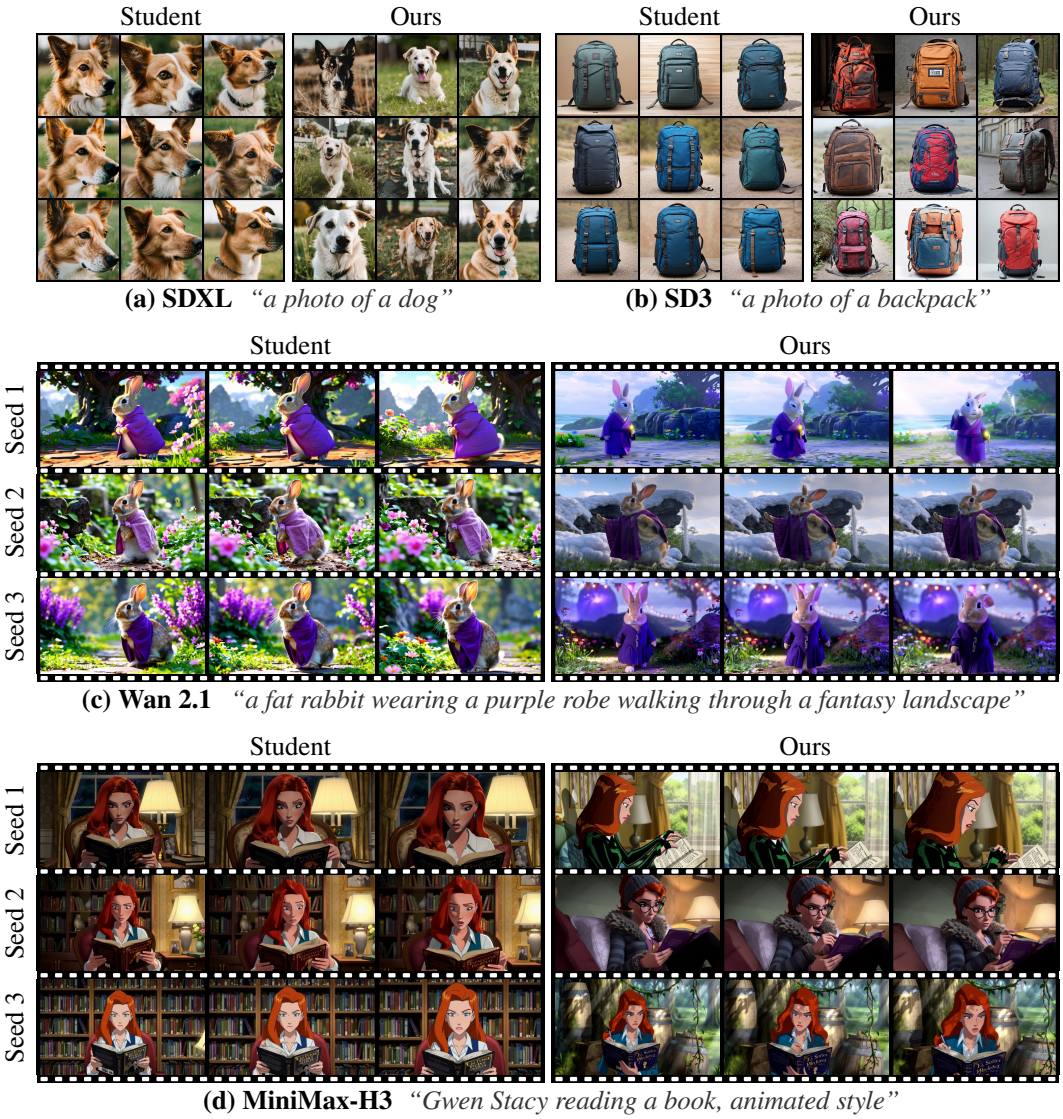
RESTORING DIVERSITY IN DISTILLED IMAGE AND VIDEO GENERATORS BY PROBING ACTIVATIONS

Anonymous authors
Paper under double-blind review



(a) SDXL *"a photo of a dog"* **(b) SD3** *"a photo of a backpack"*

(c) Wan 2.1 *"a fat rabbit wearing a purple robe walking through a fantasy landscape"*

(d) MiniMax-H3 *"Gwen Stacy reading a book, animated style"*

Figure 1: **ReDiverse improves diversity in distilled image and video generators.** For a fixed prompt, the student tends to repeat similar visual compositions, while ReDiverse produces broader variation: (a) dogs appear at different viewing distances and poses (SDXL distilled model); (b) backpacks in various styles (SD3 distilled model); and (c–d) videos show diverse environments, viewpoints, and styles (distilled from Wan 2.1 and MiniMax-H3). Each image grid shows samples from different seeds; each video row corresponds to one seed, with time progressing from left to right.

ABSTRACT

Distilling diffusion models into few-step students substantially reduces sampling latency, but induces severe diversity collapse: samples generated from different random seeds show minimal variation. In this work, we investigate the root cause

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.

of this failure by probing the internal representations of student models relative to their teachers. We find that distilled students maintain total activation variance comparable to their teachers, but concentrate it into substantially fewer directions. This directional collapse is localized to an identifiable intermediate block at the initial denoising step. Consequently, diversity collapse stems from a misallocation of activation variance rather than an inherent loss of generative capacity. Guided by this finding, we propose ReDiverse to reinstate the missing variance directions. At training time, ReDiverse-FT fine-tunes a lightweight adapter that regresses student activations onto the teacher's, reviving the teacher's feature variance characteristic. At inference time, ReDiverse-TF introduces a training-free mechanism that steers activations along precomputed teacher principal directions. Across image and video distillation benchmarks, ReDiverse consistently restores sample diversity while preserving generation quality, further supported by the user study. Anonymous project page: `https://iclr-2027-14418.github.io/`.

1 INTRODUCTION

Distillation turns diffusion-based image and video generators (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022) into fast, few-step models (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024b;a). But they can reduce the diversity of their outputs: different seeds may produce strikingly similar samples for the same prompt. We explore where in the student diversity disappears.

Broadly, there are two solutions to alleviate this problem. (1) Retrain or re-distill the student, either with a modified or regularized objective (Xu et al., 2025; Lu et al., 2025; Wu et al., 2026; Li et al., 2026b) or with real data (Chen et al., 2026). (2) Keep the student frozen and intervene at inference, either by optimizing the initial noise (Harrington et al., 2026), by perturbing intermediate features directly in the forward pass (Yadav et al., 2026; Bhat et al., 2026), or by running the teacher for the first step (Gandikota & Bau, 2026). These approaches motivate a complementary question: where does the student's internal representation cease to preserve the variation present in its teacher, and can that diagnosis guide a targeted repair?

We start by asking a simple question. What would happen to the activations across many seeds when the same prompt is given to a teacher and its student? Surprisingly, we found a similar pattern emerging in every model we tested, across both image and video families. In particular, at each block, we measure the total activation variance across seeds and its participation ratio, which summarizes how broadly that variance is distributed across directions. We find that in the initial blocks, the student's total variance and participation ratio closely match the teacher's. However, deeper down, the student's total variance equals or exceeds the teacher's, but it is restricted to far fewer directions in a particular block, resulting in a cascaded effect on downstream activations. Interestingly, we show that the collapse in diversity can be attributed to these lost directions and misallocation of variance within a single block of the student's model.

This diagnosis identifies where to supervise the student to recover a richer representation. We introduce **ReDiverse** with two complementary interventions. In **ReDiverse-FT**, we insert LoRA modules into all layers up to and including the identified block, keeping the base model frozen. We train these modules by regressing the student's activation at this block during its first denoising step toward the teacher's activation at the corresponding block during its τ-th denoising step. At inference, the LoRA modules are active only during the student's first step; subsequent steps use the original student. In **ReDiverse-TF**, no training is required: we perturb activations at the identified block along teacher principal directions computed once per prompt, with random coefficients scaled by the teacher's spread along each direction. The two interventions can be used independently or combined.

To summarize, our contributions are threefold:

- We profile and probe the variance characteristics of distilled students against their teachers, block by block under identical inputs, and show that distillation keeps the total variance but concentrates it into fewer directions at an identifiable block. This locates the collapse, which output metrics cannot, and separates students distilled from the same teacher under different recipes.
- We introduce ReDiverse to restore variation at the identified block through two complementary interventions: ReDiverse-FT trains LoRA modules in the layers up to and including the identified

block to align first-step student activations with later-step teacher activations, while ReDiverse-TF perturbs activations along teacher principal directions without training.

- We evaluate both interventions and their combination on four teacher–student pairs spanning image (SDXL (Podell et al., 2024), SD3 (Esser et al., 2024)) and video (Wan 2.1 (Wan et al., 2025), MiniMax-H3 (MiniMax, 2026)) generators, where ReDiverse restores sample diversity while preserving generation quality, further supported by a user study.

2 Related Work

Distillation and Diversity. Distillation compresses iterative diffusion models (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Podell et al., 2024) into few-step students by matching teacher trajectories (Salimans & Ho, 2022), consistency endpoints (Song et al., 2023; Luo et al., 2024), or output distributions via adversarial (Sauer et al., 2024) and distribution-matching (DMD) losses (Wang et al., 2023a; Yin et al., 2024b;a; Luo et al., 2025), recently extended to video transformers (Huang et al., 2026; Zheng et al., 2026b;a; Shaul et al., 2026; FastVideo Team, 2026; lightx2v, 2026). Because these objectives supervise terminal outputs rather than seed-conditioned activations, distilled students prematurely freeze global structure at the first step (Gandikota & Bau, 2026) and generate near-identical outputs across seeds (Shi et al., 2026; Li et al., 2026b). This collapse has been attributed to mode-seeking reverse KL (Minka et al., 2005; Xu et al., 2025; Lu et al., 2025), adversarial training (Arora et al., 2017), constrained step budgets (Salmona et al., 2022; Fan et al., 2026), or over-guided high-noise targets (Li et al., 2026a). Existing remedies either retrain the student with modified divergences (Xu et al., 2025; Lu et al., 2025), consistency–DMD hybrids (Zheng et al., 2026b; Li et al., 2026b), adjusted first-step targets (Wu et al., 2026; Li et al., 2026a), or real data (Chen et al., 2026), incurring substantial re-distillation costs and fidelity trade-offs (Li et al., 2026b); or intervene at inference on a frozen student via noise optimization (Harrington et al., 2026; Shi et al., 2026), teacher–student hybrid sampling (Gandikota & Bau, 2026), or feature perturbations (Bhat et al., 2026; Yadav et al., 2026). Crucially, neither paradigm localizes where diversity vanishes internally. ReDiverse identifies this breakdown and is agnostic to the distillation recipe.

Interpreting diffusion representations. Prior work has shown that the intermediate activations of a diffusion model encode attributes of the samples it generates: the U-Net bottleneck forms a semantic space (Kwon et al., 2023), whose editing directions can be found from principal components of these features (Haas et al., 2024), the encoder's pullback metric (Park et al., 2023), or the low-rank Jacobian of the posterior mean (Chen et al., 2024); all study a single model. In distilled students, Gandikota & Bau (2026) find that the directions needed for diversity survive but go unused, and Yadav et al. (2026) inject noise along the student's own principal directions, spreading variance it still has. However, we are the first to interpret student and teacher activations together, both to explain why diversity collapses and to remedy it.

3 Probing Distilled Diffusion Models

To identify the internal mechanisms driving diversity collapse, we analyze the representations of distilled diffusion models across network depth and denoising steps. Specifically, we trace the propagation of seed sensitivity by comparing intermediate activations between teacher models and their few-step student counterparts under identical initial latents and conditioning prompts.

3.1 Same Variance, Fewer Directions

To characterize the internal representation of block b, we evaluate three geometric quantities derived from the sample covariance matrix Σ_b across random noise initializations: total variance (TV) $\mathrm{Tr}(\Sigma_b)$, measuring overall dispersion; the participation ratio $\mathrm{PR}(\Sigma_b)$ (Gao et al., 2017), quantifying the effective dimensionality spanned by this variance; and the concentration ratio $\mathrm{Tr}(\Sigma_b)/\mathrm{PR}(\Sigma_b)$, measuring the average variance allocated per active dimension.

We find that diversity collapse is driven by a directional bottleneck rather than a loss of variance: distilled students preserve total activation variance, but confine it to fewer directions starting at an intermediate layer (Fig. 2). Across all depths, the student's total variance $\mathrm{Tr}(\Sigma_b)$ matches or exceeds that of the teacher at most blocks, confirming that overall variance is not depleted. While early layers track the teacher, the participation ratio $\mathrm{PR}(\Sigma_b)$ drops at an intermediate layer—the *collapsed block*.

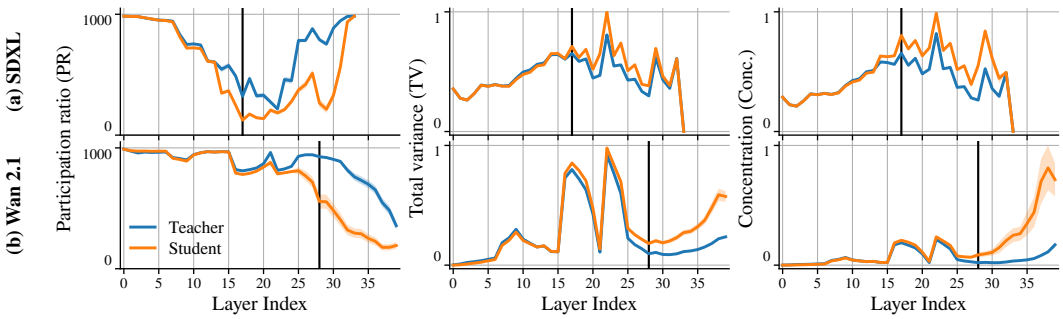


Figure 2: **Variance characteristic along the depth.** Participation ratio, total variance and concentration of teacher and student activations at every layer, for (a) an image model (SDXL) and (b) a video model (Wan 2.1). Total variance and concentration are min–max normalized to $[0, 1]$ over teacher and student (SDXL's in log space). The solid vertical line marks the collapsed block.

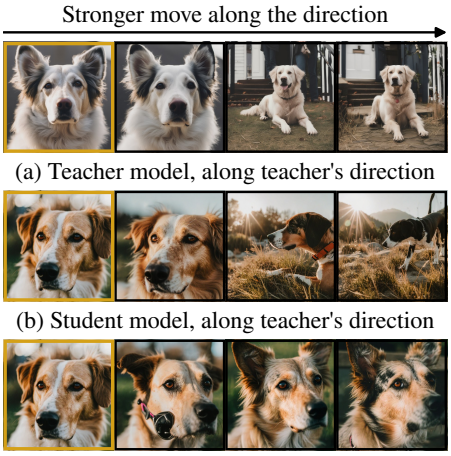
Tab. 1 divides each student's value at the collapsed block by its teacher's, showing how much the student amplifies or shrinks each quantity, for every model. Across all evaluated architectures and modalities, students preserve or even amplify total variance ($1.03\times$–$2.92\times$) while retaining merely 37%–60% of the teacher's directional span. Consequently, the surviving axes are overloaded with up to $8\times$ the variance observed in the teacher, driving a sharp surge in concentration that propagates through all downstream layers to the final output. While the exact layer index depends on network topology, this directional compression occurs universally, identifying it as an intrinsic hallmark of few-step distillation.

Table 1: **Student/teacher ratios.** The student's value divided by the teacher's at the collapsed block.

Student	PR	TV	Conc.
DMD2 (SDXL)	0.40	1.63	4.76
TDM (SD3)	0.49	1.03	2.68
lightx2v (Wan)	0.60	1.41	2.54
Turbo (H3)	0.37	2.92	8.08

3.2 Fewer Directions, Less Diversity

Because downstream layers deterministically map intermediate features to the final output, the directional span of activation variance bounds attainable generative diversity. To test whether the directions lost during distillation govern macroscopic diversity and whether the student can still decode them, we perform activation steering at the collapsed block during the initial denoising step. Specifically, we extract the top principal direction of activation covariance across random noise initializations and perturb intermediate activations along it with increasing strength α, scaled by the empirical standard deviation.



(a) Teacher model, along teacher's direction

(b) Student model, along teacher's direction

(c) Student model, along student's direction

Figure 3: **Moving along a direction, in the teacher and in the student model.**

Perturbing the teacher along its own principal axis dynamically alters macroscopic attributes, including camera framing, subject pose, and scene composition (Fig. 3a). Crucially, injecting this same teacher direction into the student produces matching structural changes (Fig. 3b), recovering diverse poses and layouts without modifying any student weights. In contrast, steering the student along its own leading axis alters only appearance-level details, such as coat pattern and localized lighting, leaving global geometry invariant (Fig. 3c). This functional asymmetry confirms two key insights: (i) the directions suppressed during distillation are precisely those that dictate structural diversity, and (ii) downstream student layers retain the capacity to decode these modes once restored. Diversity collapse is therefore not an irreversible loss of generative capacity, but a geometric misallocation of activation variance.

In addition, we found that two students distilled from the same teacher (MiniMax-H3 (MiniMax, 2026)) reach different output diversity, and that their participation ratios at the collapsed block differ in the same order (Fig. 4). Both track the teacher until just before the collapsed block (block 45), where FastVideo (FastVideo Team, 2026) keeps 61% of the teacher's participation ratio and Turbo (lightx2v, 2026) only 28%, and FastVideo is correspondingly more diverse (DINO 52.3 vs. 35.8; teacher 55.3). The retained participation ratio at the collapsed block thus tracks output diversity.

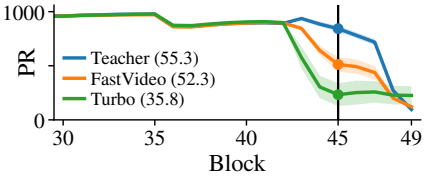


Figure 4: **PR along depth.** MiniMax-H3 and two students (DINO diversity score in parentheses); solid line: collapsed block.

4 REDIVERSE: RESTORING THE LOST DIRECTIONS

We introduce ReDiverse, which operationalizes our diagnostic findings through two complementary mechanisms: (i) ReDiverse-FT (Sec. 4.1), a targeted fine-tuning framework that regresses student activations onto the teacher's at the collapsed block to revive the teacher's variance characteristic, and (ii) ReDiverse-TF (Sec. 4.2), a training-free inference approach that injects variance along precomputed teacher principal directions to restore the directions the student lost; both hence improve diversity by adjusting the student's feature variance characteristic.

4.1 REDIVERSE-FT: DIRECTIONAL ALIGNMENT VIA TARGETED ADAPTATION

Feature regression. To align the student's feature variance characteristic with the teacher's, ReDiverse-FT regresses the student's activation at the collapsed block ℓ onto that of the frozen teacher. Let $h_{\ell,t}$ denote the activation at block ℓ during denoising step t. Both models start from the same initial noise ϵ and prompt c: the student is read at its first step (t=1), where it forms the layout of the sample from pure noise, and the teacher is run for $\tau-1$ steps and read at step τ:

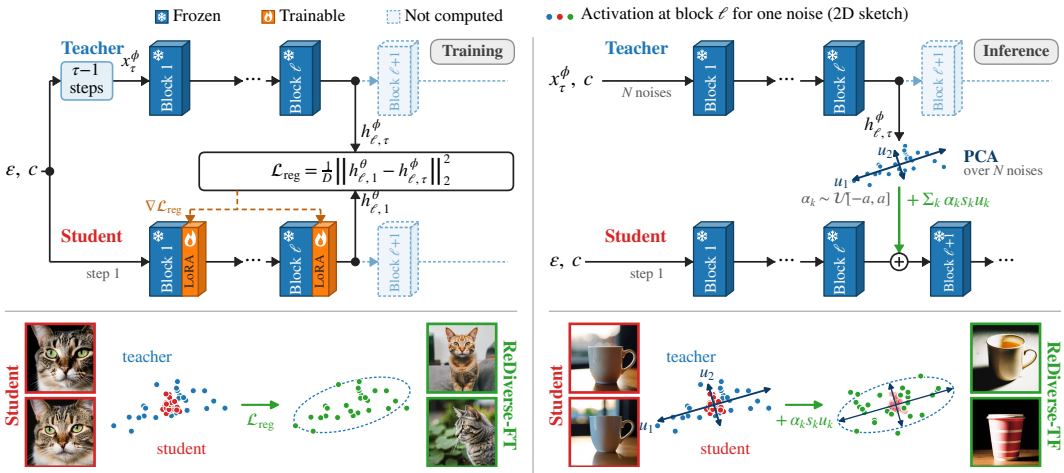
$$\mathcal{L}_{\text{reg}}(\theta) \;=\; \mathbb{E}_{\epsilon,\,c}\left[\frac{1}{D}\left\|h^{\theta}_{\ell,1}(\epsilon,c) - h^{\phi}_{\ell,\tau}(x^{\phi}_{\tau},c)\right\|_2^2\right], \quad \text{where } x^{\phi}_{\tau} = \mathcal{S}^{\phi}_{\tau-1}(\epsilon,c). \tag{1}$$

Here, $\mathcal{S}^{\phi}_{\tau-1}$ runs $\tau-1$ teacher steps from ϵ, and D is the dimension of the activation, so the loss is the mean squared error over its elements. Each prompt is paired with M noises, so matching the activations noise by noise matches the teacher's variation across noises rather than a single activation. This aligns the orientation and relative energy of the feature manifold, natively restoring the suppressed eigenspectrum without requiring auxiliary covariance regularizers or adversarial objectives. The teacher step τ, the teacher's guidance (Ho & Salimans, 2021) up to that step, the training prompts and M are design choices (Apps. E.2 and E.3, Sec. 5.1); by default the teacher is guided up to step τ and read in its conditional branch at step τ.

Localized LoRA adaptation. We implement ReDiverse-FT using LoRA (Hu et al., 2022) inserted only into the blocks up to and including the collapsed block ℓ: as these blocks map the initial noise into intermediate representations, adapting only this region enables the student to reproduce the teacher's variance characteristic before activations propagate downstream. Downstream layers already retain the capacity to decode diverse structural modes (Sec. 3.2); leaving them frozen preserves native synthesis fidelity without disrupting the rendering pipeline. Because restoring the span redistributes variance across suppressed axes rather than learning new concepts, low-rank adaptation provides an inductive bias that recovers diversity without overfitting (Gan & Isola, 2026).

4.2 REDIVERSE-TF: TRAINING-FREE STOCHASTIC STEERING

Because downstream layers can decode the teacher's principal directions without parameter updates (Sec. 3.2), we can restore diversity directly at test time by steering intermediate activations along these missing axes. To identify the exact directions to inject and calibrate their energy so that activations remain in distribution, we extract this directional subspace with principal component analysis (Bishop & Nasrabadi, 2006) on the teacher, once per prompt before sampling. Specifically, for a given prompt, we run the teacher for $\tau-1$ steps across N random noise initializations and record its centered activations at the collapsed block ℓ during step τ, matching the regression target of Eq. (1). Arranging these activations into a matrix $X \in \mathbb{R}^{N\times D}$, the eigendecomposition of the



(a) ReDiverse-FT ("a photo of a cat") (b) ReDiverse-TF ("a photo of a cup")

Figure 5: **ReDiverse overview.** The student loses the teacher's variation across noises at the collapsed block ℓ, while the later blocks can still render it; ReDiverse restores it there, by training (a) or at inference (b). Below each diagram, the activations (blue: teacher, red: student, green: ReDiverse) and samples, before and after. (a) ReDiverse-FT regresses the student's activation onto the teacher's for the same noise, updating only LoRA adapters. (b) ReDiverse-TF finds the teacher's directions of variation by PCA and moves the student along them.

sample Gram matrix $XX^\top$ yields eigenvalues λ_k and eigenvectors v_k. The k-th principal direction u_k and the teacher's empirical spread s_k along it are given by:

$$u_k = \frac{X^\top v_k}{\|X^\top v_k\|_2}, \qquad s_k = \sqrt{\frac{\lambda_k}{N}}. \tag{2}$$

Here, the unit vector u_k defines the orientation of structural variation, while the standard deviation s_k provides a principled scaling factor matching the teacher's natural dispersion.

During inference, we perturb the student's intermediate representation using these precomputed directions to induce diversity without modifying model weights. For each generated sample i, we independently draw scaling coefficients $\alpha_{i,k} \sim \mathcal{U}[-a, a]$ across the top-K principal directions and inject this calibrated variance directly into the student's activation at the collapsed block, $\tilde{h}^{(i)}_{\ell,1} = h^{(i)}_{\ell,1} + \sum_{k=1}^{K} \alpha_{i,k}\, s_k\, u_k$.

Sampling $\alpha_{i,k}$ independently per generation disperses sampling trajectories into distinct structural modes. We restrict this perturbation strictly to the first denoising step: early steps govern global layout and camera perspective, whereas subsequent steps refine fine-grained textures. Gating the intervention to step 1 steers structural mode selection while allowing subsequent steps to render outputs with native synthesis fidelity. ReDiverse-TF introduces no parameter updates and can be applied either to the base student or atop ReDiverse-FT; the teacher is needed only once per prompt, to estimate the directions that all samples of that prompt share.

5 Experiments

5.1 Evaluation Setup

Implementation Details. Our evaluation spans four teacher–student pairs across two modalities, with U-Net (Ronneberger et al., 2015) and DiT (Peebles & Xie, 2023) backbones (Tab. 2, which also gives each pair's collapsed block ℓ and teacher step τ); all students are the publicly released 4-step checkpoints, none re-distilled. *ReDiverse-FT* trains LoRA adapters (Hu et al., 2022) on DiffusionDB

Table 2: **Teacher–student design choices.**

	SDXL	SD3-M	Wan 2.1	H3
Teacher	2.6B UNet	2B MMDiT	14B DiT	33B DiT
Steps / CFG	50 / 8	28 / 7	50 / 5	50 / –
Student (4 steps)	DMD2	TDM	lightx2v	Turbo
Resolution	1024^2	1024^2	832×480	1344×768
Frames	–	–	81	124
ℓ / τ	17 / 7	12 / 9	28 / 8	45 / 16
FT: LoRA	rank 32, attention + FFN, blocks $\leq \ell$			
FT: data	100 captions $\times$ M=16 noises			
FT: training	200 iterations, batch 32			
TF: directions	top K=8 from N=25 teacher runs, a=2			

Table 4: **Main comparison results.** Diversity and quality of distilled image and video generators. **Bold** and underline mark the highest and second-highest value below the dashed line. All metrics are higher-is-better. DS: DreamSim; LP: LPIPS; TP: TPIPS averaged over its aspects; CS: CLIPScore; HPS: HPSv2; IR: ImageReward; VJ: V-JEPA 2; Cam: camera-trajectory diversity; VC: ViCLIP diversity; SC, BC: VBench subject and background consistency; VCS: ViCLIP text–video similarity.

(a) Distilled image models.

Method	Diversity					Quality		
	DINO	DS	LP	TP	Vendi	CS	HPS	IR
SDXL								
Teacher	66.3	36.0	65.9	24.1	7.66	32.9	29.2	0.695
Noise Opt.	67.0	37.9	71.2	25.1	7.83	32.7	29.7	0.714
Student	59.4	27.6	65.2	19.0	5.02	33.0	**30.4**	**0.854**
DP-DMD	58.3	22.7	64.1	17.6	4.32	32.6	30.2	0.709
ReDiverse-TF	60.1	28.4	65.6	19.6	5.30	**33.0**	30.3	0.836
ReDiverse-FT	64.6	32.9	67.8	22.6	6.74	32.8	29.8	0.694
ReDiverse	**65.0**	**33.3**	**67.9**	**23.0**	**7.03**	32.7	29.7	0.672
SD3								
Teacher	61.6	30.3	67.0	21.1	6.02	33.4	29.5	1.017
Noise Opt.	56.9	26.4	61.2	18.3	6.74	33.1	29.8	0.967
Student	48.4	19.2	55.1	13.8	3.78	33.3	30.1	**1.026**
DP-DMD	56.8	22.4	62.2	17.4	4.46	32.5	**30.5**	0.993
ReDiverse-TF	53.8	23.5	60.7	17.0	4.86	33.1	29.7	0.965
ReDiverse-FT	53.5	22.7	58.0	15.3	4.66	**33.5**	29.2	0.992
ReDiverse	**60.1**	**28.3**	**64.9**	**20.3**	**6.24**	32.9	29.0	0.872

(b) Distilled video models.

Method	Diversity					Quality		
	DINO	VJ	TP	Cam	VC	SC	BC	VCS
Wan 2.1								
Teacher	51.6	20.9	31.3	36.2	36.2	94.7	97.3	21.4
Student	31.6	11.2	23.2	36.2	22.6	96.7	95.6	21.5
DFD	26.2	10.0	17.8	25.1	18.5	**97.4**	96.4	**22.0**
ReDiverse-TF	34.6	12.5	26.3	35.8	26.1	96.3	95.2	21.5
ReDiverse-FT	46.8	17.0	28.7	**40.6**	34.7	96.4	**96.4**	20.8
ReDiverse	**52.2**	**18.7**	**32.6**	40.0	**40.0**	95.0	95.9	20.5
MiniMax-H3								
Teacher	52.5	20.7	30.4	37.7	37.2	92.8	94.6	20.7
Student	36.5	12.4	23.0	26.9	28.4	94.0	95.1	19.9
DFD	35.6	12.3	22.2	26.7	27.2	94.3	**95.8**	20.1
ReDiverse-TF	42.2	13.4	26.4	29.4	33.6	93.4	94.1	19.8
ReDiverse-FT	44.1	15.9	26.2	33.2	32.2	**94.8**	95.5	**20.1**
ReDiverse	**50.9**	**16.2**	**29.9**	**35.6**	**38.6**	94.2	94.0	19.8

captions (Wang et al., 2023b) with one recipe for all four pairs and is active only at the first denoising step; *ReDiverse-TF* estimates the teacher directions of Eq. (2) per prompt and injects them at the first step. Their settings are in Tab. 2, details in App. D.

Baselines. Besides the teacher and student models, we compare to (i) for images, Noise Optimization (Harrington et al., 2026), which searches for initial noises that maximize a diversity score at test time, and DP-DMD (Wu et al., 2026), which distills a new student whose first step is regressed toward the teacher's latent at an earlier point of its trajectory, and thus needs a full distillation run; and (ii) for videos, DFD (Chen et al., 2026), which post-trains the student on real videos and thus needs real data and further training (reproduction details in App. G). Since the teacher is the original multi-step model and Noise Optimization directly maximizes the diversity score, we report both separately as upper bounds, whereas ReDiverse needs only a short fine-tuning (ReDiverse-FT), and ReDiverse-TF is training-free, using the teacher's directions without any optimization.

Table 3: **Evaluation data.**

	Images	Video
Prompts	GenEval (553)	VBench (300)
Seeds / prompt	100	10

Dataset and Evaluation Metrics. Tab. 3 lists the prompts, from GenEval (Ghosh et al., 2023) and VBench (Huang et al., 2024); all arms of a backbone share one noise bank, so every comparison is paired. A *diversity* scorer takes all samples of a prompt and returns the mean pairwise distance between their embeddings, one minus cosine similarity (DINO (Caron et al., 2021; Oquab et al., 2024), DreamSim (Fu et al., 2023), TPIPS(Wang et al., 2026), V-JEPA 2 (Assran et al., 2025), ViCLIP(Wang et al., 2024)) or a perceptual distance (LPIPS (Zhang et al., 2018)); video frames are also compared across clips frame by frame, Vendi (Friedman & Dieng, 2023) returns the effective number of distinct samples, and camera-trajectory diversity compares estimated camera paths (Chen et al., 2026). A *quality or alignment* scorer takes one sample and its prompt and returns a CLIP image–text cosine (CLIPScore (Radford et al., 2021)), a learned preference reward (HPSv2 (Wu et al., 2023), ImageReward (Xu et al., 2023)) or a VBench quality dimension. Diversity is averaged over prompts, quality and alignment over samples (App. F).

5.2 Results and Analysis

ReDiverse improves diversity on images. Tab. 4a compares the 4-step methods on SDXL and SD3. Both students lose substantial diversity relative to their teachers, while ReDiverse recovers most of this gap and achieves the highest diversity among 4-step methods on every metric for both backbones. ReDiverse-FT and ReDiverse-TF each recover part of the gap, with their combination

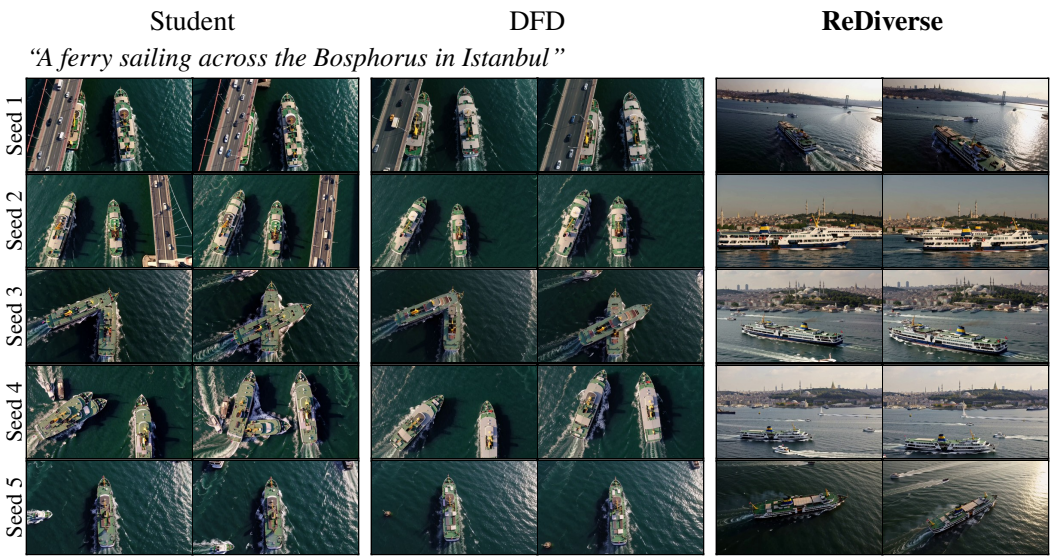


Figure 7: **Qualitative results on MiniMax-H3.** Rows are seeds, each clip shown by its first and last frame, same noise for all three columns. Best viewed as videos available in supplementary material.

providing the largest gain. ReDiverse even exceeds the teacher on LPIPS for SDXL and Vendi for SD3, and surpasses Noise Optimization on all diversity metrics except Vendi for SD3. DP-DMD falls below the student on SDXL and provides smaller gains than ReDiverse on SD3. CLIPScore changes by less than 0.5 points from the student, while the reward scores remain close to the teacher on SDXL. Qualitative results are shown in Fig. 1 and App. I.2, and Fig. 6 shows the diversity within individual images.

ReDiverse improves diversity on video. Tab. 4b compares the 4-step methods on Wan 2.1 and MiniMax-H3. Both students show a large drop in diversity relative to their teachers, and ReDiverse recovers most of this gap. On Wan 2.1, ReDiverse matches or exceeds the teacher on all diversity metrics except V-JEPA 2, while on MiniMax-H3 it approaches the teacher and surpasses it on Vi-CLIP diversity. The two students also exhibit different diversity failures: the Wan 2.1 student largely preserves camera-trajectory diversity but loses content diversity, whereas the H3 student loses both. ReDiverse recovers these missing variations in both cases, without explicitly targeting any particular type of variation, by matching the teacher's activation structure. In contrast, DFD reduces diversity on Wan 2.1 and provides little change on MiniMax-H3. This diversity recovery does not substantially affect quality: subject and background consistency remain close to the teacher, while text–video similarity stays within one point of the student. Seed-matched examples are shown in Fig. 1 and Fig. 7, with additional results for both video backbones in App. I.1.

Diversity within an image. Diversity collapse also occurs within individual images: when a prompt requires multiple instances of the same object, the student often represents them with nearly the same appearance. The directional collapse explains this behavior: with fewer available activation directions, different spatial positions have less freedom to represent distinct instances and tend to fall into the same mode. We tested this on "a photo of two vases" and found that ReDiverse restores distinct appearances for the two vases while preserving their count, layout, and framing (Fig. 6). Thus, restoring the missing directions improves diversity not only across samples, but also between instances within a single image.



Figure 6: **Diversity within an image.** The SD3 student renders the same vase twice, whereas ReDiverse increases the diversity while keeping the same count. Same column, same noise.

User and VLM-as-a-Judge Study. Automatic metrics provide an indirect measure of diversity, so we also evaluate whether people actually perceive the samples as more diverse. This is particularly important for videos, where diversity also involves differences in motion. We therefore ask raters

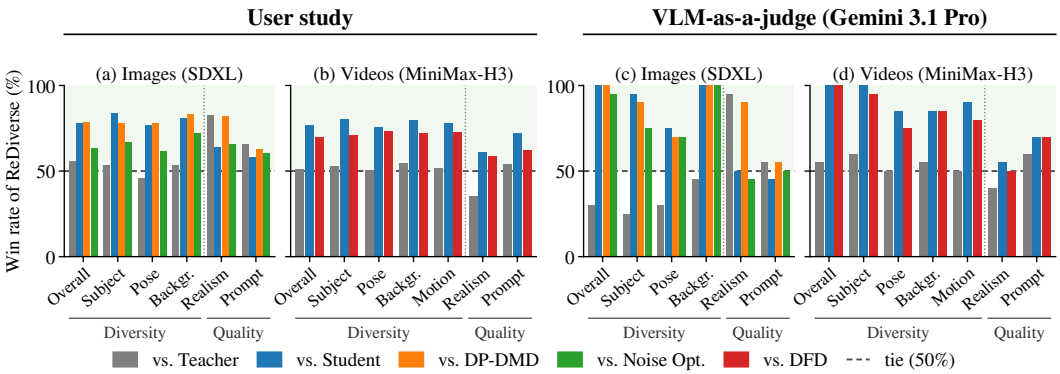


Figure 8: **User and VLM-as-a-judge study.** Win rate of ReDiverse against each method on diversity (overall and per aspect: subject, pose, background, and motion for videos), realism and prompt following; a tie counts half, and above the dashed 50% line ReDiverse is preferred.

Table 5: **Ablation on SD3.** Sand rows: defaults; small numbers: change from the default.

(a) ReDiverse-FT: where and when to regress.

Setting	DINO↑	TPIPS$_{avg}$ ↑	IR↑
Teacher	61.0	21.7	0.885
Student	46.9	13.9	0.955
Earlier block: 8	48.9 −2.7	14.6 −0.8	0.921 −0.013
Later block: 16	55.4 +3.8	17.0 +1.6	0.839 −0.095
Regress at all blocks	57.8 +6.2	26.6	0.801 −0.133
Regressed steps: t=1–4	50.0 −1.6	15.6 +0.2	−0.564 −1.498
Adapter steps: t=1–4	51.7 +0.1	16.1 +0.7	0.380 −0.554
ReDiverse-FT (ℓ=12, t=1)	51.6	15.4	0.934

(b) ReDiverse-TF: which directions, and how many.

Setting	DINO↑	TPIPS$_{avg}$ ↑	IR↑
Teacher	61.0	21.7	0.885
Student	46.9	13.9	0.955
Student's own top K=8	49.5 −3.5	14.7 −2.6	0.954 +0.076
Random K=8, teacher's s_k	57.8 +4.8	18.4 +1.1	0.719 −0.159
Teacher's top K=1	48.2 −4.8	14.6 −2.7	0.941 +0.063
ReDiverse-TF (K=8, a=2)	53.0	17.3	0.878

to compare ReDiverse with one method at a time, using the same prompt and initial noises, and judge which is more diverse (overall and by aspect), more realistic, and more faithful to the prompt. Gemini 3.1 Pro (Google DeepMind, 2026) performs the same comparisons as an automatic judge (App. J). In the user study (Fig. 8), raters prefer ReDiverse over every 4-step method in diversity, both overall and across all aspects, for both images and videos. They rate ReDiverse on par with the teacher, even slightly higher overall, and also prefer it in prompt following against every method. The VLM judge agrees that ReDiverse is more diverse than the 4-step methods (Fig. 8). More details about the user study are provided in App. J.

Ablation Study. Tab. 5 changes one design choice at a time on SD3 (full sweeps in App. E). (i) ReDiverse-FT should regress where the teacher's directions are lost, at the collapsed block: at an earlier block the student still matches the teacher, so there is less to correct and the gain shrinks, while at a later block the adapter also reshapes the blocks that render the image, adding diversity at a cost in quality (App. E.1). As the layout is fixed at the first step, regressing at every step or keeping the adapter on after it adds no diversity and lowers prompt alignment (App. E.4). (ii) For ReDiverse-TF, what matters is which directions receive variance: the student's own directions change appearance but not layout (Fig. 3), a single teacher direction is too few because the lost variance is spread over many axes, and random directions of the same scale make samples differ mostly in background texture, at a cost in prompt alignment (App. E.7).

6 Discussion

Probing the activations of distilled generators shows that diversity is lost at a single collapsed block, where the student's features lose the teacher's principal directions of variation that the downstream blocks can still render. ReDiverse restores them with a short, localized fine-tuning (ReDiverse-FT) or without training (ReDiverse-TF), on images and videos. A limitation is that quality still drops at times, most visibly in reward-model scores on images (Tab. 4) and in video realism against the teacher (Fig. 8), as ReDiverse corrects the student after distillation rather than during it. In future work, we will use this probing to design a new distillation method that keeps the teacher's directions at the collapsed block during distillation, so that the student is diverse at the first shot.

AI USE STATEMENT

We used generative AI tools to polish the text of the manuscript and to write the code that produces plots from our own measurements. Additionally, we used generative AI tools for literature search and retrieval; all retrieved works were read and analysed by the authors. Every AI-assisted passage was checked, and all AI-assisted code was tested, by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work modifies publicly released text-to-image and text-to-video models. Restoring diversity lets a distilled model cover the range of outputs its teacher already produces, countering the homogenized outputs of distillation; it adds no new generative capability, so its potential for misuse and its biases are those of the underlying models. The user study was completely anonymous: raters only compared generated images and videos, and no personal data was collected.

REPRODUCIBILITY STATEMENT

All quantitative claims rest on a frozen evaluation protocol. The measurement of the variance characteristic, including the number of prompts and latents per model, the block read, and how the participation ratio, total variance and concentration and their ratios are computed, is specified in App. A. The steering experiment of Sec. 3.2 is specified together with the further examples in App. B, and the two-student comparison in App. C. ReDiverse-FT and ReDiverse-TF are defined by Eqs. (1) and (2) and the injection in Sec. 4, and their design choices are ablated one at a time in App. E. The datasets, baselines, implementation details and evaluation metrics of the main results are given in Sec. 5.1, with checkpoints and training details in App. D, the evaluation protocol and metric definitions in App. F, the baseline reproductions in App. G, and the user-study design in App. J. We will release the code and trained modules upon publication.

REFERENCES

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pp. 224–232. PMLR, 2017.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Pradhaan S Bhat, Rishubh Parihar, Abhijnya Bhat, and R Venkatesh Babu. Don't settle at the mode! mitigating diversity collapse in pretrained flow models via feature self-guidance. *arXiv preprint arXiv:2606.27371*, 2026.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. *Advances in neural information processing systems*, 37:27340–27371, 2024.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026. URL https://openreview.net/forum?id=06SfXJk2VM.

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

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41667–41676, 2026.

FastVideo Team. Fastvideo: Fasth3 4-step dense data-free distillation. https://huggingface.co/FastVideo, 2026.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://openreview.net/forum?id=dF0g-5k05h_.

Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DEiNSfh1k7.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=92oF5bU4cU.

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. IEEE, 2026.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement, 2017. URL https://doi.org/10.1101/214262.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2024.

Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43124–43134, 2026.

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. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pd1P2eUBVfq.

Haoyu Li, Tingyan Wen, Lin Qi, Zhe Wu, Yihuang Chen, Xing Zhou, Lifei Zhu, Xueqian Wang, and Kai Zhang. 1.x-distill: Breaking the diversity, quality, and efficiency barrier in distribution matching distillation. *arXiv preprint arXiv:2604.04018*, 2026a.

Zian Li, Litong Gong, Borui Liao, Pengfei Liu, Xinyu Wang, Xinyuan Wei, Yifan Gao, Tiezheng Ge, and Muhan Zhang. Duet: A diversity-quality duet of distillation experts for two-step video generation. *arXiv preprint arXiv:2608.09637*, 2026b.

lightx2v. Minimax-h3 turbo: 4-step lora distillation. https://huggingface.co/lightx2v/Minimax-h3-Turbo, 2026.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2024. URL https://openreview.net/forum?id=duBCwjb68o.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17719–17728. IEEE, 2025.

MiniMax. Minimax-h3. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026. Text-to-video-with-audio transformer; model card.

Tom Minka et al. Divergence measures and message passing. 2005.

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.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

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

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 *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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. IEEE, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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.

Antoine Salmona, Valentin De Bortoli, Julie Delon, and Agnes Desolneux. Can push-forward generative models fit multimodal distributions? *Advances in Neural Information Processing Systems*, 35:10766–10779, 2022.

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

Neta Shaul, Chao Liu, Arash Vahdat, and Julius Berner. Parallel decoding distillation for fast image and video generation. *arXiv preprint arXiv:2607.26004*, 2026.

Qitan Shi, Cheng Jin, Ziyuan Liu, and Yuantao Gu. Manifold-constrained noise optimization for diverse diffusion sampling. *arXiv preprint arXiv:2607.23937*, 2026.

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

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.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Sheng-Yu Wang, Yotam Nitzan, Aaron Hertzmann, Jun-Yan Zhu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. The many senses of visual similarity: A text-prompted image perceptual metric. *arXiv preprint arXiv:2607.18237*, 2026.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023a.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 893–911, 2023b.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=R3JzLU2qZj.

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

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

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Ankit Yadav, Arpit Garg, Ta Duc Huy, and Lingqiao Liu. Stride: Training-free diversity guidance via pca-directed feature perturbation in single-step diffusion models. *arXiv preprint arXiv:2605.11494*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo 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. IEEE, 2024b.

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

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026a.

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 *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026b.

SUPPLEMENTARY MATERIAL

The supplementary material is organised as follows.

A Measuring the Variance Characteristic **17**

B Steering: Protocol and Further Examples **20**

C Two Students of One Teacher **21**

D Implementation Details **22**

D.1 Backbones and Collapsed Blocks 22

D.2 ReDiverse-FT: Adapter Training and Regression Targets 22

D.3 ReDiverse-TF: Direction Estimation and Injection 22

E Design Choices of ReDiverse **23**

E.1 Which Block to Regress 23

E.2 Which Student Step and Which Teacher Step to Regress 24

E.3 Teacher Guidance During Regression 25

E.4 Which Steps to Open the LoRA at Inference 25

E.5 Which Blocks Carry the LoRA 26

E.6 How Many Directions to Inject 26

E.7 Whose Directions: Teacher, Student or Random 26

F Evaluation Protocol and Metrics **28**

F.1 Prompts, Seeds, and Evaluation Subsets 28

F.2 Diversity Metrics 28

F.3 Video-Specific Diversity 28

F.4 Quality, Alignment, and Reporting 29

G Baseline Reproductions **30**

G.1 Image Baselines 30

G.2 Video Baselines 30

H Additional Quantitative Results **31**

H.1 Paired Comparison with the Student 31

H.2 Per-Aspect TPIPS 32

H.3 Foreground and Background Diversity 32

I Additional Qualitative Results **34**

I.1 Video 34

I.2 Images 38

J User Study Details **40**

A MEASURING THE VARIANCE CHARACTERISTIC

Our starting question was the following: the teacher generates different-looking outputs for different noises, but the student does not; for the same set of noises it generates similar-looking outputs. Our idea was therefore to look at the internal representations behind these outputs, to understand how different noises end up as different-looking outputs.

To measure this, we feed many initial noises through the network, noises that generate different outputs in the teacher and similar outputs in the student, and look at what they produce inside the network. At an internal block each noise gives one activation. We flatten each activation into a vector, so from N noises we have N activation vectors, that is, N points in activation space, and we measure the characteristics of the distribution formed by these points. We do this at every block of the network. With $X_b \in \mathbb{R}^{N\times D}$ holding the N vectors at block b, centred over the noises, the covariance is

$$\Sigma_b = \frac{1}{N} X_b^\top X_b, \tag{3}$$

and its eigenvalues λ_j give the variance along each principal direction of the distribution.

How we measure. We give the same $N = 1000$ noises to the teacher and to the student as the input of their first denoising step, run the conditional branch only, and keep the image or video positions of every block. We use 100 prompts for SDXL and SD3, 20 for Wan 2.1 and MiniMax-H3. Since D is very large, instead of estimating the $D \times D$ covariance matrix we estimate the $N \times N$ Gram matrix $G_b = X_b X_b^\top / N$, whose entries are the inner products between the centred activations of every pair of noises:

$$(G_b)_{ij} = \frac{1}{N} \left(h_i - \bar{h}\right)^\top \left(h_j - \bar{h}\right), \qquad \bar{h} = \frac{1}{N} \sum_{i=1}^{N} h_i, \tag{4}$$

where $h_i \in \mathbb{R}^D$ is the flattened activation of noise i at block b. The Gram matrix has the same nonzero eigenvalues as Σ_b.

Total variance. We first measure the total variance of this distribution, the sum of the eigenvalues:

$$\mathrm{Tr}(\Sigma_b) = \sum_j \lambda_j. \tag{5}$$

It measures how far the points spread, that is, how much the activation moves when the noise changes. The student's variance is not lost (Fig. 9); in fact it is higher than the teacher's at most blocks. This is surprising: since the student's outputs collapse to a few modes, we expected its activations to collapse in variance as well.

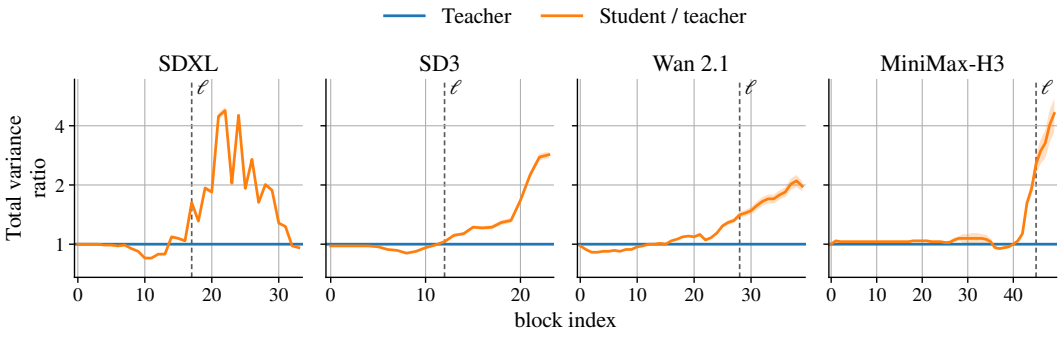


Figure 9: **Total variance**, student divided by teacher, at every block of the first denoising step; the teacher is the line at 1. Mean of the per-prompt ratios with 95% confidence interval; dashed line: the collapsed block ℓ.

Participation ratio. Since we could not find a variance collapse in the internal blocks of the network, we started to check whether this variance is collapsed into a few directions or covers many

directions. An effective way of measuring this is the participation ratio, which counts how many directions the spread effectively occupies:

$$\mathrm{PR}(\Sigma_b) = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2}. \tag{6}$$

A distribution that extends evenly in many directions has a high value, and one stretched along a few directions a low one. The student follows the teacher in the early blocks and then drops below it in the intermediate blocks (Fig. 10). So the variance is not lost, but it is collapsed into fewer directions than in the teacher.

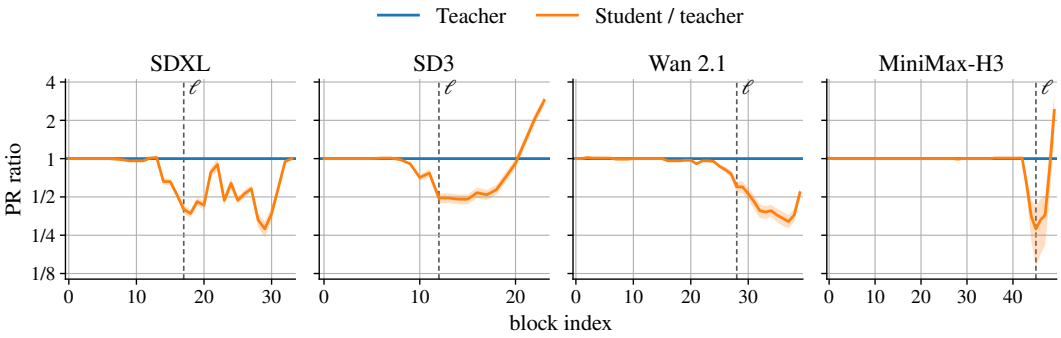


Figure 10: **Participation ratio**, student divided by teacher, at every block of the first denoising step; the teacher is the line at 1. Mean of the per-prompt ratios with 95% confidence interval; dashed line: the collapsed block ℓ.

Concentration. The student keeps the variance but spreads it over fewer directions, so we asked how much variance each of these directions carries. We measure this with the concentration, the total variance divided by the participation ratio,

$$\mathrm{Conc}(\Sigma_b) = \frac{\mathrm{Tr}(\Sigma_b)}{\mathrm{PR}(\Sigma_b)}, \tag{7}$$

and indeed each remaining direction of the student carries more variance than the teacher's (Fig. 11).

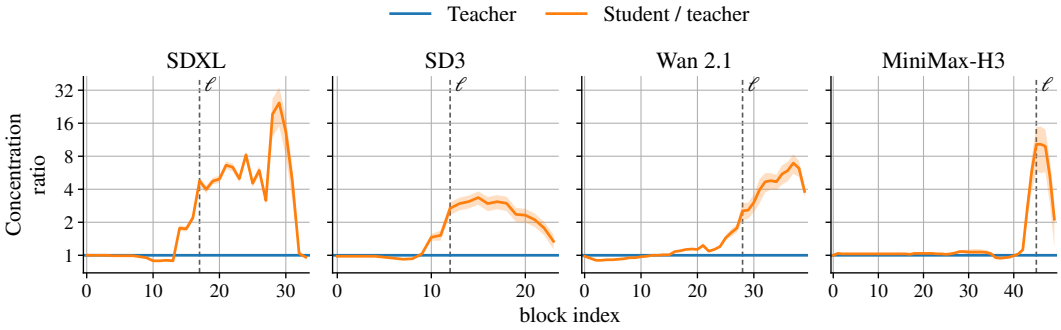


Figure 11: **Concentration**, student divided by teacher, at every block of the first denoising step; the teacher is the line at 1. Mean of the per-prompt ratios with 95% confidence interval; dashed line: the collapsed block ℓ.

Choosing the collapsed block. The drop in participation ratio does not happen at a single block; it builds up over a few blocks (Fig. 10). We choose the collapsed block inside this region, at the block where the concentration makes a clear step up (Fig. 11): there the student's variance has moved into fewer directions, each carrying much more variance than in the teacher, and in almost all following blocks the student stays more concentrated than the teacher. This is also the block that ReDiverse-FT regresses, and our ablation on SD3 supports this choice: regressing earlier constrains a representation that has not lost its directions yet and recovers less diversity, while regressing later recovers more diversity but also changes the blocks that render the image and lowers the output quality (App. E.1).

Eigenvalues at the collapsed block. We can also see this directly in the eigenvalues. Fig. 12 shows, at the collapsed block, which percentage of the total variance each of the leading eigenvalues carries. In the student, the leading eigenvalues carry a much larger share of the variance than in the teacher on all four models, and the first ten of them carry a larger share than the teacher's on every prompt: most of the student's variance sits in a few directions, while the teacher spreads it more evenly over many.

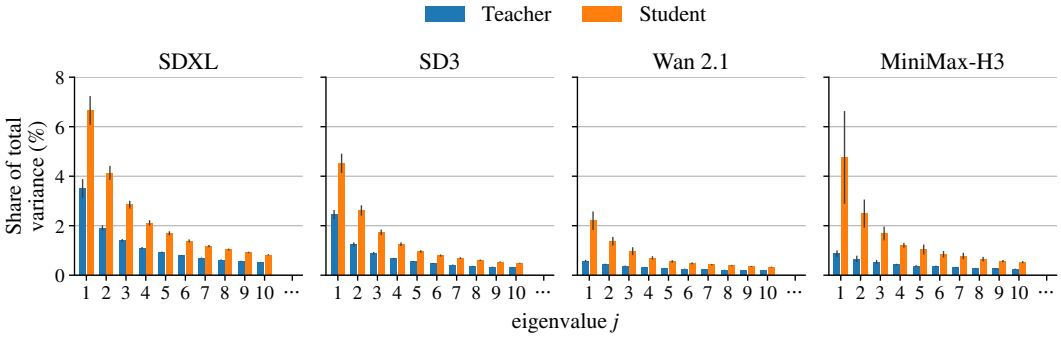


Figure 12: **Share of the total variance per eigenvalue** at the collapsed block, first denoising step, for the 10 largest eigenvalues. Mean over prompts; whiskers: 95% confidence interval.

The next steps. So far we compared the teacher and the student at the first denoising step. Since one student step replaces several teacher steps, we also ran the teacher along its own denoising steps from the same noises and measured its participation ratio and concentration at the collapsed block at each of its steps up to step τ, the teacher step whose activation ReDiverse-FT regresses ($\tau = 7$ for SDXL and 8 for Wan 2.1), against the student's first step (Fig. 13). Along its steps the teacher's participation ratio decreases and its concentration increases, but the student's first step still has, on average, fewer directions and a higher concentration than the teacher at every step up to τ.

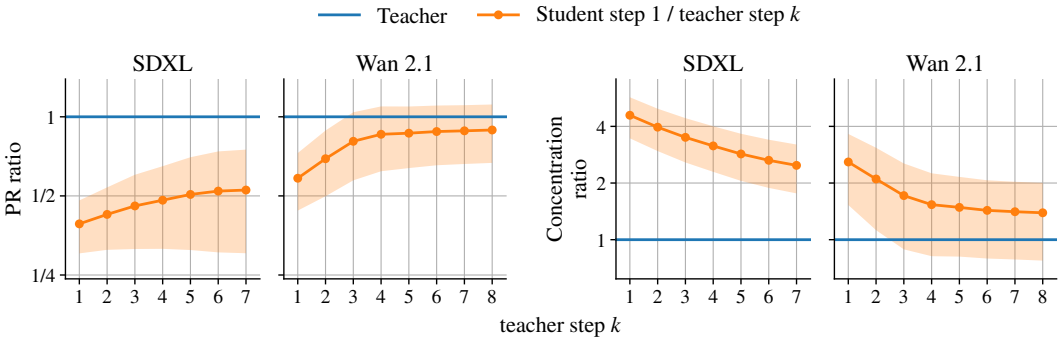


Figure 13: **Participation ratio and concentration along the teacher's steps** at the collapsed block: the student's first step divided by the teacher's step t, for the teacher's first τ steps; the teacher is the line at 1. Mean of the per-prompt ratios with 95% confidence interval.

After the fine-tuning. ReDiverse-FT regresses the student's first-step activation at the collapsed block onto the teacher's, so the quantity it should move is the one measured above. We repeat the measurement on SDXL and Wan 2.1 for the student carrying the adapter, at the first denoising step and against the teacher's first step. The span at the collapsed block widens: the participation ratio rises from 120 to 190 on SDXL and from 535 to 593 on Wan 2.1, against 317 and 918 for the teacher.

B STEERING: PROTOCOL AND FURTHER EXAMPLES

In Sec. 3.2 we asked whether the directions the student loses are the ones that make the outputs different, and whether the student can still use them. We take one noise, move its activation at the collapsed block along a single direction, and look at how the output changes: if its layout changes, this direction carries the diversity the student is missing; if only small details change, it does not.

Directions. For a prompt we feed $N{=}25$ noises, other than the one we steer, to a network and stack their activations at the collapsed block, centred over the noises, as the rows of $X \in \mathbb{R}^{N\times D}$. Since D is very large, we use the Gram matrix $XX^\top$ instead of the covariance matrix: with its top eigenvalue λ_1 and eigenvector v_1, the top principal direction and the spread of the noises along it are

$$u_1 = \frac{X^\top v_1}{\|X^\top v_1\|_2}, \qquad s_1 = \sqrt{\lambda_1/N}, \tag{8}$$

where s_1 is the standard deviation of the noises along u_1. We compute these directions for both the teacher and the student, to check how each model responds to its own direction and how the student responds to the teacher's direction.

Steering. We take one noise from our evaluation noise bank and, during the first forward pass only, replace the activation h at the collapsed block by $h + \alpha\, s_1 u_1$, leaving the rest of the sampling unchanged; the dose α is in units of s_1. The teacher is steered along its own direction with its usual guidance, the shift added to its conditional branch; the student is steered along the teacher's direction and along its own. We find that steering the student along the teacher's direction creates diversity, while steering it along its own direction does not (Fig. 14): along the teacher's direction the cat close-up opens into a view of the cat in a room, along its own only colour, contrast and lighting change.

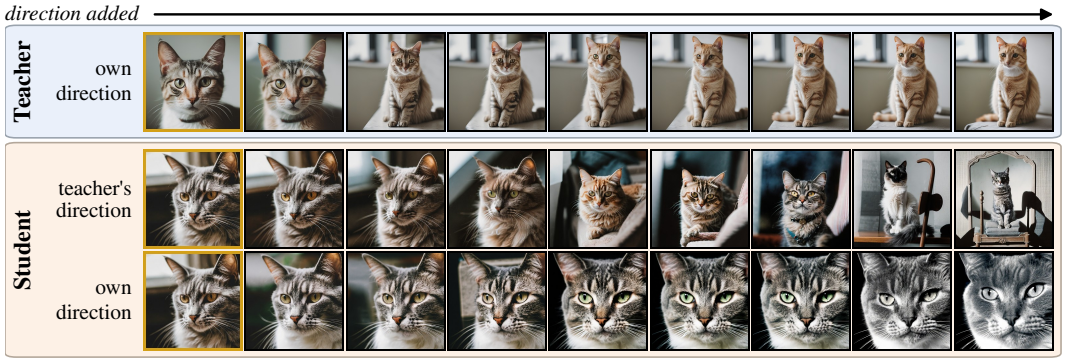


Figure 14: **Steering on SDXL.** The top box moves the teacher along its own direction and the bottom box moves the student along the teacher's direction and along its own direction. Same noise in every row; gold frame: no steering; the dose grows from left to right.

C TWO STUDENTS OF ONE TEACHER

In Sec. 3.2 we saw that two students of the same teacher, MiniMax-H3, reach different output diversity, and that the more diverse one keeps more directions at the collapsed block.

Students. Both are 4-step students of MiniMax-H3, used as released. FastVideo (FastVideo Team, 2026) distills all weights of the teacher with data-free DMD2; Turbo (lightx2v, 2026) is a LoRA on the teacher's weights.

Measurement. As in App. A, for each prompt we feed the same 1000 noises to the teacher and to both students and compute the total variance, PR and concentration at every block of the first step, on the conditional branch only; we use 5 prompts and divide each student's value by the teacher's. For diversity, every model generates 5 clips for each of 20 VBench prompts from the same noises.

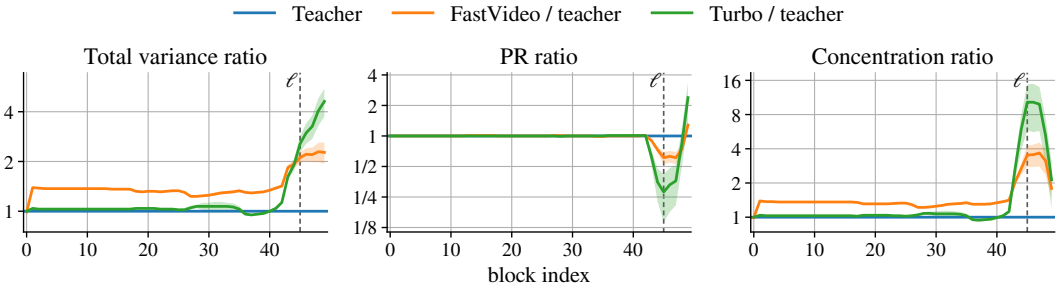


Figure 15: **Two students of MiniMax-H3 along depth.** Total variance, PR and concentration of each student divided by the teacher's at every block of the first step; mean over 5 prompts with 95% confidence interval; dashed line: the collapsed block ℓ (block 45).

Results. Both students keep the teacher's PR until block 42 (Fig. 15). From block 43 on both lose directions, but Turbo collapses much more than FastVideo: its PR drops further and its variance is concentrated on fewer directions. The outputs follow the same order (Tab. 6, Fig. 16): FastVideo is more diverse than Turbo on every diversity metric and close to the teacher.

Table 6: **Two students of MiniMax-H3.** PR at the collapsed block divided by the teacher's, measured on the 5 prompts of this section; Tab. 1 reports 0.37 for the same student on the 20 prompts of App. A. Diversity and text alignment on 20 VBench prompts with 5 clips each. All scores $\times 100$ except Vendi.

		Diversity					Alignment
Model	**PR ratio at ℓ**	**DINO↑**	**V-JEPA2↑**	**TPIPS$_{avg}$ ↑**	**ViCLIP$_{div}$ ↑**	**Vendi↑**	**ViCLIP-S↑**
Teacher	1.00	55.3	21.8	31.0	39.9	2.77	19.7
FastVideo	0.61	**52.3**	**17.5**	**30.1**	**39.0**	**2.74**	18.8
Turbo	0.28	35.8	11.3	22.9	27.7	2.21	**20.3**

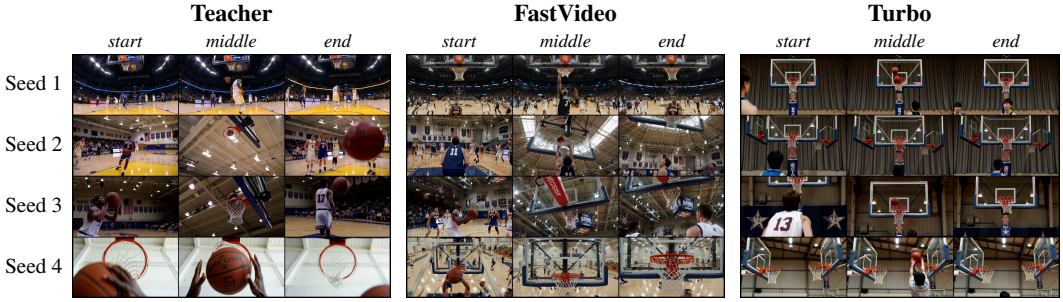


Figure 16: **"A person is dunking basketball" on MiniMax-H3.** Start, middle and end frame of 4 clips per model, same noise in each row.

D IMPLEMENTATION DETAILS

This section provides the implementation details underlying Sec. 5.1. We describe the teacher and student checkpoints, the collapsed block used for each model pair, the training procedure for ReDiverse-FT, and the estimation and injection of directions used by ReDiverse-TF.

D.1 BACKBONES AND COLLAPSED BLOCKS

We use the publicly released 4-step student checkpoints for all four backbones without modification (Tab. 7). The collapsed block ℓ is selected as described in App. A, and τ denotes the teacher step whose activation is used as the regression target for ReDiverse-FT.

Table 7: **Checkpoints and collapsed blocks.** Teacher and student checkpoints, the teacher sampler, the collapsed block ℓ, the teacher step τ used for the regression target, and the guidance used when computing the target.

	SDXL	SD3-M	Wan 2.1	MiniMax-H3
Teacher	SDXL-base-1.0	SD3-medium	Wan2.1-T2V-14B	MiniMax-H3
Student	DMD2 (4-step UNet)	TDM (LoRA)	lightx2v (step- and CFG-distilled)	lightx2v Turbo
Teacher sampler	Euler, 50 steps	FM-Euler, 28 steps	UniPC, 50 steps	FM, 50 steps
Collapsed block ℓ	`up0.attn0`	`block 12`	`block 28`	`block 45`
Teacher step τ	7	9	8	16
Teacher guidance for the target	CFG 8	CFG 7	CFG 5	CFG-distilled

D.2 REDIVERSE-FT: ADAPTER TRAINING AND REGRESSION TARGETS

Adapter. We insert rank-32 LoRA adapters ($\alpha = 32$, no dropout, and zero-initialized B matrices) into the attention projections (query, key, value, and output) and feed-forward layers up to and including the collapsed block. The adapter is activated only in the student's first step; the remaining three steps use the original student.

Targets and loss. Given a caption and noise sample, the student performs its first step with the adapter enabled, and we extract the activation at the collapsed block. The target is the teacher activation at the same block and step τ. The teacher starts from the same noise and caption, is advanced for $\tau - 1$ steps using the corresponding guidance, and is read from its conditional branch at step τ. We minimize plain MSE between the student and teacher activations. For SDXL and Wan 2.1, the loss uses all generated sample tokens. For SD3 and MiniMax-H3, which jointly process prompt and sample tokens, we compute the loss only over image tokens and video tokens, respectively. No additional loss terms are used.

Optimization. We train on the first 100 DiffusionDB captions (Wang et al., 2023b), using 16 noise samples per caption for 1,600 caption–noise pairs, disjoint from the evaluation prompts. Each step uses 4 captions and 8 associated noises per caption, giving a batch of 32 pairs, for 200 optimization steps. We use AdamW with learning rate 10^{-4}, weight decay 0.01, gradient clipping at 10, and a constant learning rate. The forward pass uses bf16, while LoRA parameters are maintained in fp32. We retain the final checkpoint and use the same recipe for all four pairs.

D.3 REDIVERSE-TF: DIRECTION ESTIMATION AND INJECTION

For each evaluation prompt, we run the teacher on $N = 25$ noise samples, independent of the evaluation noise bank, using the conditional branch. We record the activation at step τ (7 for SDXL, 9 for SD3, 8 for Wan 2.1, and 16 for MiniMax-H3), then compute and store the top $K = 8$ directions u_k and spreads s_k from the Gram matrix following App. B.

During inference, each sample i independently draws $\alpha_{i,k} \sim \mathcal{U}[-2, 2]$ and adds $\sum_k \alpha_{i,k} s_k u_k$ to the student's activation at the collapsed block. The perturbation is applied only during the first student step.

E Design Choices of ReDiverse

This section examines the key design choices underlying ReDiverse and establishes the configuration used in our final method. We organize the study around the two components of ReDiverse: ReDiverse-FT, which determines how the student is regressed toward the teacher, and ReDiverse-TF, which determines how diversity directions are injected into the student. For ReDiverse-FT, we systematically study where and when the regression is performed (Tables 8 and 9), which teacher step provides the regression target (Table 10), how teacher guidance is applied during regression (Table 11), when the learned adapter is activated at inference (Table 12), and which blocks carry the LoRA adapter (Table 13). For ReDiverse-TF, we study how many diversity directions are injected (Table 14) and which source provides these directions (Table 15). Each experiment changes one design choice while keeping the remaining components fixed at their defaults. All ablations are conducted on SD3. Each ReDiverse-FT configuration (Tables 8–13) is fully retrained using the training recipe described in Sec. 5.1; the ReDiverse-TF ablations (Tables 14 and 15) involve no training and only change how the directions are injected at inference. We evaluate each configuration on the same 50 prompts, using 100 noise samples per prompt, for a total of 5,000 generated samples per configuration. This is a smaller evaluation set than the one behind Tab. 4, so the teacher and student values here differ from those reported there (for instance the student's DINO is 46.9 rather than 48.4); the rows of a table are comparable with each other, not across tables. We use the same evaluation metrics and protocol as in the main experiments; higher values indicate better performance for all reported metrics. The teacher is included as a reference and is not ranked, while the default configuration is highlighted throughout the tables.

E.1 Which Block to Regress

Table 8: **Regressed block.** We regress ReDiverse-FT at different SD3 blocks, or at all 24 blocks at once ("All blocks"). Block 12, the collapsed block, is used as the default configuration. The teacher is shown only as a reference.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Student	46.9	19.7	20.1	3.89	<u>31.9</u>	**29.3**	**0.955**
Block 8	48.9	21.9	21.6	4.33	<u>31.9</u>	28.3	0.921
Block 10	49.4	22.7	21.7	4.47	**32.0**	28.1	0.913
Block 12 *(default)*	51.6	23.3	22.4	4.72	**32.0**	<u>28.4</u>	<u>0.934</u>
Block 16	55.4	25.1	24.2	5.53	31.8	28.2	0.839
Block 20	55.6	26.9	25.4	5.46	31.8	27.8	0.836
Block 23	<u>57.4</u>	<u>27.8</u>	<u>26.2</u>	<u>5.97</u>	31.8	27.8	0.831
All blocks	**57.8**	**28.5**	**26.6**	**6.05**	31.7	27.7	0.801

Regressing before the collapsed block provides only modest diversity gains over the student, with little change in quality (Table 8). Regressing at later blocks recovers progressively more diversity, but quality degrades substantially, particularly in ImageReward. In contrast, the collapsed Block 12 provides a substantial diversity improvement while retaining the strongest quality among the regressed-block configurations: it achieves the highest HPSv2 and ImageReward among the ablated blocks while matching the best CLIPScore. Regressing at all blocks at once behaves like the latest blocks: it gives the highest diversity of the table but the lowest HPSv2 and ImageReward.

E.2 Which Student Step and Which Teacher Step to Regress

Table 9: **Student step of the regression.** ReDiverse-FT is regressed at different steps of the student's 4-step schedule, or at multiple steps. The adapter is active at the same steps at inference.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Student	46.9	19.7	20.1	3.89	31.9	**29.3**	**0.955**
Step 1 *(default)*	**51.6**	23.3	22.4	4.72	**32.0**	28.4	0.934
Step 2	48.2	23.0	21.2	5.13	31.2	25.8	0.477
Step 3	47.7	21.7	20.0	4.60	31.9	26.0	0.745
Step 4	46.9	19.7	20.1	3.89	31.9	**29.3**	0.954
Steps 1–2	51.2	26.3	**23.2**	6.44	31.0	24.9	0.299
All four steps	50.0	**27.7**	22.8	**7.66**	29.4	20.0	−0.564

Among the single-step configurations (Table 9), Step 1 provides the highest DINO diversity while retaining substantially better quality than Steps 2 and 3; Step 4 is effectively unchanged from the student. Activating the regression at multiple steps increases several diversity measures, particularly Vendi and DreamSim, but this comes with a pronounced degradation in CLIPScore, HPSv2, and ImageReward. Thus, Step 1 provides the strongest DINO diversity among the individual steps while avoiding the severe quality loss associated with multi-step regression.

Table 10: **Teacher step τ of the regression target.** ReDiverse-FT is regressed toward the teacher's activation at step τ of its 28-step schedule.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Student	46.9	19.7	20.1	3.89	31.9	**29.3**	**0.955**
τ=1	54.6	26.5	**24.8**	5.75	31.5	26.3	0.642
τ=3	52.8	24.2	23.6	4.87	31.9	28.6	0.906
τ=9 *(default)*	51.6	23.3	22.4	4.72	**32.0**	28.4	0.934
τ=18	52.6	24.1	22.6	4.90	**32.0**	27.6	0.820
τ=28	**58.7**	**28.8**	24.7	**7.16**	31.1	24.0	0.328

The teacher step reveals a clear trade-off between diversity and quality (Table 10). Early targets such as τ=1 substantially improve diversity over the student, but lead to lower HPSv2 and ImageReward. At the other extreme, τ=28 achieves the strongest diversity, with the highest DINO, DreamSim, and Vendi scores, but incurs a substantial drop in quality and alignment. Among these settings, τ=9 provides a strong overall balance and is therefore used as the default.

E.3 Teacher Guidance During Regression

Table 11: **Teacher target.** The regression target is obtained from the teacher's conditional branch, from the teacher rolled under classifier-free guidance at scale w, or from its unconditional branch. The released scale w=7 is the default.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Student	46.9	19.7	20.1	3.89	31.9	**29.3**	**0.955**
Conditional, no guidance	**58.6**	**27.6**	**26.4**	**5.82**	31.8	27.9	0.807
Guidance w=2	56.1	25.6	24.8	5.33	31.9	28.3	0.882
Guidance w=3.5	53.8	24.3	23.5	4.97	31.9	28.5	0.911
Guidance w=5	52.6	23.7	23.0	4.84	**32.0**	28.5	0.921
Guidance w=7 *(default)*	51.6	23.3	22.4	4.72	**32.0**	28.4	0.934
Unconditional	51.9	23.7	22.7	4.86	**32.0**	28.4	0.925

Teacher guidance controls the trade-off between diversity and quality (Table 11). Removing guidance produces substantially higher diversity but lowers ImageReward, while increasing the guidance scale progressively reduces diversity as the target becomes more aligned with the teacher's guided trajectory. The default w=7 provides lower diversity than weakly guided targets but achieves the highest ImageReward among the guided configurations while maintaining the best CLIPScore.

E.4 Which Steps to Open the LoRA at Inference

Table 12: **Adapter schedule at inference.** The adapter is activated at the indicated steps; "never" corresponds to the unmodified student. The default activates the adapter only at Step 1.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Step 1 only *(default)*	51.6	23.3	22.4	4.72	**32.0**	28.4	0.934
Step 2 only	46.9	21.2	20.8	4.31	31.7	27.9	0.788
Steps 1–2	**51.7**	25.8	23.4	5.91	31.6	26.4	0.607
All four steps	**51.7**	**27.0**	**23.8**	**7.02**	31.3	25.1	0.380
Never (student)	46.9	19.7	20.1	3.89	31.9	**29.3**	**0.955**

Activating the adapter only at Step 1 provides a substantial gain in DINO diversity while largely preserving quality and alignment (Table 12). Activating it at later steps alone does not improve diversity over Step 1 and can reduce quality, particularly ImageReward. Applying the adapter across multiple steps further increases several diversity metrics, most notably Vendi, but consistently degrades HPSv2 and ImageReward. These results support the default Step-1-only schedule as a balanced trade-off between diversity and quality.

E.5 Which Blocks Carry the LoRA

Table 13: **Adapter placement.** We vary the blocks and modules carrying the LoRA adapter. "attn" denotes attention projections, "FFN" denotes feed-forward layers, and ℓ=12 denotes the collapsed block. The default covers all blocks up to ℓ in both attention and FFN layers.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Student	46.9	19.7	20.1	3.89	31.9	**29.3**	**0.955**
Blocks $\leq \ell$, attn + FFN *(default)*	**51.6**	**23.3**	**22.4**	**4.72**	**32.0**	28.4	0.934
Blocks $\leq \ell$, attn only	50.5	**23.3**	**22.4**	4.62	**32.0**	28.4	0.923
Blocks 6–12	50.3	22.7	21.9	4.55	**32.0**	28.3	0.921
Block ℓ only	47.8	20.8	20.3	4.16	**32.0**	28.6	0.935

The placement study (Table 13) shows that adapter coverage primarily affects diversity, while quality and alignment remain comparatively stable across configurations. Expanding the adapter from the collapsed block alone to a broader range of preceding blocks generally increases diversity, with the full placement through the collapsed block achieving the highest DINO, DreamSim, TPIPS, and Vendi scores among the tested configurations.

E.6 How Many Directions to Inject

Table 14: **Number of injected directions K.** ReDiverse-TF injects the top K teacher directions into the student; K=0 corresponds to the unmodified student.

	Diversity				Quality & Alignment		
Setting	DINO↑	DreamSim↑	TPIPS↑	Vendi↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	30.2	6.26	31.8	28.6	0.885
Student (K=0)	46.9	19.7	20.1	3.89	**31.9**	**29.3**	**0.955**
K=1	48.2	20.7	20.8	4.12	**31.9**	29.2	0.941
K=2	49.1	21.5	21.5	4.31	31.8	29.1	0.929
K=4	50.6	22.6	22.5	4.67	31.7	28.9	0.907
K=8 *(default)*	53.0	24.5	24.1	5.18	31.6	28.8	0.878
K=16	56.3	27.3	26.4	5.85	31.3	28.6	0.811
K=24	**58.4**	**29.8**	**28.2**	**6.57**	31.1	28.5	0.761

Increasing K consistently improves diversity across all four metrics (Table 14), but this gain comes with a gradual decline in image quality and alignment. The trade-off becomes increasingly pronounced as more directions are injected, with the largest direction sets showing noticeable degradation in quality. We therefore use K=8 as the default, as it provides a substantial diversity improvement while preserving image quality and alignment more effectively than larger direction sets.

E.7 Whose Directions: Teacher, Student or Random

Table 15: **Source of the injected directions.** We inject K=8 directions obtained from the teacher, the student itself, or random directions with the teacher's spreads, using the same injected energy.

	Diversity								Quality & Alignment		
Setting	DINO↑	DreamSim↑	Vendi↑	$\text{TPIPS}_{\text{ove}}$ ↑	$\text{TPIPS}_{\text{subj}}$ ↑	TPIPS_{bg} ↑	$\text{TPIPS}_{\text{cam}}$ ↑	$\text{TPIPS}_{\text{light}}$ ↑	CLIPScore↑	HPSv2↑	ImageReward↑
Teacher	61.0	32.9	6.26	30.2	20.4	20.9	20.3	16.5	31.8	28.6	0.885
Student	46.9	19.7	3.89	20.1	13.9	12.9	11.9	10.7	**31.9**	29.3	**0.955**
Teacher's directions *(default)*	53.0	24.5	5.18	24.1	16.5	17.9	14.3	13.6	31.6	28.8	0.878
Student's own directions	49.5	20.7	4.33	21.0	14.5	14.5	12.4	11.3	31.8	**29.4**	0.954
Random directions	**57.8**	**25.3**	**6.16**	**24.5**	**17.6**	**20.8**	**15.0**	**14.0**	30.8	28.7	0.719

The source of the injected directions determines both the magnitude and the character of the resulting diversity (Table 15). The student's own directions provide only modest gains over the unmodified student across both aggregate and aspect-conditioned diversity, suggesting that they contribute limited additional variation beyond what the student already represents. Random directions produce the largest diversity gains, with improvements spanning subject, background, camera, and lighting, but this broader variation comes with a substantially larger loss in image quality and alignment. In contrast, teacher directions achieve strong diversity gains while maintaining substantially better alignment, and their aspect-conditioned gains are more balanced across the different visual factors rather than being dominated by a single source of variation. Together, these results indicate that teacher directions provide a more controlled form of diversity injection: they introduce meaningful variation while preserving the semantic and perceptual characteristics of the generated images. We therefore use teacher directions as the default source.

F Evaluation Protocol and Metrics

This section describes the evaluation prompts, noise construction, subsets, and metrics used throughout our experiments. Comparisons are paired: all methods are evaluated on the same prompts and the same initial noise samples.

F.1 Prompts, Seeds, and Evaluation Subsets

For images, we use the 553 GenEval prompts and generate 100 samples per prompt at 1024×1024 resolution. For video, we use 300 prompts selected from VBench with a fixed quota for each dimension and generate 10 clips per prompt. Wan 2.1 generates 81-frame clips at 832×480 and 16 fps, while MiniMax-H3 generates 124-frame clips at 1344×768 and 24 fps.

For each backbone, we use a fixed stored noise bank, so paired methods receive identical initial noise. The coefficients α used by ReDiverse-TF are also generated deterministically, ensuring that ReDiverse-TF (the student carrying the directions, with no adapter) and ReDiverse use identical coefficients. The noises used for direction estimation and the DiffusionDB captions used to train ReDiverse-FT are disjoint from all evaluation prompts and noises.

F.2 Diversity Metrics

Diversity metrics are computed within each prompt and then averaged across prompts. Except for Vendi, which evaluates the full sample set jointly, and LPIPS, which is a learned perceptual distance, the remaining metrics compare pairs of samples from the same prompt using one minus the cosine similarity between their embeddings. Thus, higher values indicate greater diversity.

DINO. For images, we use DINOv2-B patch tokens and compare corresponding spatial locations, averaging their cosine distance. This captures differences in layout, pose, and content. For video, we sample 16 frames throughout each clip, including the first and last, and average the corresponding frame distances.

DreamSim. DreamSim combines DINO, CLIP, and OpenCLIP features and is fine-tuned to human judgments of visual similarity. It captures perceptual differences in layout, pose, color, and object appearance.

LPIPS. LPIPS measures distance between deep VGG features and is sensitive to differences in texture, color, and local structure, including variations that preserve semantic content.

TPIPS. TPIPS conditions the visual embedding on the aspect being compared. We use "Represent the similarity of the image." for the overall score and "Represent the similarity of the image based on X," for the main subject, background scene, lighting and color, and camera viewpoint and framing. $\mathrm{TPIPS}_{\mathrm{avg}}$ is the mean of the overall and four aspect-specific scores. For video, we sample 8 frames and compute the same comparison frame by frame.

Vendi. Vendi measures set-level diversity rather than individual pairwise differences. Given the normalized similarity matrix K, its score is the exponential of the entropy of the eigenvalue distribution of K, reflecting the effective number of distinct samples generated for a prompt. For images, we use DINOv2-B CLS features; for video, we use ViCLIP embeddings.

F.3 Video-Specific Diversity

Video diversity metrics either compare corresponding frames or represent each clip with a single video embedding.

V-JEPA 2. We embed 64 frames from each clip using V-JEPA 2 ViT-L and average the resulting tokens. Because the model processes the frames jointly, the metric captures differences in both appearance and motion.

ViCLIP. ViCLIP is the video–text model used by VBench. We use 8 frames per clip and compute the distance between clip-level embeddings. $\text{ViCLIP}_{\text{div}}$ therefore captures differences in overall content, while the ViCLIP Vendi score measures the effective number of distinct clips per prompt.

Camera trajectory. We estimate camera trajectories with ViPE and express them relative to the first camera position. Each trajectory is resampled to 32 points and normalized by its path length, removing absolute-scale differences. CamTraj is the mean pointwise distance between two normalized trajectories, while CamEnd is the distance between their endpoints. Both are reported after the per-prompt normalization described above: the "Cam" column of Tab. 4 is the normalized CamTraj and the "CamEnd" column of Tab. 17 is the normalized CamEnd, so the two columns are not on the same scale and should not be compared with each other.

F.4 Quality, Alignment, and Reporting

Images. CLIP-S measures prompt alignment using the cosine similarity between CLIP (ViT-B/32) image and text embeddings. HPSv2 (version 2.1) and ImageReward are preference models trained on human judgments of generated images; HPSv2 uses a CLIP ViT-H/14 backbone.

Video. SubjC and BgC are VBench dimensions measuring subject and background consistency across frames using DINO and CLIP features, respectively. ViCLIP-S measures prompt alignment using cosine similarity between the generated clip and prompt embeddings.

Reporting. Diversity metrics are computed within each prompt and averaged across prompts, while quality and alignment metrics are averaged over all generated samples. All reported scores are multiplied by 100 except Vendi, ImageReward, and PickScore. In Tab. 4, Teacher and Noise Optimization are reference methods and are not ranked. Among the remaining methods, the best and second-best values are shown in bold and underlined, respectively.

G BASELINE REPRODUCTIONS

None of the baselines provides checkpoints trained with our teacher–student pairs, so we reproduced each baseline ourselves. All baselines use the same initial noise bank as the other methods and are evaluated on the full evaluation sets using the protocol described in App. F.

G.1 IMAGE BASELINES

Noise Optimization. Noise Optimization (Harrington et al., 2026) keeps the student fixed and optimizes the initial noise at inference time. It optimizes groups of four noise samples using gradient ascent on a diversity objective based on DINOv2 features, together with an HPSv2 objective to preserve image quality. We use the authors' code and, where applicable, their recommended settings. We use 50 optimization iterations for both SDXL and SD3 to keep the evaluation computationally feasible. For SDXL, the diversity objective is evaluated using DPP with 150 points and the HPSv2 objective uses a weight of 3. For SD3, we use 100 DPP points and an HPSv2 weight of 5. Since Noise Optimization directly optimizes a DINOv2-based diversity objective, we report it as an optimization-based reference in Tab. 4.

DP-DMD. DP-DMD (Wu et al., 2026) distills a new 4-step student using DMD and regresses the student's first step toward the teacher's latent representation at an earlier point in the teacher's denoising trajectory. This is intended to preserve the teacher's diversity in the distilled student. The released checkpoint is distilled from SD3.5-Medium rather than from our teacher, so we train new students for both SD3 and SDXL from scratch on DiffusionDB prompts. For SD3, we use the authors' official implementation. For SDXL, we use our own implementation based on the same training procedure, as the official implementation does not support SDXL.

G.2 VIDEO BASELINES

DFD. DFD (Chen et al., 2026) post-trains the student with DMD while incorporating real videos into the training process. In half of the updates, the real score is evaluated on a noised real video with the same caption instead of on a generated sample, encouraging the student to remain closer to the real data distribution. For Wan 2.1, we use the authors' implementation with 29k captioned real video clips. MiniMax-H3 is not supported by the released code, so we adapt the method to its student architecture and train it using 2k real video clips. For both models, we first warm up the fake score for 100 iterations and then train DFD for another 100 iterations. We also evaluate several variants of this setup and use the best configuration for DFD in the baseline comparison.

H Additional Quantitative Results

We provide three additional views of the results in Tab. 4: paired comparisons with the student, per-aspect TPIPS scores, and diversity measured separately on foreground and background regions.

H.1 Paired Comparison with the Student

Because all methods are evaluated with the same prompts and initial noises, we can compare each method with the student on a prompt-by-prompt basis. As shown in Tab. 16, ReDiverse improves diversity on the vast majority of prompts across all backbones, while alignment improves on a smaller but substantial fraction of prompts. Thus, the increased diversity is generally accompanied by only a limited loss of alignment.

Table 16: **Paired comparison with the student.** Percentage of prompts on which each method scores higher than the student using the same prompt and initial noise. We report DINO and TPIPS diversity, and CLIPScore for images or ViCLIP-S for video. All DINO differences are significant ($p < 0.001$, two-sided sign test). **Bold** and underline mark the highest and second-highest value below the dashed line; the teacher and Noise Optimization are references and are not ranked. Two methods can share a win rate while differing in how much they win by: on SD3, ReDiverse-TF and ReDiverse both beat the student on 100% of the prompts in DINO, by $+5.5$ and $+11.7$ points respectively.

Method	DINO	TPIPS	CLIPScore
SDXL			
Teacher	96	98	48
Noise Opt.	100	100	25
DP-DMD	37	24	35
ReDiverse-TF	88	85	**44**
ReDiverse-FT	**95**	**97**	41
ReDiverse	**95**	**97**	38
SD3			
Teacher	99	98	47
Noise Opt.	100	100	28
DP-DMD	97	81	21
ReDiverse-TF	**100**	**99**	31
ReDiverse-FT	94	83	**58**
ReDiverse	**100**	**99**	31

Method	DINO	TPIPS	ViCLIP-S
Wan 2.1			
Teacher	93	94	45
DFD	26	15	**59**
ReDiverse-TF	91	**98**	50
ReDiverse-FT	93	91	40
ReDiverse	**98**	**98**	36
MiniMax-H3			
Teacher	87	86	61
DFD	40	40	**58**
ReDiverse-TF	**92**	**97**	47
ReDiverse-FT	75	77	53
ReDiverse	88	91	47

H.2 PER-ASPECT TPIPS

Tab. 17 breaks down TPIPS into its individual aspects. Across the image and video backbones, ReDiverse improves diversity across multiple aspects, including subject, background, lighting, and camera viewpoint, indicating that the improvement is not concentrated in a single source of variation.

Table 17: **Per-aspect TPIPS.** TPIPS scores for the overall evaluation set and the four aspect-specific subsets. For video, we additionally report CamEnd and the ViCLIP Vendi score. The overall score is computed over all prompts, while the four aspect scores use the 100-prompt subset; CamEnd uses the 30-prompt subset. TPIPS and CamEnd are scaled by 100; Vendi is unscaled. All metrics are higher-is-better. **Bold** and underline mark the highest and second-highest value below the dashed line; the teacher and Noise Optimization are references and are not ranked.

	TPIPS				
Method	**Overall**	**Subject**	**Backgr.**	**Lighting**	**Camera**
SDXL					
Teacher	32.3	23.5	23.4	17.1	24.0
Noise Opt.	33.6	23.8	23.6	20.3	24.4
Student	25.5	17.5	18.5	15.4	18.2
DP-DMD	23.1	16.5	18.6	13.7	16.3
ReDiverse-TF	26.2	18.0	19.2	15.9	18.6
ReDiverse-FT	29.9	21.4	22.6	17.3	21.8
ReDiverse	**30.3**	**21.7**	**23.1**	**17.5**	**22.2**
SD3					
Teacher	27.7	20.2	21.1	15.8	20.9
Noise Opt.	25.5	19.1	17.1	13.1	16.9
Student	19.2	13.9	13.2	10.1	12.4
DP-DMD	22.4	16.5	18.5	13.0	16.6
ReDiverse-TF	22.8	16.4	18.1	12.9	14.7
ReDiverse-FT	21.6	15.8	14.5	10.6	14.2
ReDiverse	**26.9**	**19.5**	**21.8**	**15.4**	**18.0**

	TPIPS						
Method	**Overall**	**Subject**	**Backgr.**	**Lighting**	**Camera**	**CamEnd**	**ViCLIP Vendi**
Wan 2.1							
Teacher	41.4	31.4	29.7	25.2	28.9	69.8	3.72
Student	30.1	23.5	24.5	16.7	21.4	66.7	2.51
DFD	23.7	18.1	19.0	12.5	15.9	46.3	2.23
ReDiverse-TF	33.6	25.3	27.1	22.6	22.9	66.3	2.81
ReDiverse-FT	38.2	30.3	29.4	19.7	26.0	**78.6**	3.63
ReDiverse	**42.5**	**32.9**	**31.8**	**26.8**	**28.8**	78.1	**4.11**
MiniMax-H3							
Teacher	40.3	32.1	29.2	20.3	29.9	70.4	3.74
Student	30.2	25.1	23.5	15.1	21.3	47.6	2.97
DFD	29.5	24.1	22.7	14.4	20.5	46.9	2.88
ReDiverse-TF	34.3	28.1	26.8	19.0	23.6	54.3	3.48
ReDiverse-FT	34.8	28.3	25.9	16.7	25.1	61.7	3.34
ReDiverse	**39.1**	**31.9**	**29.5**	**20.8**	**28.0**	**66.7**	**3.98**

H.3 FOREGROUND AND BACKGROUND DIVERSITY

Tab. 18 separates each sample into its foreground, segmented with SAM 3, and the remaining background. ReDiverse improves diversity in both regions across all backbones, indicating that the recovered variation is not restricted to the background.

Table 18: **Foreground and background diversity.** Diversity of the SAM 3 foreground and of the background, measured separately with DINO and CLIP embeddings, and of the foreground's mask geometry (Pose); mean pairwise distance over the samples of a prompt, averaged over prompts, $\times 100$. Higher is better. **Bold** and underline mark the highest and second-highest value below the dashed line; the teacher and Noise Optimization are references and are not ranked.

	Foreground		Background		
Method	**DINO**	**CLIP**	**DINO**	**CLIP**	**Pose**
SDXL					
Teacher	39.0	16.7	50.1	19.2	21.3
Noise Opt.	41.8	16.3	52.1	18.7	19.2
Student	29.7	12.8	40.8	15.3	15.1
DP-DMD	27.5	12.5	36.9	14.5	13.9
ReDiverse-TF	30.9	13.3	42.5	15.8	15.5
ReDiverse-FT	35.5	15.2	47.5	18.3	19.8
ReDiverse	**36.3**	**15.5**	**49.0**	**18.8**	**20.1**
SD3					
Teacher	32.4	13.8	43.7	17.7	17.3
Noise Opt.	34.2	14.0	40.5	16.0	13.8
Student	22.6	11.2	31.2	13.3	9.9
DP-DMD	27.3	12.3	39.3	15.4	14.0
ReDiverse-TF	26.9	12.8	37.9	15.8	11.8
ReDiverse-FT	27.0	12.2	36.6	14.7	11.2
ReDiverse	**32.6**	**14.7**	**46.1**	**18.1**	**14.9**
Wan 2.1					
Teacher	45.9	15.5	54.8	17.2	17.3
Student	24.3	9.8	35.2	12.0	10.9
DFD	22.5	9.6	33.9	10.1	8.8
ReDiverse-TF	27.0	11.0	37.9	13.0	11.0
ReDiverse-FT	40.3	15.1	51.3	17.4	17.1
ReDiverse	**43.9**	**15.5**	**54.6**	**17.9**	**17.6**
MiniMax-H3					
Teacher	38.9	12.8	50.8	18.3	15.2
Student	24.7	10.0	39.0	16.8	12.0
DFD	24.4	9.7	37.3	16.1	11.7
ReDiverse-TF	28.2	11.2	43.7	18.5	12.5
ReDiverse-FT	32.9	11.7	46.7	17.5	13.5
ReDiverse	**36.2**	**12.5**	**51.3**	**18.8**	**14.4**

I Additional Qualitative Results

I.1 Video

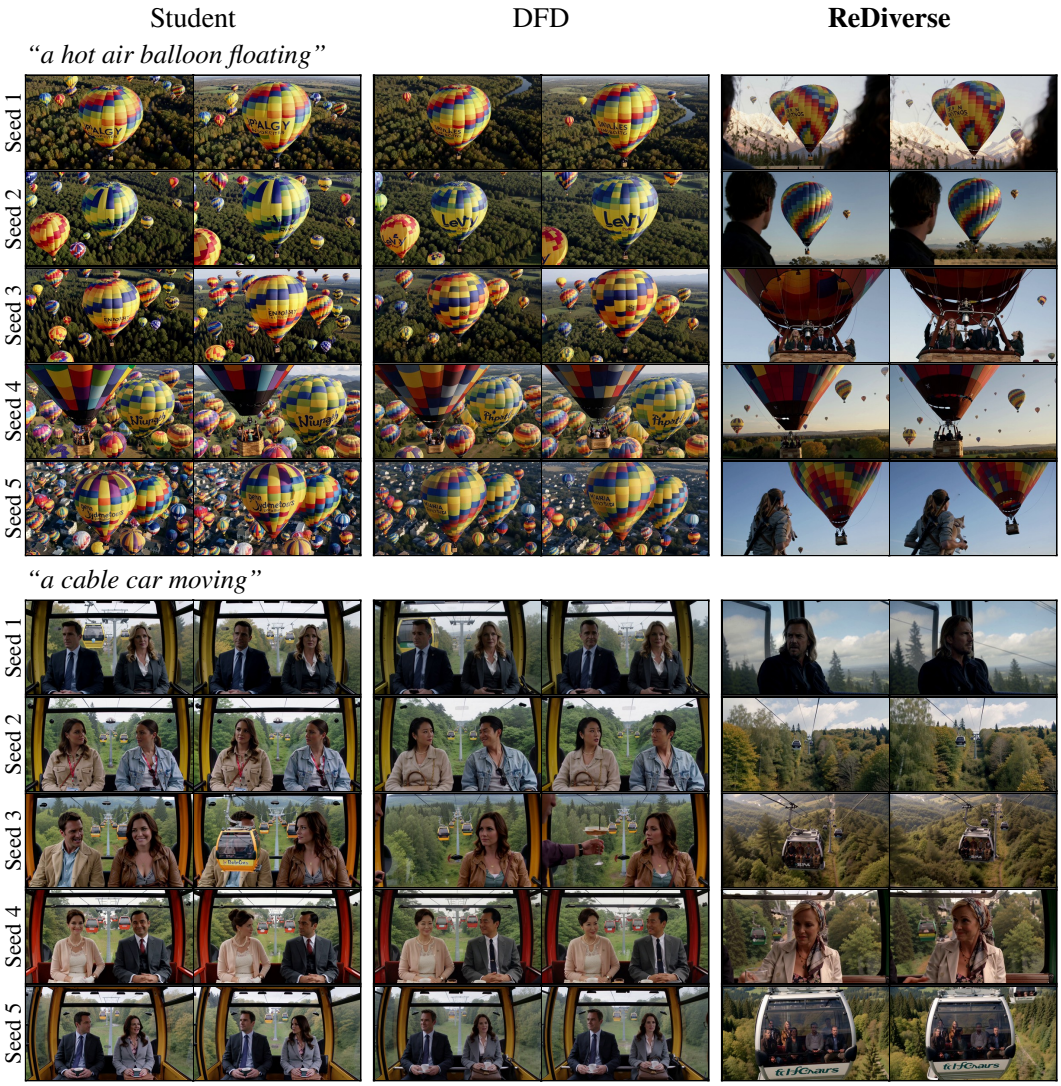


Figure 17: **Qualitative results on MiniMax-H3 (1/2).** Same protocol as Fig. 7: Rows are seeds, each clip shown by its first and last frame, same noise for all three columns. Best viewed as videos available in supplementary material.



Figure 18: **Qualitative results on MiniMax-H3 (2/2).** Same protocol as Fig. 17. Rows are seeds, each clip shown by its first and last frame, same noise for all three columns. Best viewed as videos available in supplementary material.

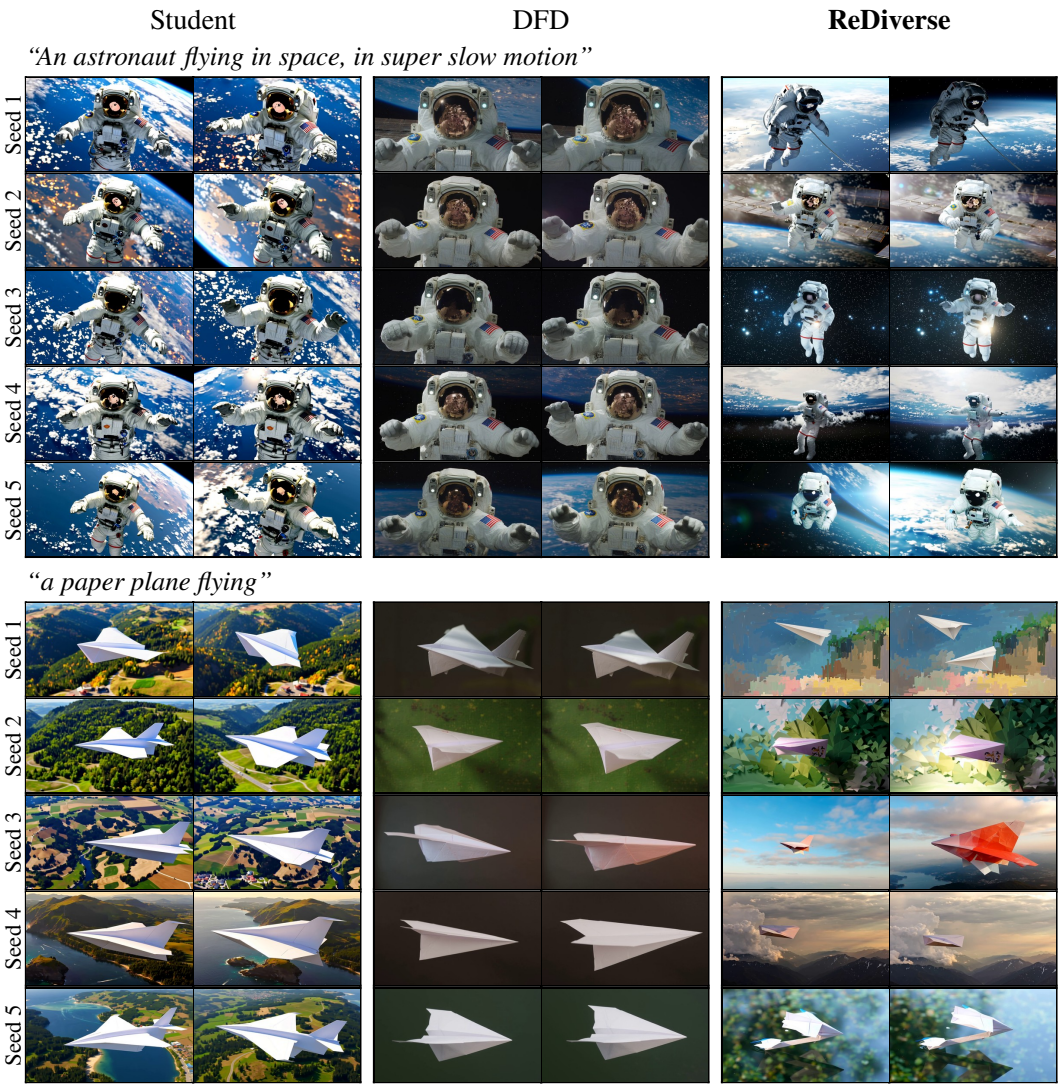


Figure 19: **Qualitative results on Wan 2.1 (1/2).** Rows are seeds, each clip shown by its first and last frame, same noise for all three columns. Best viewed as videos available in supplementary material.

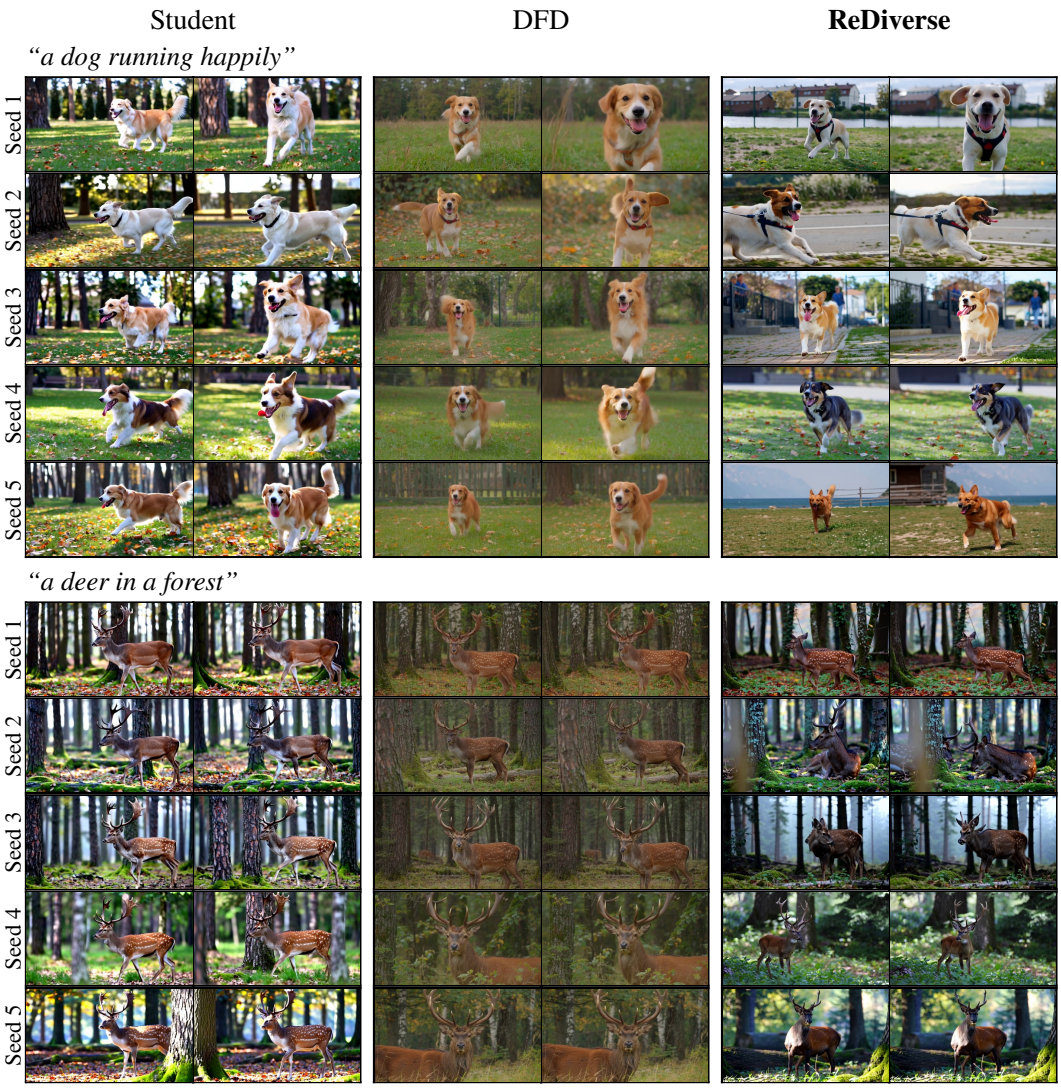


Figure 20: **Qualitative results on Wan 2.1 (2/2).** Rows are seeds, each clip shown by its first and last frame, same noise for all three columns. Best viewed as videos available in supplementary material.

I.2 IMAGES



Figure 21: **Qualitative results on SDXL.** Three prompts and four 4-step methods are shown across nine columns. Each column corresponds to one initial noise, shared across all four methods.

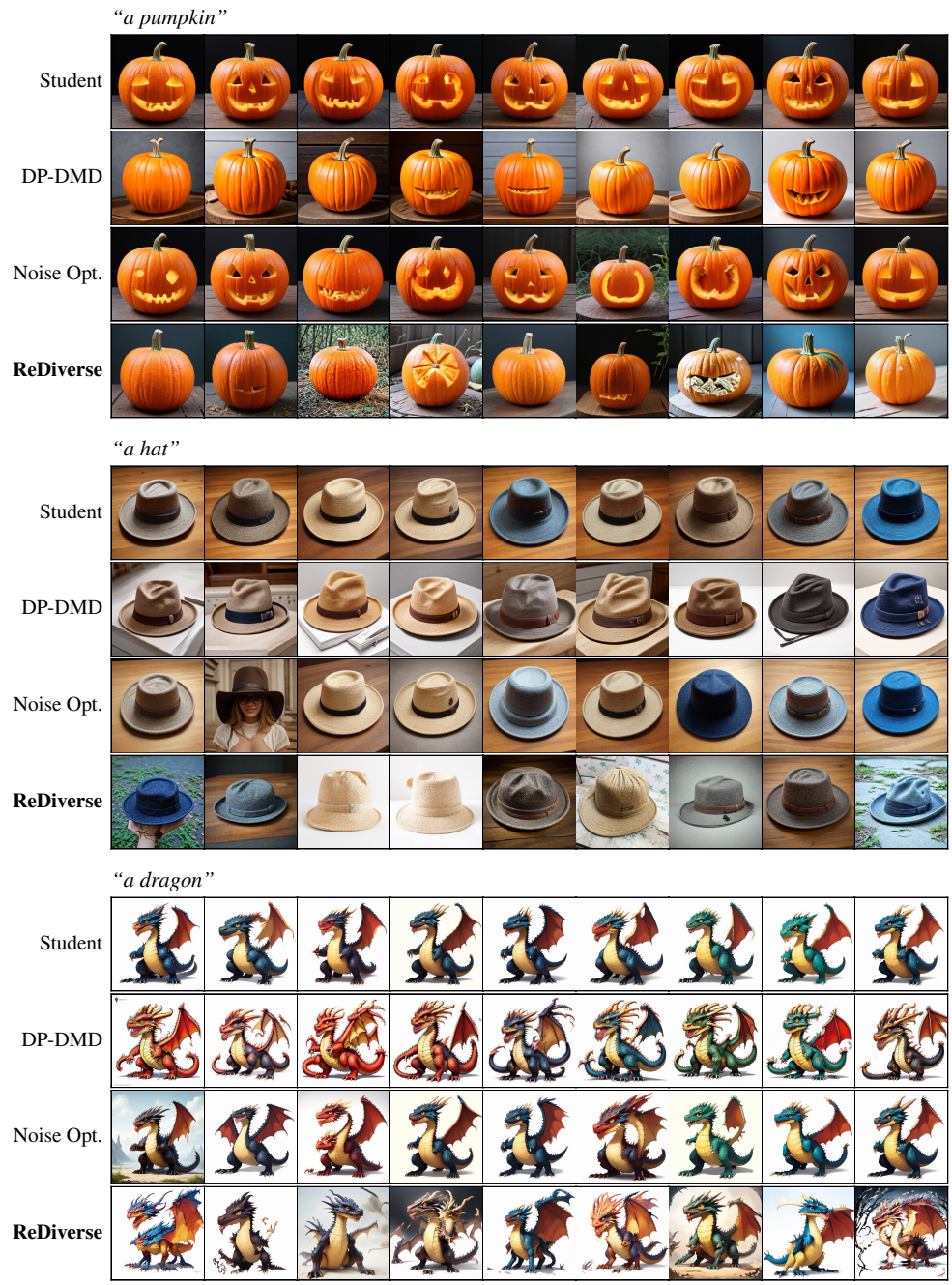


Figure 22: **Qualitative results on SD3.** Three prompts and four 4-step methods are shown across nine columns. Each column corresponds to one initial noise, shared across all four methods.

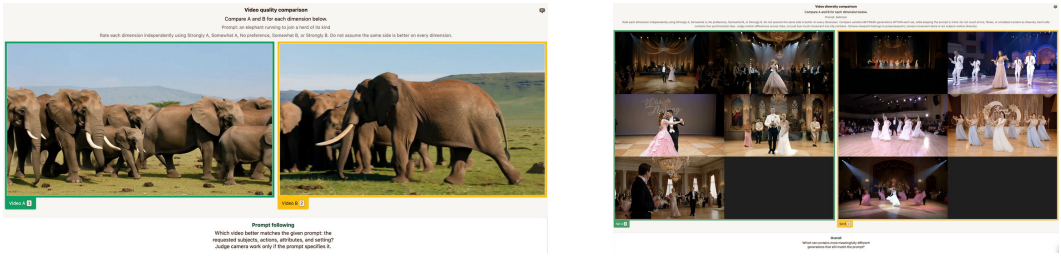


Figure 23: **The rating interface.** Left: quality comparison, with one sample per side at a shared seed. Right: diversity comparison, with five synchronized clips per side at the same five seeds. Each dimension is rated independently by choosing A, B, or No Preference. The instructions also specify what does not count as diversity.

J User Study Details

Automatic diversity metrics are only proxies and are not designed to capture motion, so we additionally evaluated diversity with human raters. Each trial compares ReDiverse with one baseline, the teacher, the student, or a few-step method (DP-DMD (Wu et al., 2026) or Noise Optimization (Harrington et al., 2026) for images; DFD (Chen et al., 2026) for videos), using one prompt and the same initial noises. We use ten prompts per model: generic single-object GenEval prompts for images and VBench scenes/actions for videos, chosen to leave room for diversity.

Instruments. Quality and diversity were collected as separate rating jobs (Fig. 23). The quality job asks about realism and prompt following. The diversity job asks about overall diversity, subject, pose and viewpoint, background, and, for videos, motion. The rubric explicitly excludes errors, flicker, and unrelated content from diversity.

Raters and aggregation. Each job was rated by 30 or 31 raters, 59 distinct people in total. The A/B order was randomized. “No preference” counts as half, so 50% represents an even split.