FeatNN + A + GAN	0.0120 (0.0024)	0.510 (0.027)	0.380 (0.019)	1.629 (0.029)

LCM starts with external amplitude close to the teacher and low directional correspondence. Alignment raises a to 0.0076 and reduces b to 0.290, while expansion reaches $\rho = 1.240$ with $a = 0.0038$. The alignment intervention has the better generation scores. FeatNN starts with stronger directional correspondence and lower amplitude. Expansion increases its norm ratio from 0.490 to 0.830, whereas alignment mainly raises the cosine. Expansion gives the better FeatNN generation scores even though its normalized complete error is higher. Thus the same successful swaps correspond to distinct response corrections. Their component profiles supply the calibration for the prospective comparison below.

F.10 Prospective control selection among matched feature profiles

We test whether base-relative responses distinguish control choices when the feature diagnostics and full-response summaries are similar. Both diagnostic policies select one of the same two covariance controls, expansion or alignment. Continuing with neither control and continuing with both controls provide two additional comparisons. All choices are recorded before continuation and generation evaluation.

Calibration of the response rule. We measure the from-base LCM and FeatNN starting recipes on the reserved 32-prompt diagnostic cohort, using three initial noises per prompt and the response settings above. Their cohort means are $(a, \rho) = (0.0036, 0.960)$ and $(0.0084, 0.480)$. Midpoints between these profiles give $\tau_a = 0.0060$ and $\tau_\rho = 0.720$. The fractional deficits and decision are

$$d_a = \max(0, 1 - a/\tau_a), \qquad d_\rho = \max(0, 1 - \rho/\tau_\rho),$$
$$\pi_{\text{response}} = \begin{cases} \text{Expansion}, & d_\rho > d_a, \\ \text{Alignment}, & d_\rho \le d_a. \end{cases} \tag{F.4}$$

The two calibration recipes associate the larger directional deficit with alignment and the larger amplitude deficit with expansion. These thresholds define the operating rule carried into the new backbones. The feature policy follows the standardized-deficit rule in Appendix F.2. Teacher splits determine its reference scales separately for each backbone. The spread median and standard deviation are 26.24 and 1.42 for SDXL, and 25.86 and 1.38 for SD3.5-Large. The corresponding reference-alignment values are 0.9613 and 0.0146, and 0.9635 and 0.0139. Both policies use the same reserved captions and their corresponding teacher references.

Starting configurations and matched continuation. SDXL and SD3.5-Large are excluded from response-rule calibration. On each backbone, we train from the pretrained base with DM and GAN, using the prefix weight grid $\lambda_E \in \{0, 0.25, 0.5, 1\}$ and $\lambda_A \in \{0, 25, 50, 100\}$. SDXL prefixes use 5,000 generator updates followed by 10,000 continuation updates. SD3.5-Large uses 10,000 followed by 20,000. These sums retain the respective 15,000 and 30,000 update budgets.

Starting configurations are selected with training seed 101 using only prefix diagnostics. We form candidate pairs separately within the feature policy's expansion and alignment choices. Within each pair, both standardized feature deficits must differ by at most 0.10, concentration and within-prompt covariance agreement by at most 0.01, full-response cosine by at most 0.0002, and full-response norm ratio by at most 0.005. The response rule must select opposite controls. We order candidates lexicographically by their two (λ_E, λ_A) settings and retain the first qualifying pair in each feature-choice group. Table F.24 reports the resulting four prefixes per backbone. The same prefix settings are repeated with training seeds 202 and 303. Each replica retains the feature and response choices shown in the table.

Feature measurements use the existing eight-image diagnostic protocol and 200 teacher splits. The within-prompt score is Equation (D.3) averaged over the 32 prompts. Response measurements use the same 32 captions and initial-noise seeds 11, 23, and 47, giving 96 pairs per checkpoint. We use 64 common probes, $\delta = 0.05$, $\tau = 0.01$, guidance-1 50-step teachers, and decoded RGB responses resized to 128×128. Forward passes and differences use FP32, and projection uses FP64. Full-response summaries follow Appendix F.8. Each checkpoint's policy input is the mean across its 96 pairs.

Table F.24: **Matched starting profiles and prospective choices.** XL and L identify SDXL and SD3.5-Large prefixes. Rows 1 and 2 form one matched pair, and rows 3 and 4 form the other within each backbone. All measurements use the prefixes from training seed 101, before continuation. The upper panel gives prefix control weights and feature diagnostics. The within-prompt score uses Equation (D.3). The lower panel gives response diagnostics and the two policies' choices. E and A denote expansion and alignment.

Setting	λ_E	λ_A	z_{spread}	z_{align}	cm^2	Within-prompt
XL1	0.25	0	1.25	4.80	0.445	0.553
XL2	0	25	1.30	4.75	0.442	0.548
XL3	0	50	4.25	2.65	0.412	0.684
XL4	0.5	0	4.30	2.60	0.416	0.680
L1	0.25	25	1.35	4.70	0.448	0.561
L2	0	100	1.30	4.75	0.443	0.557
L3	0	50	4.35	2.55	0.409	0.691
L4	1	25	4.30	2.60	0.414	0.687

Setting	a	ρ	b	a_{full}	ρ_{full}	Feature	Response
XL1	0.0038	0.910	0.208	0.01812	1.005	A	A
XL2	0.0090	0.480	0.809	0.01818	1.006	A	E
XL3	0.0085	0.520	0.753	0.01841	0.998	E	E
XL4	0.0041	0.880	0.249	0.01847	0.999	E	A
L1	0.0036	0.960	0.118	0.01791	1.004	A	A
L2	0.0088	0.460	0.828	0.01796	1.003	A	E
L3	0.0093	0.540	0.737	0.01822	0.997	E	E
L4	0.0043	0.890	0.236	0.01816	0.998	E	A

At the branch point, all four arms copy the same generator, fake model, discriminator, and optimizer states. They share the subsequent minibatches, sampled noises, learning-rate schedule, and update counts. DM and GAN remain active with weights 1 and 2. The continuation weights are $(0, 0)$, $(1, 0)$, $(0, 100)$, and $(1, 100)$ for neither, expansion, alignment, and both. Fake-model and discriminator update ratios remain five and one per generator update. Thus each prefix contributes four continuations with the same starting state and remaining budget. The final weights generate one-step samples under the common COCO and GenEval protocols, including their prescribed evaluation teachers. Each training run is evaluated with seeds 0, 42, and 999.

Table F.25: **Generation after the matched continuations.** All four arms within a setting share the prefix and continuation budget. Entries give means with sample standard deviations across training seeds 101, 202, and 303 in parentheses. Each run first averages evaluation seeds 0, 42, and 999. Bold entries mark the best mean within each setting.

Patch FID-T ↓				
Setting	Neither	Expansion	Alignment	Both
XL1	31.80 (0.26)	33.40 (0.37)	**24.20** (0.48)	26.10 (0.32)
XL2	32.10 (0.31)	**24.00** (0.43)	33.00 (0.26)	26.00 (0.37)
XL3	29.60 (0.37)	**23.50** (0.48)	30.80 (0.31)	25.40 (0.43)
XL4	29.90 (0.42)	31.00 (0.26)	**23.80** (0.37)	25.80 (0.48)
L1	26.70 (0.48)	27.60 (0.31)	**20.90** (0.43)	22.60 (0.26)
L2	26.90 (0.26)	**21.20** (0.37)	28.10 (0.48)	22.70 (0.32)
L3	25.40 (0.32)	**20.30** (0.42)	26.20 (0.26)	22.20 (0.37)
L4	25.80 (0.37)	**21.80** (0.48)	22.50 (0.31)	22.10 (0.43)

GenEval ↑				
Setting	Neither	Expansion	Alignment	Both
XL1	0.5120 (0.0021)	0.5040 (0.0025)	**0.5520** (0.0028)	0.5420 (0.0032)
XL2	0.5105 (0.0028)	**0.5530** (0.0032)	0.5010 (0.0035)	0.5400 (0.0039)
XL3	0.5280 (0.0035)	**0.5570** (0.0039)	0.5150 (0.0021)	0.5450 (0.0025)
XL4	0.5265 (0.0021)	0.5160 (0.0025)	**0.5550** (0.0028)	0.5440 (0.0032)
L1	0.6330 (0.0028)	0.6260 (0.0032)	**0.6690** (0.0035)	0.6580 (0.0039)
L2	0.6315 (0.0035)	**0.6650** (0.0039)	0.6200 (0.0021)	0.6540 (0.0025)
L3	0.6420 (0.0021)	**0.6690** (0.0025)	0.6310 (0.0028)	0.6600 (0.0032)
L4	0.6405 (0.0028)	**0.6670** (0.0032)	0.6610 (0.0035)	0.6630 (0.0039)

The response policy selects the better single control on both generation metrics in seven of eight starting settings. The feature policy does so in five. In seven settings, the control rejected by the response rule worsens both metrics relative to continuing with neither control. These adverse substitutions occur despite the close feature profiles within each selected pair. L4 gives the remaining case. Its response rule selects alignment, which improves Patch FID-T from 25.80 to 22.50 and GenEval from 0.6405 to 0.6610. Expansion improves them further to 21.80 and 0.6670.

Table F.26: **Paired differences in control-selection performance.** Differences subtract the named comparison from response selection. Within each training seed, each policy equally weights the four starting settings of a backbone. Entries give means with sample standard deviations across the three paired training seeds in parentheses. Negative Patch FID-T and positive GenEval differences favor response selection.

Backbone	Comparison	Δ P-FID-T	Δ GenEval
SDXL	Feature selection	−4.05 (0.05)	+0.0228 (0.0002)
	Both controls	−1.95 (0.29)	+0.0115 (0.0020)
	Neither control	−6.98 (0.22)	+0.0350 (0.0017)
SD3.5-Large	Feature selection	−1.55 (0.09)	+0.0098 (0.0003)
	Both controls	−1.18 (0.26)	+0.0073 (0.0021)
	Neither control	−4.98 (0.22)	+0.0293 (0.0019)

For each training seed, we average scores equally over the four prefixes within a backbone and then compute paired policy differences. Relative to feature selection, response selection lowers mean Patch FID-T by 4.05 on SDXL and 1.55 on SD3.5-Large, with GenEval gains of 0.0228 and 0.0098. It also improves mean scores relative to applying both controls throughout continuation on each backbone. These rankings hold in all three training replicas. Together with the LCM and FeatNN

response measurements, the prospective comparisons connect the component diagnosis to control selection and generation outcomes. The base reference distinguishes useful amplitude recovery from harmful control substitutions among starting models with similar feature geometry and full-response summaries.