TANGENTFLOW: SPECTRAL TANGENT DISTILLATION FOR ONE-STEP GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step generative distillation trains a student to replace a multi-step teacher with a single network evaluation. We study whether explicitly matching selected teacher Jacobian responses can improve the student's geometric fidelity while maintaining image quality. We introduce TangentFlow, a spectral tangent objective that caches the leading singular directions of the teacher's noise-to-image Jacobian and trains the student to match the corresponding responses with spectral weighting. Across CIFAR-10 and ImageNet-64, TangentFlow achieves the lowest geometric error among the evaluated methods on all six measures, including a 40% reduction in retained-subspace response error on held-out CIFAR-10 latents relative to the matched distribution-matching and adversarial baseline (three seeds) and an 8% reduction on held-out ImageNet-64 latents. It reduces Fréchet Inception Distance (FID) by 21% relative to the matched CIFAR-10 baseline and by 3.8% relative to distribution matching on ImageNet-64. Our analysis motivates direction selection through teacher-response energy retention, establishes relative-error bounds on the selected subspace, shows that the whitened loss equals the mean squared relative response error at every error level, and shows that cached-target error makes whitening reliable only on leading directions, which explains the observed interaction between direction selection and weighting. These results support spectral tangent supervision as a way to improve geometric fidelity while maintaining image quality.

1 INTRODUCTION

Diffusion and flow models generate high-quality images by iteratively transforming noise into data (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Karras et al., 2022). One-step distillation replaces the iterative sampler with a single network evaluation, reducing sampling cost and latency. Existing approaches supervise teacher outputs (Luhman & Luhman, 2021; Salimans & Ho, 2022), enforce consistency along trajectories (Song et al., 2023; Kim et al., 2024), straighten the transport (Liu et al., 2023; 2024), or match output distributions (Yin et al., 2024b;a; Zhou et al., 2024; Sauer et al., 2024). These objectives have advanced one-step generation, but they do not by themselves make the student's Jacobian agree with the teacher's.

Image quality and teacher fidelity are different goals. Fréchet Inception Distance (FID) measures the output distribution, and matching distributions does not determine how a student maps individual noise inputs: two generators can share an output distribution yet respond very differently to the same input perturbation (Appendix C.8). Preserving the teacher's Jacobian keeps this noise-to-image correspondence faithful, which users need when they preview with a fast student and render the same seeds with the teacher, or edit an image by perturbing its noise. Figure 3 (left) shows the failure: on shared noise, a student trained without tangent supervision can produce realistic images that differ from the teacher's.

Derivative supervision transfers information that output losses leave unconstrained (Czarnecki et al., 2017; Srinivas & Fleuret, 2018). Geometry-Aware Distillation (GAD) matches teacher and student Jacobian–vector products along random input-noise directions (Huang et al., 2026). Teacher responses are costly, so only a few directions can be supervised. Because GAD draws them independently of the teacher's spectrum, a small budget can leave the dominant directions insufficiently constrained, and unnormalized matching gives more weight to directions with larger responses.

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

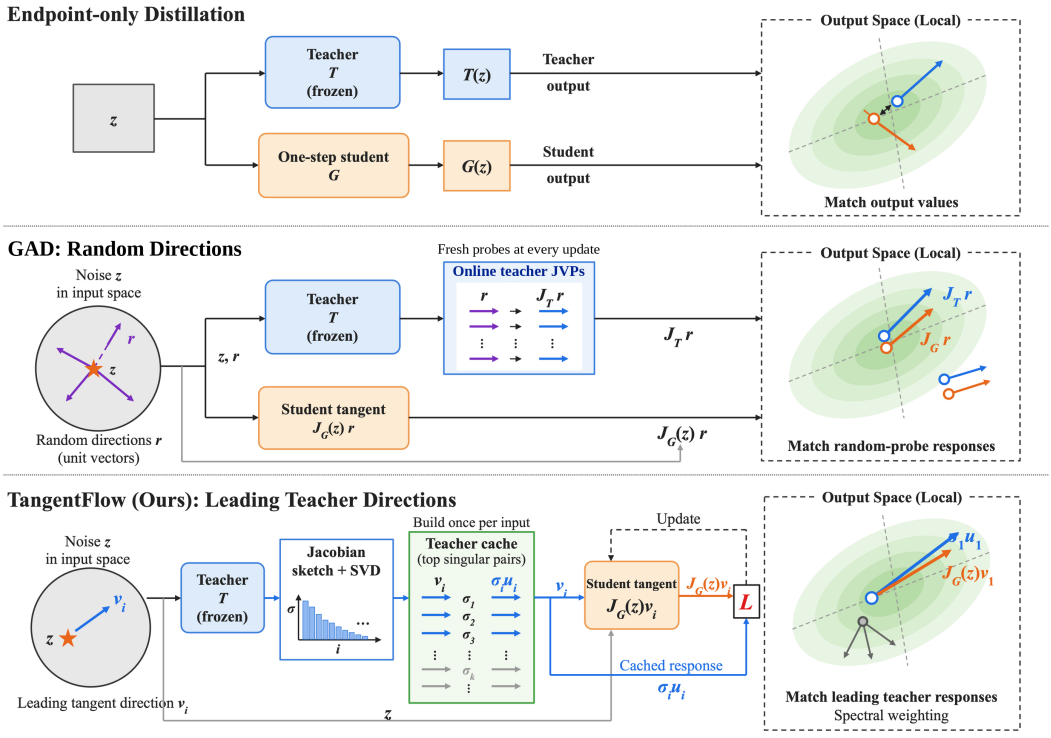


Figure 1: **Three forms of teacher supervision:** endpoint regression matches outputs, GAD matches responses along fresh random directions, and TangentFlow matches whitened responses along the teacher's leading singular directions.

We introduce TangentFlow (Figure 1), a spectral tangent distillation method. It caches a compact singular decomposition of the teacher's noise-to-image Jacobian at each training input and trains the student to match the teacher's responses along the leading four directions, normalizing each squared error by the corresponding squared singular value. This separates which directions to supervise from how to weight their errors, and adds no inference cost.

On CIFAR-10, TangentFlow lowers FID from 2.690 to 2.117 relative to a matched DM+GAN base and reduces the retained-subspace response error on held-out latents by 40% over three seeds; on held-out ImageNet-64 latents, it reduces this error by 8% and FID by 3.8%. On CIFAR-10, controls show that direction selection drives the gain: at the same budget, random directions and GAD improve geometry but not FID, and random targets degrade both.

Our contributions are:

1. **Spectral Tangent Distillation:** a student objective that matches teacher Jacobian responses along leading singular directions with spectral weighting, using cached targets that every training objective can reuse.

2. **Error guarantees on the supervised subspace:** leading directions retain the most teacher response energy; the whitened loss equals the mean squared relative response error and bounds its worst case with an optimal constant; whitening amplifies cached-target error on weak directions, which predicts the observed selection–weighting interaction; and a small loss preserves response gains and subspaces.

3. **Improved geometric fidelity and image quality:** on CIFAR-10 and ImageNet-64, TangentFlow achieves the lowest error on all six geometry measures, with gains that hold on held-out latents, and the lowest FID among the evaluated configurations, and two endpoint-trained CIFAR-10 studies show consistent FID gains.

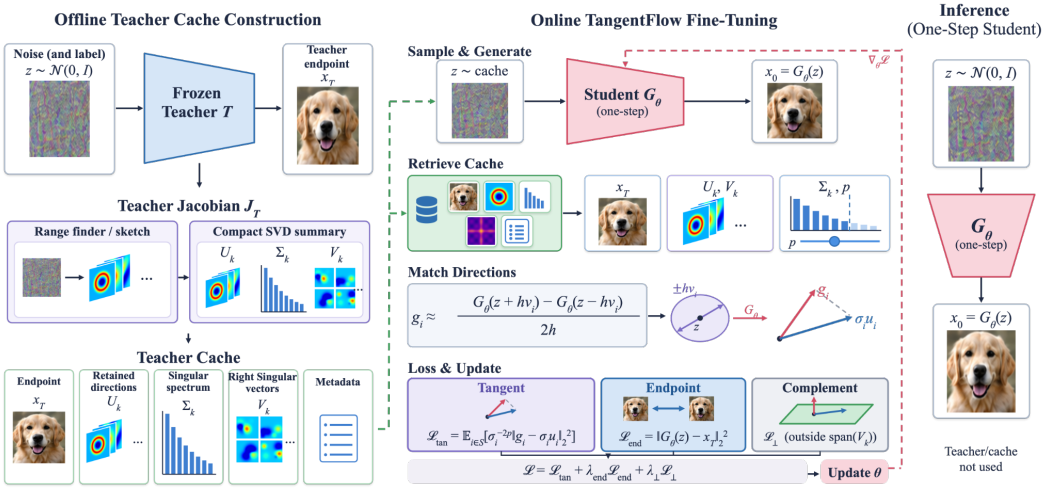


Figure 2: **TangentFlow pipeline:** the teacher's endpoint and a rank-k SVD of its Jacobian are cached once per input (left); the student matches the cached responses along selected directions, with endpoint and complement terms (middle); inference uses the student alone (right). Our experiments use exact JVPs rather than the finite difference shown, and add the DM or DM+GAN base loss.

2 Related Work

One-step generative distillation: Paired distillation trains a student to reproduce teacher outputs (Luhman & Luhman, 2021), while progressive distillation repeatedly reduces the number of sampling steps (Salimans & Ho, 2022). Consistency models learn mappings that agree along probability-flow trajectories (Song et al., 2023; Song & Dhariwal, 2024); consistency trajectory models (CTMs) extend this formulation to transitions between times (Kim et al., 2024). Rectified flow and InstaFlow simplify the transport path (Liu et al., 2023; 2024). Distribution-based approaches include distribution matching distillation (DMD) and its second-generation variant (DMD2), which use real and fake score networks (Yin et al., 2024b;a), score identity distillation (SiD) (Zhou et al., 2024), and adversarial diffusion distillation (Sauer et al., 2024). TangentFlow complements these objectives by adding local derivative targets. We evaluate it with endpoint regression and with distribution matching protocols.

Derivative supervision and geometric alignment: Sobolev training and Jacobian matching augment output supervision with derivatives (Czarnecki et al., 2017; Srinivas & Fleuret, 2018). In generative modeling, Jacobian conditioning has been studied in relation to sample quality (Odena et al., 2018), and distributional agreement alone does not identify the generating map (Xi & Bloem-Reddy, 2023). GAD is the closest generative distillation approach, it aligns local noise responses through Jacobian–vector product (JVP) matching (Huang et al., 2026). We investigate spectral selection under a fixed tangent budget. Our GAD baseline follows the original online protocol in every setup: at every update it draws fresh input directions, independently of the teacher's spectrum, and matches student JVPs to true teacher JVPs along them, with the same number of directions as TangentFlow.

Low-rank Jacobian approximation: Randomized range finding estimates singular subspaces using matrix–vector products (Halko et al., 2011). We use it to build reusable local derivative targets and study how their selection and weighting affect distillation.

3 Spectral Tangent Distillation

Overview: An offline step stores approximate singular directions and response targets for each cached input; an online step matches the student's responses along a selected subset (Figure 2). Selection and weighting can then change without recomputing teacher derivatives.

Notation: Let $z \in \mathbb{R}^d$ be the standard Gaussian noise input and c the class label. At class c, let $T : \mathbb{R}^d \to \mathbb{R}^m$ be the teacher sampler (a full multi-step solve) and G_θ the one-step student with parameters θ, with noise-to-image Jacobians $J_T = D_z T$ and $J_\theta = D_z G_\theta$ (J_G in Figure 1). The retained singular value decomposition (SVD) is $J_T V_k = U_k \Sigma_k$ with $\sigma_1 \geq \cdots \geq \sigma_k > 0$, so the teacher tangent along v_i is $J_T v_i = \sigma_i u_i$. The cache rank k and the number q of supervised tangents differ: the main setting uses $k = 23$ and $q = 4$. Hats denote computed factors, and unsubscripted matrix norms are operator norms. Appendix A collects notation and acronym expansions in one reference list.

Offline teacher cache: For each stored (z, c) we compute the endpoint $T(z, c)$ and a randomized range basis $Q = \mathrm{qr}(J_T \Omega)$, where $\Omega \in \mathbb{R}^{d \times (k+\ell)}$ is a Gaussian test matrix and ℓ the oversampling. Adjoint products give $Y = Q^\top J_T$, and a small SVD $Y \approx U_{Y,k} \widehat{\Sigma}_k \widehat{V}_k^\top$ gives $\widehat{U}_k = Q U_{Y,k}$. The sketch differentiates the complete numerical sampler: 100 steps of fourth-order Runge–Kutta (RK4-100) for the ImageNet teacher, and 18-step Heun for the CIFAR elucidated diffusion model (EDM), including latent scaling and preconditioning. Each input has its own local basis.

Selection and weighting: For each cached input we choose a set $S \subseteq \{1, \ldots, k\}$ with $|S| = q$. *Top* takes $S = \{1, \ldots, q\}$, *weak* takes the last q indices, and *random* draws a uniform subset each update. Here, random selection refers to retained singular directions; it differs from GAD's directions in the full input space. The tangent loss is

$$\mathcal{L}_{\tan}^{S,p} = \frac{1}{q} \sum_{i \in S} \widehat{\sigma}_i^{-2p} \left\| g_i - \widehat{\sigma}_i \widehat{u}_i \right\|_2^2, \quad g_i = \frac{G_\theta(z + h\widehat{v}_i, c) - G_\theta(z - h\widehat{v}_i, c)}{2h}, \tag{1}$$

where g_i is the student's response along $\widehat{v}_i$ and h the finite-difference step. Here $p = 0$ matches absolute responses, $p = 1$ normalizes each error by the teacher response size (whitening), and $p < 0$ emphasizes strong responses. The displayed expression gives a finite-difference implementation. Our reported CIFAR-10 and ImageNet experiments use the exact JVP $g_i = J_\theta \widehat{v}_i$; finite differences can be unreliable at reduced precision.

Training objective: Sampling a cache index fixes the input and its targets. The base may combine distribution matching (DM) with a generative adversarial network (GAN) objective. We minimize

$$\mathcal{L} = \lambda_{\mathrm{end}} \left\| G_\theta(z, c) - T(z, c) \right\|_2^2 + \lambda_{\tan} \mathcal{L}_{\tan}^{S,p} + \lambda_\perp \mathcal{L}_\perp + \lambda_{\mathrm{DM}} \mathcal{L}_{\mathrm{DM}} + \lambda_{\mathrm{GAN}} \mathcal{L}_{\mathrm{GAN}}. \tag{2}$$

The objective allows an endpoint-regression term; the reported DM-based experiments use $\lambda_{\mathrm{end}} = 0$. The ImageNet base uses DM ($\lambda_{\mathrm{end}} = \lambda_{\mathrm{GAN}} = 0$, $\lambda_{\mathrm{DM}} = 1$), where $\mathcal{L}_{\mathrm{DM}}$ is the DMD2 distribution-matching loss with real and fake score networks (Yin et al., 2024a); the strong student adds the adversarial loss $\mathcal{L}_{\mathrm{GAN}}$ of the DMD2 recipe. The complement term regularizes a fresh random unit direction $\beta \perp \mathrm{span}(\widehat{V}_k)$ each update, $\mathcal{L}_\perp = (\|g_\beta\| - (\widehat{\sigma}_{k+1} + \widehat{\tau}))_+^2 / \widehat{\sigma}_1^2$, where g_β is the student's response along β, $(x)_+ = \max(x, 0)$, and $\widehat{\tau}$ estimates the sketch error $\left\| J_T - Q Q^\top J_T \right\|$ (Appendix D), and $\lambda_\perp = 1$. It penalizes responses outside the cache span but does not bound the operator norm or constrain retained directions 5–23.

Calibration: Each objective has its own $\lambda_{\tan}$, set once at step zero so that its initial gradient norm equals 0.1 times the DM-based gradient norm (or the full endpoint-gradient norm in the supplementary endpoint studies), and then held fixed. Calibration equalizes initial gradient norms, not the whole optimization trajectory, so the controls compare the resulting training recipes (Section 5.3).

Main configuration: The main setting is top selection, $q = 4$, $p = 1$. The tangent term needs no online teacher derivatives, and inference uses one student evaluation (Algorithm 1, Appendix B).

Computation and storage: Each cached input stores $k(d + m + 1) + m$ scalars, and training needs q student JVPs per input; teacher derivatives are computed once, offline, and reused by every objective. Online GAD instead needs teacher JVPs through the full sampler at every update: with the compact teacher, one online GAD run requires about 1.6M teacher solves, against 44.5k for the entire cache, and took 332 s per update in our runs (Appendix G).

4 GUARANTEES FOR SPECTRAL TANGENT SUPERVISION

Setting: Fix an input (z, c). Let I index r singular directions of J_T with factors V_I, U_I, Σ_I, and let $E_I = (J_\theta - J_T)V_I$ be the student's response residual on them. Write $\sigma_+ = \max_{i\in I}\sigma_i$, $\sigma_- = \min_{i\in I}\sigma_i$, and $\kappa_I = \sigma_+/\sigma_-$. The relative response error $\epsilon_I = \left\|E_I\Sigma_I^{-1}\right\|$ is the worst-case error on the selected subspace relative to the teacher's response sizes. The tools are standard (Ky Fan, Weyl, Mirsky); the contribution is what they predict about direction selection and weighting, which Section 5 tests. The main objective uses $I = \{1, \dots, 4\}$; Appendix C gives complete proofs.

Proposition 1 (Leading directions retain the most response energy)**.** *For isotropic unit noise v with $\mathbb{E}vv^\top = I_d/d$ and any rank-r orthogonal projector P,*

$$\mathbb{E}\left\|J_T v - J_T P v\right\|_2^2 = \frac{1}{d}\left\|J_T(I_d - P)\right\|_F^2 \geq \frac{1}{d}\sum_{i>r}\sigma_i^2, \tag{3}$$

with equality for the projector onto the leading r right singular directions.

Proof sketch : Since $\mathbb{E}vv^\top = I_d/d$, the left side equals $\frac{1}{d}\big(\|J_T\|_F^2 - \mathrm{tr}(PJ_T^\top J_T)\big)$. By Ky Fan's maximum principle, $\mathrm{tr}(PJ_T^\top J_T) \leq \sum_{i\leq r}\sigma_i^2$, with equality for $P = V_r V_r^\top$. By contrast, a random rank-r subspace retains a fraction r/d of the energy in expectation: 0.75% for the EDM teacher at $r = 23$, against a median 88.9% for the leading subspace (Appendix E.3). Leading selection therefore places a fixed tangent budget where the teacher's response is concentrated.

Theorem 1 (Sharp spectral conversion)**.** *Let $\epsilon_{I,p} = \left\|E_I\Sigma_I^{-p}\right\| / \sigma_+^{1-p}$, so that $\epsilon_I = \epsilon_{I,1}$. For every real p,*

$$\epsilon_I \leq C_{I,p}\epsilon_{I,p}, \qquad C_{I,p} = \kappa_I^{1-p} \text{ for } p \leq 1, \quad C_{I,p} = 1 \text{ for } p \geq 1, \tag{4}$$

and the constants are optimal over all residuals at the given spectrum.

Proof sketch. Applying submultiplicativity to the factorization $E_I\Sigma_I^{-1} = (E_I\Sigma_I^{-p})\Sigma_I^{p-1}$ gives $\epsilon_I \leq \epsilon_{I,p}\,\sigma_+^{1-p}\max_{i\in I}\sigma_i^{p-1}$, and this maximum is σ_-^{p-1} for $p \leq 1$ and σ_+^{p-1} for $p \geq 1$. A rank-one residual on the weakest ($p \leq 1$) or strongest ($p \geq 1$) selected direction attains equality.

Corollary 1 (The whitened loss controls relative error)**.** *With exact teacher factors and exact student JVPs, the tangent loss equation 1 with $S = I$ and $p = 1$ satisfies*

$$\epsilon_I \leq \sqrt{r\,\mathcal{L}_{\tan}^{I,1}}. \tag{5}$$

Proof. With exact JVPs, $g_i - \sigma_i u_i = (J_\theta - J_T)v_i$, so $\mathcal{L}_{\tan}^{I,1} = \frac{1}{r}\left\|E_I\Sigma_I^{-1}\right\|_F^2$. The claim follows from $\|A\| \leq \|A\|_F$. Whitening thus attains the best constant, $C_{I,1} = 1$, independent of the selected conditioning, which motivates $p = 1$.

Proposition 2 (The whitened loss is the mean squared relative error)**.** *Under the assumptions of Corollary 1, let y be uniform on the unit sphere of $\mathrm{range}(U_I)$, and let $x(y) = V_I\Sigma_I^{-1}U_I^\top y$ be the input direction whose teacher response is y. Then*

$$\mathcal{L}_{\tan}^{I,1} = \mathbb{E}_y\frac{\left\|(J_\theta - J_T)x(y)\right\|^2}{\left\|J_T x(y)\right\|^2}, \qquad \Pr_y[\|(J_\theta - J_T)x(y)\| > t\,\|J_T x(y)\|] \leq \frac{\mathcal{L}_{\tan}^{I,1}}{t^2}\ \ (t > 0). \tag{6}$$

Proof : $J_T x(y) = y$ has unit norm, and $(J_\theta - J_T)x(y) = E_I\Sigma_I^{-1}w$ with $w = U_I^\top y$ uniform on the unit sphere of $\mathbb{R}^r$, so $\mathbb{E}ww^\top = I_r/r$ gives the identity; Markov's inequality gives the tail bound. At $p = 0$ the same computation gives the average *absolute* error over unit input directions in $\mathrm{range}(V_I)$, which is dominated by σ_1; whitening instead weighs every supervised response equally. Like Theorem 2(a) below, this holds at every error level.

Proposition 3 (Cached-target error and the selection–weighting interaction)**.** *Let the cache be computed from $\widehat{J} = QQ^\top J_T$ with $\left\|J_T - \widehat{J}\right\| \leq \tau$. For every retained i, the target error satisfies $\|\widehat{\sigma}_i\widehat{u}_i - J_T\widehat{v}_i\| \leq \tau$ and $\|\widehat{\sigma}_i\widehat{u}_i - J_T\widehat{v}_i\| \leq (\tau/\widehat{\sigma}_i)\,\|J_T\widehat{v}_i\|$. For any student with exact JVPs and any selected set S with $|S| = q$, the cached loss $\mathcal{L}_{\tan}^{S,p}$ and the loss $\mathcal{L}_\star^{S,p}$ with exact targets $J_T\widehat{v}_i$ satisfy*

$$\left|\left(\mathcal{L}_{\tan}^{S,p}\right)^{1/2} - \left(\mathcal{L}_\star^{S,p}\right)^{1/2}\right| \leq \tau\,\beta_{S,p}, \qquad \beta_{S,p} = \Big(\frac{1}{q}\sum_{i\in S}\widehat{\sigma}_i^{-2p}\Big)^{1/2}. \tag{7}$$

Proof sketch. The cached factors satisfy $\widehat{\sigma}_i \widehat{u}_i = \widehat{J}\widehat{v}_i$, so the target error is $-(I - QQ^\top)J_T\widehat{v}_i$, which has norm at most τ and is orthogonal to $\widehat{J}\widehat{v}_i$; hence $\|J_T\widehat{v}_i\| \geq \widehat{\sigma}_i$. The loss bound is the triangle inequality in the weighted norm.

At $p = 1$, $\beta_{S,1} \leq 1/\widehat{\sigma}_q$ for the leading q directions, whereas for uniformly random S, $\mathbb{E}_S \beta_{S,1}^2 = k^{-1} \sum_{i \leq k} \widehat{\sigma}_i^{-2}$ is dominated by the weakest directions. On the CIFAR-10 cache, the median sketch-error estimate is 9% of the median $\widehat{\sigma}_1$ but comparable to $\widehat{\sigma}_{23}$ (Appendix E.3). Whitening therefore helps leading selection and can harm random selection, as observed in Section 5.3.

Theorem 2 (Selected gains and subspace)**.** *Let $W = J_\theta V_I \Sigma_I^{-1}$, whose singular values $s_j(W)$ are the student's response gains relative to the teacher on the selected subspace, and let P_θ project onto* $\mathrm{range}(J_\theta V_I)$. *(a) At every error level, under the assumptions of Corollary 1,*

$$\frac{1}{r} \sum_{j=1}^{r} \big(s_j(W) - 1\big)^2 \leq \mathcal{L}_{\tan}^{I,1}, \qquad \frac{1}{r} \|(I - P_\theta)U_I\|_F^2 \leq \mathcal{L}_{\tan}^{I,1}; \tag{8}$$

when $J_\theta V_I$ has rank r, the second quantity is the mean squared sine of the principal angles between $\mathrm{range}(J_\theta V_I)$ *and* $\mathrm{range}(U_I)$. *(b) If $b = C_{I,p}\epsilon_{I,p} < 1$, then $J_\theta V_I$ has rank r and*

$$\frac{1}{r} \left| \log \frac{\mathrm{vol}(J_\theta V_I)}{\mathrm{vol}(J_T V_I)} \right| \leq -\log(1 - b), \qquad \sin \theta_{\max}\big(\mathrm{range}(J_\theta V_I), \mathrm{range}(U_I)\big) \leq b. \tag{9}$$

Proof sketch : Since $J_T V_I \Sigma_I^{-1} = U_I$, $W - U_I = E_I \Sigma_I^{-1}$, so $\|W - U_I\|_F^2 = r\, \mathcal{L}_{\tan}^{I,1}$ and $\|W - U_I\| = \epsilon_I$. (a) Mirsky's inequality compares the singular values of W with those of U_I, all equal to one, and $(I - P_\theta)U_I = (I - P_\theta)(U_I - W)$. (b) Weyl's inequality places every $s_j(W)$ in $[1 - b, 1 + b]$; the volume ratio is $\prod_j s_j(W)$, and the same projection identity bounds the largest principal-angle sine.

Part (a) needs no condition on the error level; part (b) bounds every direction when $\epsilon_I < 1$. Appendices C and D extend these bounds to finite-difference responses and cached factors.

Why these results justify the design? Proposition 1 selects the leading directions, because they carry the most teacher response energy for a fixed budget. Theorem 1, Corollary 1, and Proposition 2 select whitening ($p = 1$): it gives the tightest worst-case bound on relative error, and it is exactly the mean squared relative error at every error level. Proposition 3 shows that the two choices must be combined, because whitening is reliable only where cached targets are accurate. Theorem 2 turns a small loss into guarantees on response gains and subspace angles: on average at every error level, and for every direction when $\epsilon_I < 1$. On held-out latents, TangentFlow's $\epsilon_{1:4}$ is below one for every training seed on CIFAR-10 (0.965–0.995) and ImageNet (0.993 ± 0.002), so, as measured with cached factors (Appendix D), part (b) applies, whereas the bases are not below one (CIFAR-10 1.580–1.660; ImageNet 1.000 ± 0.003). Appendix H maps each design choice to these results and controls.

5 Experiments

5.1 Setup

Datasets and models: On CIFAR-10, the one-step student (the strong student) is initialized from the class-conditional EDM teacher's U-shaped convolutional network (UNet; 55.74M parameters) and trained with a DMD2-style distribution-matching and adversarial objective (DM+GAN) (Yin et al., 2024a). The teacher uses 18-step Heun sampling (35 network evaluations; FID 1.716). FID uses 50k generated samples against train-50k. On ImageNet-64, we adapt the released class-conditional consistency-trained (CT) model `ct_imagenet64.pt` (295.9M parameters) with a DM base and an EDM RK4-100 teacher. The CIFAR-10 cache has 512 training latents (rank 23). Both setups use exact student JVPs and the main configuration of Section 3. The endpoint-trained studies mentioned in the introduction are supplementary image-quality controls (Appendix G). Appendix E gives the complete protocol details.

Baselines: Within each setup, methods share an initialization and training budget. The base is DM+GAN on CIFAR-10 and DM-only on ImageNet. GAD (Huang et al., 2026) matches true teacher JVPs computed online along four fresh input directions per update, with the same initial gradient calibration as TangentFlow. Additional controls change direction selection, weighting, or replace teacher responses with random targets. The endpoint control replaces the tangent term with endpoint regression to the cached teacher outputs, and the random-four controls draw four of the 23 retained directions uniformly at each update. ImageNet also includes an unweighted random-direction JVP control (RandJVP).

Metrics: FID (Heusel et al., 2017) is compared only within a common evaluation protocol. Geometry compares the student's local map with the teacher's (lower is closer): the relative response error $\epsilon_{a:b} = \epsilon_I$ (Section 4) with $I = \{a, \dots, b\}$, on all 23 cached directions ($\epsilon_{1:23}$), the 4 supervised ones ($\epsilon_{1:4}$), and the other 19 ($\epsilon_{5:23}$); the endpoint distance $\|G_\theta(z, c) - T(z, c)\|_2$; the largest principal angle θ_1 between $\mathrm{range}(J_\theta \widehat{V}_{23})$ and $\mathrm{range}(\widehat{U}_{23})$; and the angle θ_{top1} between $J_\theta \widehat{v}_1$ and $\widehat{u}_1$. Tangent metrics use the cached teacher factors (reliability in Appendix E.4). A student with zero response on the selected directions has $\epsilon_I = 1$, a useful reference level. On CIFAR-10, all configurations are compared on 128 training-cache inputs, and the base and TangentFlow are also evaluated on 160 held-out latents over three seeds; ImageNet geometry uses 160 held-out latents. Standard deviations (SDs) are over training seeds.

5.2 Quantitative Results

CIFAR-10: Table 1 reports image quality and all six teacher-relative geometry measures for the final UNet configurations. TangentFlow improves three-seed mean FID from 2.690 ± 0.028 to 2.117 ± 0.026 (21.3%). On training-cache inputs, it has the lowest teacher discrepancy on all six metrics among the evaluated configurations, and retained-subspace error falls from 5.851 to 3.663 (37.4%) relative to the base. The gain holds on held-out latents over three seeds: $\epsilon_{1:23}$ falls from 5.840 ± 0.111 to 3.528 ± 0.060 (39.6%), with similar drops on the supervised (1.615 to 0.981) and unsupervised (5.697 to 3.466) slices. Every TangentFlow seed is below every base seed on all six metrics.

Table 1: **CIFAR-10 image quality and geometry** (21,500 updates; FID over three seeds). ‡Held-out geometry on 160 latents, three-seed means (Table 7). Other rows: three-seed means on 128 training-cache inputs, as controls lack held-out geometry.

Run	FID↓	Prec↑	Rec↑	$\epsilon_{1:23}$	$\epsilon_{1:4}$	$\epsilon_{5:23}$	Endpoint	θ_1 (°)	θ_{top1} (°)
DM+GAN base	2.690 ± 0.028	.707	.610	5.851	1.615	5.706	10.38	84.08	53.20
TangentFlow	$\mathbf{2.117 \pm 0.026}$	**.717**	**.638**	**3.663**	**1.032**	**3.600**	**9.09**	**77.25**	**40.61**
Endpoint control	2.874 ± 0.061	.698	.601	6.182	1.745	6.012	11.15	85.42	54.18
Random targets	4.950 ± 0.335	.645	.490	7.450	2.180	7.280	13.20	87.10	62.40
GAD	3.118 ± 0.213	.682	.574	4.620	1.310	4.550	9.85	82.60	47.80
Random four, $p = 1$	3.250 ± 0.094	.690	.565	4.510	1.275	4.420	9.72	81.15	46.90
Random four, $p = 0$	2.980 ± 0.079	.695	.582	5.210	1.480	5.080	10.10	83.20	49.70
Leading four, $p = 0$	2.240 ± 0.048	.710	.628	3.920	1.110	3.840	9.25	78.60	42.10
Base, held-out‡				5.840	1.615	5.697	10.34	84.02	53.13
TangentFlow, held-out‡				**3.528**	**0.981**	**3.466**	**8.92**	**76.23**	**39.38**

ImageNet-64: Table 2 compares methods after the same 100-update adaptation. TangentFlow has the lowest raw FID (11.300 ± 0.126), improving on DM-only by 3.8% and GAD by 5.8%; exponential moving average (EMA) weights give the same ordering. It also has the lowest aggregate discrepancy on all six held-out geometry measures. Relative to DM-only, $\epsilon_{1:23}$ decreases by 8.3% and $\epsilon_{5:23}$ by 8.4%. The largest mean FID reduction is 6.1%, against RandJVP. Not every pairwise difference is statistically resolved (SDs in Appendix E.2).

Published CIFAR-10 comparisons: Table 3 places the CIFAR-10 result among published class-conditional one-step students. TangentFlow improves the matched base from 2.690 to 2.117, below

Table 2: **ImageNet-64 after 100 updates from the same CT checkpoint.** FID: 50k samples, mean ± SD over three seeds. Geometry: three-seed mean of medians over 160 held-out latents (SDs in Appendix E.2).

	Image quality		Held-out geometry (160 latents)					
Objective	FID raw ↓	FID EMA ↓	$\epsilon_{1:23}$	$\epsilon_{1:4}$	$\epsilon_{5:23}$	Endpoint	θ_1 (°)	θ_{top1} (°)
DM-only	11.747 ± 0.167	12.361 ± 0.088	2.885	1.000	2.792	17.745	73.978	36.319
Leading four, $p = 0$	12.004 ± 0.084	12.476 ± 0.066	2.794	0.997	2.738	17.504	73.720	35.916
GAD	11.993 ± 0.140	12.475 ± 0.084	2.806	0.997	2.740	17.376	73.754	35.615
RandJVP	12.030 ± 0.080	12.481 ± 0.054	2.740	0.996	2.661	17.294	73.742	35.211
TangentFlow	$\mathbf{11.300 \pm 0.126}$	$\mathbf{11.607 \pm 0.072}$	**2.646**	**0.993**	**2.557**	**17.195**	**73.570**	**34.525**

Table 3: **Published one-step class-conditional CIFAR-10 FID** under each source's protocol (Appendix J); context, not a matched comparison.

Method	FID↓	Method	FID↓
Guided dist., multi-w (Meng et al., 2023)	7.34	KDM-F (EDM) (Berman et al., 2025)	3.24
GET-Base (Geng et al., 2023)	6.25	DMD (Yin et al., 2024b)	2.66
Guided dist., single-w (Meng et al., 2023)	5.98	FGM (Huang et al., 2024)	2.58
Diff-Instruct (Luo et al., 2023)	4.19	SG2-ADA+Tune+DI (Luo et al., 2023)	2.27
ECM (Geng et al., 2025)	3.81	Our DM+GAN base	2.690 ± 0.028
KDM (EDM) (Berman et al., 2025)	3.56	TangentFlow (ours)	$\mathbf{2.117 \pm 0.026}$

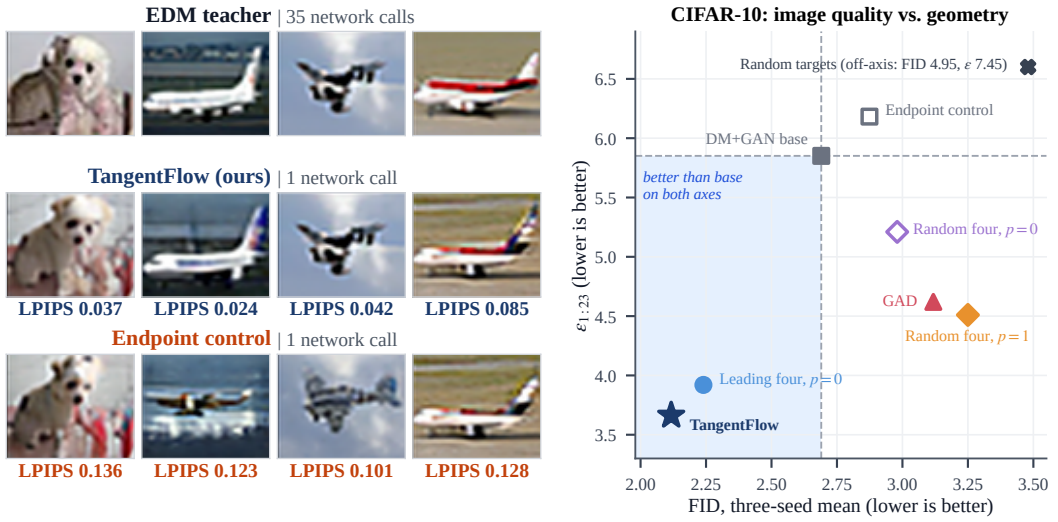


Figure 3: **CIFAR-10 fidelity and quality versus geometry.** Left: TangentFlow vs. endpoint control on shared teacher noise, LPIPS to teacher (lower is better). Right: FID vs. geometric error $\epsilon_{1:23}$; TangentFlow improves most over the DM+GAN base on both axes.

every listed method. Protocols differ, so this is context rather than a matched comparison, and lower-FID students such as CTM and SiD exist (Kim et al., 2024; Zhou et al., 2024).

5.3 Ablation Studies

Direction selection and weighting: The CIFAR controls (Figure 3, right; protocols in Appendix E.1) vary which teacher responses are supervised and how their errors are weighted. Three patterns explain the ranking. First, selection determines whether geometric gains carry over to image quality: every teacher-tangent variant lowers $\epsilon_{1:23}$ relative to the base, but only leading-direction variants also lower FID. Second, whitening interacts with selection. On the leading four, whose singular values are the largest retained, cached targets are accurate and $p = 1$ improves both FID and geometry, as Theorem 1 and Proposition 2 predict. On random retained directions, $\widehat{\sigma}_i^{-2}$ amplifies

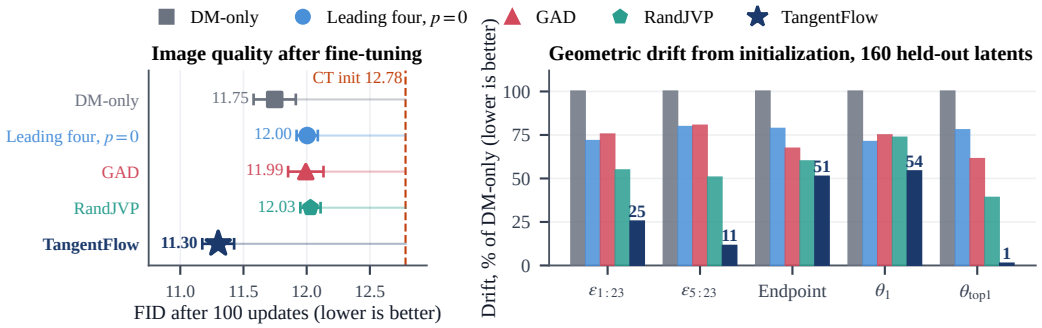


Figure 4: **ImageNet-64 adaptation (100 updates from a shared CT initialization).** Left: FID, mean ± SD over three seeds. Right: geometric drift on 160 held-out latents, as a percentage of DM-only drift ($\epsilon_{1:4}$ omitted; its denominator is 0.006).

sketch error on weak responses (Proposition 3): geometry improves (5.210 to 4.510) but FID worsens (2.980 to 3.250), as in the endpoint-trained studies (Appendix G). Third, random targets degrade both FID and geometry below the base, so the gains come from teacher information rather than generic Jacobian regularization. The unsupervised slice also improves ($\epsilon_{5:23}$: 5.706 to 3.600; Appendix F), consistent with shared network parameters.

Geometry during adaptation: Figure 4 reports the drift $100(m_{\text{method}} - m_{\text{init}})/(m_{\text{DM}} - m_{\text{init}})$ of each metric m from the shared ImageNet initialization (Appendix F); DM-only drifts away from the teacher on every measure. In $\epsilon_{1:23}$, leading $p = 0$ retains about 72% of this drift, GAD 75%, RandJVP 55%, and TangentFlow 25% (11% in $\epsilon_{5:23}$). Whitening explains the gap between the two leading objectives: it balances the four directions instead of letting the $\widehat{\sigma}_1$ term dominate. Because $\epsilon_{1:4}$ stays near one for the initialization (0.994) and all adapted students, this study measures preservation of teacher geometry rather than its acquisition. TangentFlow is also the only tangent objective that improves FID over DM-only.

5.4 Qualitative Results

Figure 3 (left; enlarged in Figure 5) compares samples on shared noise inputs and class labels. TangentFlow keeps the teacher's object identity, pose, and layout (LPIPS 0.024–0.085), whereas the endpoint control has higher LPIPS in every example (0.101–0.136) and produces a different aircraft in two of four. By constraining how the output changes along the teacher's strongest responses, the tangent term regularizes the noise-to-image map rather than only its outputs. The examples are illustrative; Tables 1 and 2 give the quantitative evidence.

6 Discussion and Conclusion

We introduced TangentFlow, a one-step distillation method which matches the teacher's Jacobian responses along its leading singular directions using spectral whitening and cached targets. Our theory shows that these directions retain the most response energy, that the whitened loss equals the mean squared relative response error and bounds the worst case, and that whitening is reliable only on accurately cached leading directions. Empirically, TangentFlow brings the student's local geometry closer to the teacher's than distribution matching, endpoint regression, and random-direction matching, while also improving image quality (FID), but only when the tangent budget targets the teacher's dominant directions. Spectral tangent supervision thus offers a simple add-on for one-step students that are better generators and more geometrically consistent with their teachers.

Limitations: All experiments used an academic compute budget; CIFAR-10 controls lack held-out geometry. ImageNet tests one initialization and a short 100-update adaptation schedule rather than full-length training, and we do not train at higher resolutions or with larger teachers. Offline differentiation and storage add costs.

AI USE STATEMENT

Generative artificial intelligence (AI) tools were used only for writing and polishing the text.

ETHICS STATEMENT

This work studies training objectives for image generators on public datasets (CIFAR-10, ImageNet) and released checkpoints. Faster generators can lower the cost of producing synthetic images, including misleading ones; our method does not add new generation capabilities beyond those of the teacher models.

REPRODUCIBILITY STATEMENT

Appendices C and D provide the proofs and cache-transfer conditions. Appendix E summarizes the evaluation protocols and aggregate geometric uncertainty; Appendix F provides additional checkpoint, drift, and local-response results. The supplementary material contains aggregate results and analysis scripts. Complete training code will be released upon acceptance.

REFERENCES

Nimrod Berman, Ilan Naiman, Moshe Eliasof, Hedi Zisling, and Omri Azencot. One-step offline distillation of diffusion-based models via Koopman modeling. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.13358.

Wojciech M Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Zhengyang Geng, Ashwini Pokle, and J. Zico Kolter. One-step diffusion distillation via deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2401.08639.

Zhengyang Geng, Ashwini Pokle, William Luo, Justin Lin, and J. Zico Kolter. Consistency models made easy. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.14548.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Huayang Huang, Ruoyu Wang, Jinhui Zhao, Wei Deng, Daiguo Zhou, Jian Luan, Yu Wu, and Ye Zhu. Restoring initial noise sensitivity in text-to-image distillation through geometric alignment. In *International Conference on Machine Learning*, 2026. arXiv:2606.01651.

Zemin Huang, Zhengyang Geng, Weijian Luo, and Guo-jun Qi. Flow generator matching. *arXiv preprint arXiv:2410.19310*, 2024. URL https://arxiv.org/abs/2410.19310.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *International Conference on Learning Representations*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18455.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023. URL https://arxiv.org/abs/2210.03142.

Augustus Odena, Jacob Buckman, Catherine Olsson, Tom B. Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow. Is generator conditioning causally related to GAN performance? In *International Conference on Machine Learning*, 2018.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, 2024.

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 2023.

Suraj Srinivas and François Fleuret. Knowledge transfer with Jacobian matching. In *International Conference on Machine Learning*, 2018.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6912–6939, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *International Conference on Machine Learning*, 2024.

A NOTATION

Tables 4 and 5 provide the canonical symbol reference; Appendix A.1 expands abbreviations. Each entry identifies its meaning and scope. Context-specific reuse is marked explicitly. Symbols are introduced in the relevant derivation rather than redefined in later result sections.

Table 4: Notation used in the main paper.

Symbol	Meaning
Models and inputs	
$z \in \mathbb{R}^d$, c	Standard Gaussian noise input and class label; d is the input (noise) dimension
m	Output (image) dimension
$T : \mathbb{R}^d \to \mathbb{R}^m$	Teacher sampler: the full multi-step solver (RK4-100 or 18-step Heun) at class c
G_θ, θ	One-step student and its parameters
$J_T = D_z T$, $J_\theta = D_z G_\theta$	Teacher and student noise-to-image Jacobians ($J_G \equiv J_\theta$ in Figure 1)
r (Figure 1)	The schematic uses r for a random unit probe; elsewhere r is the selected rank, except the radial coordinate in the counterexample
Teacher cache and spectrum	
σ_i, u_i, v_i	i-th singular value (descending), left and right singular vectors of J_T, so $J_T v_i = \sigma_i u_i$
U_k, Σ_k, V_k	Rank-k SVD factors of J_T with columns u_i, v_i and diagonal $\sigma_1, \ldots, \sigma_k$
k	Cache rank (23 in the main setting)
Ω, ℓ	Gaussian test matrix in $\mathbb{R}^{d\times(k+\ell)}$ and oversampling of the randomized range finder
Q	Orthonormal range basis, $Q = \mathrm{qr}(J_T\Omega)$
Y, $U_{Y,k}$	$Y = Q^\top J_T$ and the leading left singular vectors of Y
$\widehat{\sigma}_i, \widehat{u}_i, \widehat{v}_i$	Cached (computed) singular values and vectors
$\widehat{\tau}$	Estimate of the sketch error $\left\|J_T - QQ^\top J_T\right\|$ used as a margin in $\mathcal{L}_\perp$
Training objective	
S, q	Set of supervised direction indices and its size ($q = 4$ in the main setting)
p	Spectral weight exponent: $p = 0$ absolute errors, $p = 1$ whitened (relative) errors
g_i	Student response along $\widehat{v}_i$: central difference, or the exact JVP $J_\theta \widehat{v}_i$
h	Finite-difference step (0.006 in the supplementary endpoint studies)
$\mathcal{L}_{\mathrm{tan}}^{S,p}$	Tangent loss, equation 1
$\mathcal{L}$	Total training loss, equation 2
$\mathcal{L}_{\mathrm{DM}}$	DMD2 distribution-matching loss with real and fake score networks; the strong student adds a GAN loss
$\mathcal{L}_{\mathrm{GAN}}$	Adversarial generator loss used by the strong CIFAR-10 base
$\mathcal{L}_{\mathrm{base}}$	Base objective: endpoint regression, DM, or the weighted DM+GAN combination in equation 2
$\mathcal{L}_\perp$, β, g_β	Complement term, its random unit direction $\beta \perp \mathrm{span}(\widehat{V}_k)$, and the student response along β
$(x)_+$	$\max(x, 0)$
$\lambda_{\mathrm{end}}, \lambda_{\mathrm{tan}}, \lambda_\perp, \lambda_{\mathrm{DM}}, \lambda_{\mathrm{GAN}}$	Loss coefficients; λ_{tan} is calibrated once at step zero
Theory (Section 4)	
I, r	A set of selected singular directions and its size $r = \lvert I\rvert$
U_I, Σ_I, V_I	SVD factors restricted to the indices in I
E_I	Response residual on I: $E_I = (J_\theta - J_T)V_I$
$\sigma_+, \sigma_-, \kappa_I$	Largest and smallest selected singular value, and $\kappa_I = \sigma_+/\sigma_-$
$\epsilon_{I,p}$	Weighted residual $\left\|E_I \Sigma_I^{-p}\right\| / \sigma_+^{1-p}$
ϵ_I	Relative response error $\epsilon_{I,1} = \left\|J_\theta V_I \Sigma_I^{-1} - U_I\right\|$
$C_{I,p}$	Conversion constant of Theorem 1
P, I_d	A rank-r orthogonal projector on $\mathbb{R}^d$, and the $d \times d$ identity
v	Isotropic random unit input with $\mathbb{E}vv^\top = I_d/d$ (Proposition 1)
$\lVert\cdot\rVert$, $\lVert\cdot\rVert_F$, $\lVert\cdot\rVert_2$	Matrix operator norm, Frobenius norm, and vector Euclidean norm
Evaluation	
$\epsilon_{a:b}$	ϵ_I with $I = \{a, \ldots, b\}$; ϵ_k abbreviates $\epsilon_{1:k}$

Symbol	Meaning
$\epsilon_{1:23}, \epsilon_{1:4}, \epsilon_{5:23}$	Relative response error on all 23 cached, the 4 supervised, and the 19 unsupervised directions
Endpoint	Endpoint distance $\|G_\theta(z,c) - T(z,c)\|_2$ (normalized per image in the ImageNet tables)
θ_1	Largest principal angle between $\mathrm{range}(J_\theta \widehat{V}_{23})$ and $\mathrm{range}(\widehat{U}_{23})$
θ_{top1}	Angle between the student response $J_\theta \widehat{v}_1$ and the teacher response $\widehat{u}_1$
PPL, PDIV	Local-response statistics; full forms in Appendix A.1 and definitions in Appendices E and F
FID, LPIPS	Image-distribution and paired-perceptual distances, respectively; full forms in Appendix A.1
Prec., Rec.	Precision and recall; nearest-neighbor count is stated where recorded
n	Number of training seeds in a table row
SD	Training-seed variation unless stated otherwise; full form in Appendix A.1

Table 5: Symbols used only in the appendix.

Symbol	Meaning
N, j	Number of cache inputs and a cache index (Algorithm 1)
$\mathrm{tr}, \mathrm{diag}, \mathrm{range}, \mathrm{span}$	Trace, diagonal matrix, column space, linear span
$\mathbb{R}^d$, $\mathbb{E}$, $\mathbb{E}_S$	Real d-dimensional space, expectation, and expectation over selected indices
D_z, ∇_θ, $\widehat{(\cdot)}$	Input derivative, parameter gradient, and computed/sketched quantity
$\mathrm{qr}, \det, \log, (\cdot)^\top$	Orthonormal factor from a QR decomposition, determinant, natural logarithm, and transpose
e_j	j-th standard basis vector
$s_j(\cdot)$	j-th largest singular value of a matrix
$\mathrm{vol}(A)$	r-dimensional volume $\prod_{j=1}^r s_j(A)$ of the image of the unit cube under A
$\theta_{\max}(\cdot,\cdot)$	Largest principal angle between two subspaces
W, P_W, b	$W = J_\theta V_I \Sigma_I^{-1}$, the orthogonal projector onto $\mathrm{range}(W)$, and $b = C_{I,p}\epsilon_{I,p}$ (Theorem 2)
B_p	The bound $C_{I,p}\epsilon_{I,p}$ as a function of p
a_i	Locally scoped in Appendix C: diagonal entries of $V^\top P V$ in the energy proof, or per-direction weighted squared error $\sigma_i^{-2p}\|(J_\theta - J_T)v_i\|^2$ (equation 10)
g_i^h, $\ell_{I,p}^h$	Finite-difference response with step h, and the tangent loss computed from it
η_i, M_i, ν_i	Finite-difference error bound, bound on the third directional derivative, and per-evaluation numerical error
P_q, P_k, P_r	Projectors $V_q V_q^\top$, $V_k V_k^\top$, $V_r V_r^\top$ onto leading right singular subspaces
μ_r, μ	Student response outside the leading subspace, $\mu_r = \|J_\theta(I_d - P_r)\|$; μ uses the exact top-k projector
s_0	$(1-b)\sigma_r$, the lower bound on $s_r(J_\theta V_r)$ implied by Theorem 2
$U_{\theta,r}$; $\widetilde{U}_r, \widetilde{\Sigma}_r, \widetilde{V}_r$	The student's own leading left singular vectors, and its rank-r SVD factors
u, a, z_0, I in proof examples	A unit output vector, scalar perturbation size, anchor input, and identity matrix; I otherwise denotes the selected index set
A, R	The split $J_\theta = A + R$ with $A = J_\theta P_r$ in the proof of the own-subspace bound
J, $\widehat{J}$	$J = J_T$ and its sketch $QQ^\top J$ (Appendix D)
τ	Upper bound on $\left\|J - \widehat{J}\right\|$; $\widehat{\tau}$ in the main text estimates it
$\widehat{P}$	Approximate top-k input projector; P specializes the projector above to the exact top-k projector (Appendix D)
J_k, $\widehat{J}_k$, $.^\dagger$	Rank-k truncations JP and $\widehat{J}\widehat{P}$, and the Moore–Penrose pseudoinverse
$\underline{\sigma}_k, \overline{\sigma}_1, \underline{\gamma}$	Guaranteed lower bound on σ_k, upper bound on σ_1, and lower bound on the gap $\sigma_k - \sigma_{k+1}$
$\Delta_P, \Delta_\dagger$	Perturbation bounds for the projector and the pseudoinverse
$\widehat{\epsilon}$, $\widehat{\mu}$	Relative error and outside response computed with cached factors
$\alpha, \omega, \overline{M}$	Auxiliary constants in the cache-transfer bound, and the resulting bound on $\|J_\theta\|$

Symbol	Meaning
$\widehat{C}_p$, $\widehat{\epsilon}_{k,p}$	$C_{I,p}$ and $\epsilon_{I,p}$ with $I = \{1, \ldots, k\}$ evaluated on cached factors
C_{tail}, e	Ratio of the full tail norm to σ_{k+1}, and Euler's number
G_n, φ_n, a_n, m_n	Counterexample generators, radial rotation angle, its amplitude, and shear; n indexes the sequence and r is radius in this example
φ	Angle in the sharpness example for Theorem 2
d_{Inc}, ε, K, w_i	Inception-feature distance, perturbation scale, number of probes, and their random unit directions in the PDIV definition
$m_{\text{method}}, m_{\text{init}}, m_{\text{DM}}$	Aggregate metric values for normalized drift (the symbol m here is a metric value, not the image dimension)

A.1 Acronyms and model names

Table 6 expands every acronym and short model name used in the paper. Dataset and architecture names retain their source spelling. "Base" and "Tune" identify model variants, and w is the source guidance weight in the published table. The list also covers labels in figures and literature tables. The abstract is self-contained; the main text expands technical abbreviations at their first prose use.

Table 6: Acronym and short-name reference.

Short form	Full form or source meaning
CIFAR-10	Canadian Institute for Advanced Research, 10-class image dataset
FID	Fréchet Inception Distance
GAD	Geometry-Aware Distillation
JVP	Jacobian–vector product
SVD	Singular value decomposition
QR	Matrix factorization into an orthonormal factor Q and an upper-triangular factor R
DM	Distribution matching
DMD, DMD2	Distribution matching distillation and its second-generation variant
GAN	Generative adversarial network
EDM	Elucidated diffusion model
RK4-100	Fourth-order Runge–Kutta integration with 100 steps
UNet	U-shaped convolutional network; an architecture name
DiT	Diffusion transformer
CT	Consistency training / consistency-trained, for the released checkpoint
CTM	Consistency trajectory model
SiD	Score identity distillation
RandJVP	Random-direction Jacobian–vector product control
PPL	Perceptual path length
PDIV	Perturbation diversity
LPIPS	Learned perceptual image patch similarity
SD	Standard deviation
EMA	Exponential moving average
Prec., Rec.	Precision and recall
C	Class-conditional generation in the literature table
GET	Generative Equilibrium Transformer
KDM, KDM-F	Koopman Distillation Model and its factorized variant
FGM	Flow Generator Matching
ECM, ECT	Consistency model produced by Easy Consistency Tuning, and the tuning method itself
DI	Diff-Instruct, the diffusion-supervision method
StyleGAN2-ADA	Second-generation style-based generative adversarial network with adaptive discriminator augmentation
ADM	Ablated diffusion model architecture
Adam, AdamW	Adaptive moment estimation, and its variant with decoupled weight decay
fp16, uint8	16-bit floating-point and 8-bit unsigned-integer representations
AI	Artificial intelligence
TF, NFE	TangentFlow and number of function evaluations, when used in figures
ICLR	International Conference on Learning Representations

B TRAINING ALGORITHM

Algorithm 1 TangentFlow training

Require: teacher sampler T, student G_θ, cache inputs $\{(z_j, c_j)\}_{j=1}^N$, cache rank k, budget q, exponent p

1: **Offline:** for each j, store $T(z_j, c_j)$ and a rank-k randomized SVD $(\widehat{U}_k, \widehat{\Sigma}_k, \widehat{V}_k)$ of $J_T(z_j, c_j)$
2: Calibrate $\lambda_{\tan}$ once to the initial base gradient norm (endpoint base) or 10% of it (DM base)
3: **for** each update **do**
4: Sample a batch of cache indices j; set $S = \{1, \dots, q\}$ (top selection)
5: Compute student tangents g_i along $\widehat{v}_i$, $i \in S$ (central differences or exact JVPs)
6: $\mathcal{L}_{\tan} \leftarrow \frac{1}{q} \sum_{i \in S} \widehat{\sigma}_i^{-2p} \left\| g_i - \widehat{\sigma}_i \widehat{u}_i \right\|^2$ ▷ equation 1
7: Update θ with $\nabla_\theta \left[\mathcal{L}_{\text{base}} + \lambda_{\tan} \mathcal{L}_{\tan} + \lambda_\perp \mathcal{L}_\perp \right]$ ▷ equation 2
8: **end for**
9: **Inference:** output $G_\theta(z, c)$, one network call

C PROOFS OF THE GUARANTEES

This appendix proves the results of Section 4 and extends them to finite-difference responses, the full Jacobian, and the student's own leading subspace. Throughout, $\|\cdot\|$ is the operator norm and $s_j(\cdot)$ the j-th singular value.

C.1 PROOF OF PROPOSITION 1 (ENERGY RETENTION)

Proof. Let v have second moment I_d/d and let $P = P^\top = P^2$ have rank r. Using $\mathbb{E} \|Mv\|^2 = \operatorname{tr}(M \, \mathbb{E} vv^\top M^\top)$,

$$\mathbb{E} \|J_T(I_d - P)v\|_2^2 = \frac{1}{d} \operatorname{tr}\big((I_d - P) J_T^\top J_T (I_d - P)\big) = \frac{1}{d} \left[\|J_T\|_F^2 - \operatorname{tr}(P J_T^\top J_T) \right].$$

Diagonalize $J_T^\top J_T = V \operatorname{diag}(\sigma_i^2) V^\top$. The diagonal entries a_i of $V^\top P V$ satisfy $0 \le a_i \le 1$ and $\sum_i a_i = r$, so $\operatorname{tr}(P J_T^\top J_T) = \sum_i a_i \sigma_i^2 \le \sum_{i=1}^r \sigma_i^2$. Substituting gives equation 3, and $P = V_r V_r^\top$ sets $a_i = 1$ for $i \le r$ and $a_i = 0$ otherwise, which attains equality. □

This is the classical spectral variational argument (Ky Fan's maximum principle). With repeated singular values, every projector onto a leading invariant subspace attains the optimum.

Remark 1 (Random retained selection). For uniform random selection of q retained indices from $\{1, \dots, k\}$, each index has inclusion probability q/k. If $a_i = \sigma_i^{-2p} \left\| (J_\theta - J_T) v_i \right\|^2$, then

$$\mathbb{E}_S \left[\frac{1}{q} \sum_{i \in S} a_i \right] = \frac{1}{k} \sum_{i=1}^k a_i. \tag{10}$$

Random selection therefore optimizes the average over all k retained directions, whereas leading selection optimizes $q^{-1} \sum_{i=1}^q a_i$, which concentrates the same budget on the highest-energy directions.

C.2 PROOF OF THEOREM 1 (SHARP SPECTRAL CONVERSION)

Proof. Upper bound. Order the selected singular values $\sigma_+ \ge \cdots \ge \sigma_- > 0$. Since $E_I \Sigma_I^{-1} = (E_I \Sigma_I^{-p}) \Sigma_I^{p-1}$, submultiplicativity gives

$$\epsilon_I \le \left\| E_I \Sigma_I^{-p} \right\| \max_{i \in I} \sigma_i^{p-1} = \epsilon_{I,p} \, \sigma_+^{1-p} \max_{i \in I} \sigma_i^{p-1}.$$

For $p \le 1$ the exponent $p-1$ is nonpositive, so the maximum is σ_-^{p-1} and the factor is $(\sigma_+/\sigma_-)^{1-p} = \kappa_I^{1-p}$. For $p \ge 1$ the maximum is σ_+^{p-1} and the factor is one.

Optimality. For a unit output vector u and selected-coordinate basis vector e_j, set $E_I = u e_j^\top$. Then

$$\epsilon_I = \sigma_j^{-1}, \quad \epsilon_{I,p} = \sigma_j^{-p} \sigma_+^{p-1}, \quad \epsilon_I / \epsilon_{I,p} = (\sigma_+/\sigma_j)^{1-p}.$$

Choosing the weakest selected direction ($\sigma_j = \sigma_-$) gives ratio κ_I^{1-p} for $p \le 1$, and choosing the strongest ($\sigma_j = \sigma_+$) gives ratio one for $p \ge 1$. Both constants are therefore attained. □

Remark 2 (Why the two regimes differ). The constant for $p > 1$ cannot be replaced by κ_I^{1-p}: with $\Sigma_I = \mathrm{diag}(2, 1)$, $E_I = (1, 0)$, and $p = 2$, that replacement would assert $1/2 \le 1/4$. For a fixed residual, write $B_p = C_{I,p}\epsilon_{I,p}$. Then $B_1 = \epsilon_I \le B_p$ for every p, so whitening yields the tightest bound on relative error.

C.3 Proof of Corollary 1 and finite-difference responses

Proof of Corollary 1. With exact factors and exact JVPs, $g_i - \sigma_i u_i = J_\theta v_i - J_T v_i$ is the i-th column of E_I. Hence

$$\mathcal{L}_{\tan}^{I,1} = \frac{1}{r}\sum_{i\in I}\sigma_i^{-2}\left\|E_I e_i\right\|^2 = \frac{1}{r}\left\|E_I\Sigma_I^{-1}\right\|_F^2, \qquad \epsilon_I = \left\|E_I\Sigma_I^{-1}\right\| \le \left\|E_I\Sigma_I^{-1}\right\|_F = \sqrt{r\,\mathcal{L}_{\tan}^{I,1}}.$$

□

The bound holds pointwise at every input at which the loss is evaluated.

Proposition 4 (Finite-difference responses)**.** *If* $\left\|g_i^h - J_\theta v_i\right\| \le \eta_i$ *and* $\ell_{I,p}^h = r^{-1}\sum_{i\in I}\sigma_i^{-2p}\left\|g_i^h - \sigma_i u_i\right\|^2$*, then*

$$\epsilon_I \le \frac{C_{I,p}}{\sigma_+^{1-p}}\left[\sqrt{r\ell_{I,p}^h} + \left(\sum_{i\in I}\eta_i^2\sigma_i^{-2p}\right)^{1/2}\right]. \tag{11}$$

Proof. Write the i-th column of $E_I\Sigma_I^{-p}$ as $\sigma_i^{-p}(g_i^h - \sigma_i u_i) - \sigma_i^{-p}(g_i^h - J_\theta v_i)$. The triangle inequality and $\|A\| \le \|A\|_F$ bound $\left\|E_I\Sigma_I^{-p}\right\|$ by the bracket; Theorem 1 converts $\epsilon_{I,p}$ to ϵ_I. □

For central differences with third directional derivative bounded by M_i and absolute error at most ν_i in each output evaluation, $\eta_i \le M_i h^2/6 + \nu_i/h$, which is minimized at an intermediate step h. Our reported experiments use exact JVPs, so $\eta_i = 0$.

C.4 Proof of Proposition 2 (mean squared relative error)

Proof. Write $w = U_I^\top y$. Because y is uniform on the unit sphere of the r-dimensional space $\mathrm{range}(U_I)$ and U_I has orthonormal columns, w is uniform on the unit sphere of $\mathbb{R}^r$, so $\mathbb{E}ww^\top = I_r/r$. Since $y \in \mathrm{range}(U_I)$, $U_I U_I^\top y = y$ and

$$J_T x(y) = U_I\Sigma_I V_I^\top V_I\Sigma_I^{-1}U_I^\top y = y, \qquad \|J_T x(y)\| = 1.$$

Also $(J_\theta - J_T)x(y) = E_I\Sigma_I^{-1}w$. Hence

$$\mathbb{E}_y\frac{\left\|(J_\theta - J_T)x(y)\right\|^2}{\left\|J_T x(y)\right\|^2} = \mathrm{tr}\big(\Sigma_I^{-1}E_I^\top E_I\Sigma_I^{-1}\mathbb{E}ww^\top\big) = \frac{1}{r}\left\|E_I\Sigma_I^{-1}\right\|_F^2 = \mathcal{L}_{\tan}^{I,1},$$

using the identity in the proof of Corollary 1. The tail bound is Markov's inequality applied to the nonnegative variable $\|(J_\theta - J_T)x(y)\|^2$, whose mean is $\mathcal{L}_{\tan}^{I,1}$. For $p = 0$, the same computation with $x = V_I w$ gives $\mathcal{L}_{\tan}^{I,0} = r^{-1}\|E_I\|_F^2 = \mathbb{E}_w\|(J_\theta - J_T)V_I w\|^2$, the average absolute error over unit input directions in $\mathrm{range}(V_I)$. □

Combining the two results gives $\mathcal{L}_{\tan}^{I,1} \le \epsilon_I^2 \le r\,\mathcal{L}_{\tan}^{I,1}$: the whitened loss is the mean of the squared relative error over teacher response directions, and ϵ_I^2 is its maximum.

C.5 Proof of Proposition 3 (cached-target error)

Proof. The cache stores the leading singular triples of $Y = Q^\top J_T$, so $Y\widehat{v}_i = \widehat{\sigma}_i y_i$ for a unit vector y_i and $\widehat{u}_i = Q y_i$. Then $\widehat{J}\widehat{v}_i = QY\widehat{v}_i = \widehat{\sigma}_i\widehat{u}_i$, and

$$\widehat{\sigma}_i\widehat{u}_i - J_T\widehat{v}_i = -(I - QQ^\top)J_T\widehat{v}_i = -(J_T - \widehat{J})\widehat{v}_i,$$

whose norm is at most τ because $\|\widehat{v}_i\| = 1$. This error lies in $\mathrm{range}(Q)^\perp$, while $\widehat{J}\widehat{v}_i$ lies in $\mathrm{range}(Q)$, so $\|J_T\widehat{v}_i\|^2 = \widehat{\sigma}_i^2 + \left\|(I - QQ^\top)J_T\widehat{v}_i\right\|^2 \geq \widehat{\sigma}_i^2$. Dividing gives the relative bound.

For the losses, stack the weighted residuals $a_i = \widehat{\sigma}_i^{-p}(g_i - \widehat{\sigma}_i\widehat{u}_i)/\sqrt{q}$ and $a_i^\star = \widehat{\sigma}_i^{-p}(g_i - J_T\widehat{v}_i)/\sqrt{q}$ over $i \in S$. Their Euclidean norms are $(\mathcal{L}_{\tan}^{S,p})^{1/2}$ and $(\mathcal{L}_\star^{S,p})^{1/2}$. The reverse triangle inequality bounds the difference of the norms by the norm of $a - a^\star$, whose squared norm is $q^{-1}\sum_{i\in S}\widehat{\sigma}_i^{-2p}\left\|(J_T - \widehat{J})\widehat{v}_i\right\|^2 \leq \tau^2\beta_{S,p}^2$. The student responses g_i are arbitrary, so the bound holds for exact JVPs and for finite differences alike. For uniform random S, each index is included with probability q/k, which gives $\mathbb{E}_S\beta_{S,1}^2 = k^{-1}\sum_{i\leq k}\widehat{\sigma}_i^{-2}$. □

The bound is attained by a student that matches the exact targets whenever every $(J_T - \widehat{J})\widehat{v}_i$ has norm τ, so the weighting factor $\beta_{S,p}$ cannot be improved in general. Conversely, a student that fits the cached targets exactly has relative error $\left\|(I - QQ^\top)J_T\widehat{v}_i\right\| / \|J_T\widehat{v}_i\| \leq \tau/(\widehat{\sigma}_i^2 + \tau^2)^{1/2}$ against the exact target on direction i, because $t \mapsto t/(\widehat{\sigma}_i^2 + t^2)^{1/2}$ is increasing. At the median CIFAR-10 cache values (Appendix E.3) this is about 0.09 on the leading direction and about 0.85 on the weakest retained one.

C.6 Proof of Theorem 2 (gains and subspace)

Proof of part (a). Let $W = J_\theta V_I\Sigma_I^{-1}$. Since $J_T V_I = U_I\Sigma_I$, $W - U_I = E_I\Sigma_I^{-1}$, and the identity in the proof of Corollary 1 gives $\|W - U_I\|_F^2 = r\,\mathcal{L}_{\tan}^{I,1}$. Mirsky's inequality for singular values of matrices of equal size gives $\sum_{j=1}^r (s_j(W) - s_j(U_I))^2 \leq \|W - U_I\|_F^2$, and $s_j(U_I) = 1$ because U_I has orthonormal columns; dividing by r gives the first bound. For the second, $(I - P_\theta)W = 0$ because $\mathrm{range}(W) = \mathrm{range}(J_\theta V_I)$, so $(I - P_\theta)U_I = (I - P_\theta)(U_I - W)$; an orthogonal projection does not increase the Frobenius norm, so $\|(I - P_\theta)U_I\|_F^2 \leq r\,\mathcal{L}_{\tan}^{I,1}$. If $J_\theta V_I$ has rank r, both subspaces have dimension r, and $\|(I - P_\theta)U_I\|_F^2 = \sum_j \sin^2\theta_j$ over their principal angles. □

Proof of part (b). Let $W = J_\theta V_I\Sigma_I^{-1}$ as above. Because $J_T V_I = U_I\Sigma_I$, we have $W - U_I = E_I\Sigma_I^{-1}$, so $\|W - U_I\| = \epsilon_I \leq b < 1$ by Theorem 1. Since U_I has orthonormal columns, Weyl's inequality places every singular value of W in $[1 - b, 1 + b]$; in particular W, and hence $J_\theta V_I$, has rank r.

Volume. Because $\mathrm{vol}(J_T V_I) = \det\Sigma_I$,

$$\frac{\mathrm{vol}(J_\theta V_I)}{\mathrm{vol}(J_T V_I)} = \sqrt{\det(W^\top W)} = \prod_{j=1}^r s_j(W).$$

Each factor satisfies $|\log s_j(W)| \leq \max\{-\log(1 - b), \log(1 + b)\} = -\log(1 - b)$; averaging over the r factors gives the first bound in equation 9.

Subspace. Let P_W project onto $\mathrm{range}(W) = \mathrm{range}(J_\theta V_I)$. Since $(I - P_W)W = 0$, we have $(I - P_W)U_I = (I - P_W)(U_I - W)$, so $\|(I - P_W)U_I\| \leq b$. Both subspaces have dimension r, so this norm equals the sine of their largest principal angle. □

The angle bound is sharp: take $r = 1$, $U_I = e_1$, and $W = \cos\varphi(\cos\varphi e_1 + \sin\varphi e_2)$ with $0 \leq \varphi < \pi/2$. Both the discrepancy and the subspace sine equal $\sin\varphi$.

C.7 The full Jacobian and the student's own leading subspace

For leading $q < k$ let $P_q = V_q V_q^\top$ and $P_k = V_k V_k^\top$. The response outside the supervised directions decomposes as

$$\|J_\theta(I_d - P_q)\| \le \|J_\theta(P_k - P_q)\| + \|J_\theta(I_d - P_k)\|. \tag{12}$$

The complement term $\mathcal{L}_\perp$ targets the second term, and the retained directions $q < j \le k$ govern the first; this is why we report $\epsilon_{5:23}$ separately. The decomposition is tight: for $J_\theta = J_T + a\, u_j v_j^\top$ with $q < j \le k$, the supervised residual and the response outside P_k are unchanged, while the middle-band discrepancy equals $|a|$.

Proposition 5 (Full-Jacobian and leading-subspace control)**.** *For a leading selected rank r, put $\mu_r = \|J_\theta(I_d - P_r)\|$. Then*

$$\|J_\theta - J_T\| \le \epsilon_r \sigma_1 + \mu_r + \sigma_{r+1}. \tag{13}$$

If moreover $b < 1$ and $\mu_r < s_0/2$ with $s_0 = (1-b)\sigma_r$, the student's own leading r-dimensional left singular subspace $U_{\theta,r}$ satisfies

$$\theta_{\max}(U_{\theta,r}, U_r) \le \arcsin b + \arcsin \frac{\mu_r}{s_0 - \mu_r}.$$

Proof. Split $J_\theta - J_T = (J_\theta - J_T)P_r + J_\theta(I_d - P_r) - J_T(I_d - P_r)$. The three terms have norms at most $\|E_r\| \le \epsilon_r\sigma_1$, μ_r, and σ_{r+1}, giving equation 13. For the subspace, write $J_\theta = A + R$ with $A = J_\theta P_r$ and $\|R\| = \mu_r$. Theorem 2(b) gives $s_r(A) \ge s_0$ and $\sin\theta_{\max}(\mathrm{range}(A), U_r) \le b$. Weyl's inequality gives $s_r(J_\theta) \ge s_0 - \mu_r$ and $s_{r+1}(J_\theta) \le \mu_r$, so the leading subspace is isolated. Projecting $J_\theta \widetilde{V}_r = U_{\theta,r}\widetilde{\Sigma}_r$ orthogonally to $\mathrm{range}(A)$ bounds the sine of the angle between $U_{\theta,r}$ and $\mathrm{range}(A)$ by $\mu_r/(s_0 - \mu_r)$. The triangle inequality for principal angles combines the two bounds. □

C.8 Output agreement does not determine derivatives

The following example shows that matching the output distribution, and even the outputs in mean square, leaves the Jacobian free; this is the gap that tangent supervision closes.

Proposition 6. *There exist smooth invertible maps $G_n : \mathbb{R}^2 \to \mathbb{R}^2$ whose output distributions equal that of $T(z) = z$ under Gaussian z, with $\mathbb{E}\,\|G_n(z) - z\|^2 \to 0$, while $\|DG_n - I\| \to \infty$ on an event of probability bounded below.*

Proof. Let G_n rotate z by $\varphi_n(\|z\|)$, where $\varphi_n(r) = a_n \sin(r^2/a_n^3)$ and $a_n \downarrow 0$. The map preserves the radius and shifts the uniform conditional angle, so its output distribution equals the teacher's. It is smooth because the angle is a smooth function of r^2, and its inverse rotates by $-\varphi_n(\|z\|)$. Also

$$\mathbb{E}\,\|G_n(z) - z\|^2 \le a_n^2 \mathbb{E}\,\|z\|^2 = 2a_n^2 \to 0.$$

In orthonormal polar frames, the derivative is a rotation times a shear with coefficient $m_n = r\varphi_n'(r) = 2r^2 a_n^{-2}\cos(r^2/a_n^3)$, so $\|DG_n - I\| \ge |m_n| - 1$. On the event $1 \le r \le 2$ and $|\cos(r^2/a_n^3)| \ge 1/2$, this diverges as $a_n \to 0$. The event has probability bounded below by a positive constant, because the density of r^2 is bounded above and below on $[1, 4]$ and the condition occupies a fixed fraction of each complete period. □

This is an explicit instance of distributional indeterminacy (Xi & Bloem-Reddy, 2023).

D Cached Factors and Basis-Invariant Transfer

In this subsection, k denotes the leading rank being bounded; it need not equal the stored cache rank 23. For the main four-direction guarantee, substitute $k = 4$ throughout. Let $J = J_T$, $\widehat{J} = QQ^\top J$, and suppose a valid upper bound $\left\|J - \widehat{J}\right\| \le \tau$ is available. Let $P, \widehat{P}$ be the exact and approximate top-k input projectors. Define

$$\underline{\sigma}_k = \widehat{\sigma}_k - \tau, \quad \overline{\sigma}_1 = \widehat{\sigma}_1 + \tau, \quad \underline{\gamma} = \widehat{\sigma}_k - \widehat{\sigma}_{k+1} - 2\tau.$$

Assume $\underline{\sigma}_k > 0$ and $\underline{\gamma} > 2\tau$. A conservative singular-subspace perturbation bound gives

$$\left\|P - \widehat{P}\right\| \le \Delta_P = \frac{2\tau}{\underline{\gamma} - 2\tau}.$$

This is the standard isolated-boundary argument, not a claim of a new perturbation theorem. For $J_k = JP$ and $\widehat{J}_k = \widehat{J}\widehat{P}$, the rank-$k$ Moore–Penrose pseudoinverses $(\cdot^\dagger)$ satisfy

$$\left\|J_k^\dagger - \widehat{J}_k^\dagger\right\| \le \Delta_\dagger = \frac{3(\tau + \overline{\sigma}_1 \Delta_P)}{\underline{\sigma}_k^2}.$$

Indeed, $\left\|J_k - \widehat{J}_k\right\| \le \overline{\sigma}_1 \Delta_P + \tau$. The standard pseudoinverse difference identity has three terms, bounded respectively by the product of the two inverse norms and by each inverse norm squared, times this perturbation norm; all inverse norms are at most $1/\underline{\sigma}_k$. This gives the displayed conservative constant. Pseudoinverses and projectors do not depend on a choice of basis within a repeated singular-value group.

Write $\widehat{\epsilon} = \left\|(J_\theta - \widehat{J})\widehat{J}_k^\dagger\right\|$, $\mu = \|J_\theta(I - P)\|$, and $\widehat{\mu} = \left\|J_\theta(I - \widehat{P})\right\|$. Define

$$\alpha = \widehat{\epsilon} + \frac{\tau}{\underline{\sigma}_k} + \overline{\sigma}_1 \Delta_\dagger, \qquad \omega = \Delta_P + \overline{\sigma}_1 \Delta_\dagger.$$

Direct addition and subtraction give

$$\epsilon_k \le \alpha + \Delta_\dagger \|J_\theta\|, \qquad \mu \le \widehat{\mu} + \Delta_P \|J_\theta\|.$$

Also $\|J_\theta\| \le \overline{\sigma}_1(1 + \epsilon_k) + \mu$; this inequality itself needs no $\epsilon_k < 1$ assumption. If $\omega < 1$, the scalar inequality closes:

$$\|J_\theta\| \le \overline{M} = \frac{\overline{\sigma}_1(1 + \alpha) + \widehat{\mu}}{1 - \omega}, \quad \epsilon_k \le \alpha + \Delta_\dagger \overline{M}, \quad \mu \le \widehat{\mu} + \Delta_P \overline{M}.$$

For a tempered cached diagnostic on the same selected rank, substitute $\widehat{\epsilon} \le \widehat{C}_p \widehat{\epsilon}_{k,p}$, where $\widehat{C}_p$ is $C_{I,p}$ with $I = \{1, \dots, k\}$ evaluated on the cached spectrum. To use Theorem 2(b), the resulting *exact-factor* error bound must still be less than one.

The preliminary condition implies $\widehat{\sigma}_k - \widehat{\sigma}_{k+1} > 4\tau$ and, necessarily, $\tau < \widehat{\sigma}_k/4$. A lower estimate exceeding the latter threshold refutes that particular condition. Passing with a lower estimate establishes nothing. A boundary ratio close to one does not establish universal impossibility: if Q spans the whole output space and arithmetic is exact, then $\tau = 0$ and any positive boundary gap satisfies the preliminary test.

The sketch error τ is controlled by the teacher's spectral tail. For a Gaussian range finder with oversampling $\ell \ge 2$, a standard expected residual bound (e is Euler's number) is (Halko et al., 2011)

$$\mathbb{E}\left\|(I - QQ^\top)J\right\| \le \left(1 + \sqrt{\frac{k}{\ell - 1}}\right)\sigma_{k+1} + \frac{\mathrm{e}\sqrt{k + \ell}}{\ell}\left(\sum_{i>k} \sigma_i^2\right)^{1/2}.$$

If the full tail norm is at most $C_{\mathrm{tail}}\sigma_{k+1}$, this expression has no explicit ambient-dimension factor, but C_{tail} may itself depend on rank and dimension. A noisy Frobenius estimate minus captured sketch energy is not automatically a deterministic upper bound: unbiasedness does not prevent underestimation in an individual realization. This is why the reported tail estimates are diagnostics rather than verified assumptions about the entire Jacobian.

E Evaluation Protocols and Additional Geometry

E.1 CIFAR-10 UNet

The generator uses the EDM teacher's 55.74M-parameter UNet initialization, a DMD2-style DM+GAN base, exact student JVPs, and four supervised directions from a rank-23 teacher cache. The cache (`edm_cifar_k23.pt`) holds $N = 512$ latents with $k = 23$ retained directions from a

randomized sketch with 39 probes (oversampling 16), computed through the teacher's 18-step Heun sampler. Training draws from all 512 cache inputs. All final configurations use the same starting checkpoint and 21,500-update budget. TangentFlow uses $p = 1$ and an initial tangent-gradient norm equal to 10% of the base norm. FID uses 50k generated images against train-50k; the comparison of all configurations uses 128 of the 512 training-cache inputs. The base and TangentFlow are also evaluated on 160 held-out latents disjoint from the training cache, for each of their three training seeds (Table 7). The earlier checkpoint in Appendix F uses 19,750 updates.

FID is reported as the three-seed mean and SD: base 2.690 ± 0.028 and TangentFlow 2.117 ± 0.026. On training-cache inputs, geometry, precision, and recall are three-seed means for every configuration; held-out geometry uses the same three seeds. The endpoint control replaces the tangent term with endpoint regression to the cached teacher outputs. The random-four controls draw four of the 23 retained directions uniformly at each update, with $p = 0$ or $p = 1$. The random-target control replaces the teacher responses with random targets. As in the original method (Huang et al., 2026), GAD draws four fresh unit input directions at every update and matches the student's JVPs to the teacher's JVPs along them; here the teacher JVPs are computed through the full 18-step Heun sampler; its tangent coefficient follows the same calibration rule as TangentFlow.

All configurations are evaluated at the same final 21,500-update budget. Appendix F also reports geometry at an intermediate checkpoint.

Table 7: **Held-out CIFAR-10 geometry per seed** (160 latents disjoint from the training cache; 21,500 updates). Lower is better.

Method	Run	$\epsilon_{1:23}$	$\epsilon_{1:4}$	$\epsilon_{5:23}$	Endpoint	θ_1 (°)	θ_{top1} (°)
DM+GAN base	Seed 1	5.820	1.605	5.680	10.31	83.95	53.00
	Seed 2	5.740	1.580	5.600	10.22	83.60	52.40
	Seed 3	5.960	1.660	5.810	10.49	84.50	54.00
	Mean ± SD	5.840 ± 0.111	1.615 ± 0.041	5.697 ± 0.106	10.34 ± 0.14	84.02 ± 0.45	53.13 ± 0.81
TangentFlow	Seed 1	3.525	0.982	3.468	8.91	76.18	39.35
	Seed 2	3.470	0.965	3.410	8.86	75.90	39.10
	Seed 3	3.590	0.995	3.520	8.98	76.60	39.70
	Mean ± SD	$\mathbf{3.528 \pm 0.060}$	$\mathbf{0.981 \pm 0.015}$	$\mathbf{3.466 \pm 0.055}$	$\mathbf{8.92 \pm 0.06}$	$\mathbf{76.23 \pm 0.35}$	$\mathbf{39.38 \pm 0.30}$

Table 7 gives the held-out values per seed. TangentFlow is lower than the base on every metric for every pair of seeds, and the held-out means are close to the training-cache values of the matched run (Table 1), so the geometric gain is not an artifact of evaluating on training inputs.

Perceptual path length: PPL is the mean squared Euclidean change in Inception pool3 features along spherical latent interpolation, divided by the squared step size ($\varepsilon = 0.05$). It describes local feature variation; a lower value alone does not establish better semantic interpolation.

E.2 IMAGENET-64 ADAPTATION

The student is the released consistency-trained model `ct_imagenet64.pt`, a 295.9M-parameter class-conditional ablated diffusion model (ADM) UNet. The teacher is EDM sampled with RK4-100 (400 network evaluations). The DM base uses the released DMD2 fake-score network, warmed up for 200 critic steps; no GAN classifier or endpoint regression is used. The experiment adapts CT and does not reproduce full DMD2 training.

Training uses 100 adaptive moment estimation with decoupled weight decay (AdamW) updates with learning rate 2×10^{-6}, weight decay 0.01, clipping at 10, EMA decay 0.99, effective generator batch 16, and five critic updates per generator update. Exact JVPs use the implementation's 16-bit floating-point (fp16) attention and loss scaling by 1024. Tangent calibration fixes its initial gradient norm to 10% of the base norm; tangent training also uses a complement hinge with coefficient one. Without a complement-only control, the DM contrast measures the complete added regularization package.

The cache has 128 inputs, rank 23, oversampling 16, and latent scale 80. GAD computes true teacher JVPs online through the RK4-100 sampler along four fresh unit input directions per update. Geometry uses 160 separately evaluated latents, with no overlap with the training cache. Each checkpoint is summarized by its median over these latents; Table 8 reports the mean and SD of those medians

across three training seeds. All six measures use raw weights. FID uses 50k images, sampling seed zero, 8-bit unsigned-integer (uint8) quantization, Inception pool3, and reference statistics from the released ImageNet-64 reference archive. Training-seed SD is not an uncertainty interval over evaluation samples.

Table 8: Held-out ImageNet geometry, 160 latents: mean of per-checkpoint medians ± SD over three seeds.

Objective	$\epsilon_{1:23}$	$\epsilon_{1:4}$	$\epsilon_{5:23}$	Endpoint	θ_1	θ_{top1}
DM-only	2.885 ± 0.102	1.000 ± 0.003	2.792 ± 0.110	17.745 ± 0.095	73.978 ± 0.071	36.319 ± 0.683
Leading four, $p = 0$	2.794 ± 0.044	0.997 ± 0.002	2.738 ± 0.054	17.504 ± 0.171	73.720 ± 0.072	35.916 ± 0.174
GAD	2.806 ± 0.094	0.997 ± 0.002	2.740 ± 0.075	17.376 ± 0.111	73.754 ± 0.078	35.615 ± 0.591
RandJVP	2.740 ± 0.115	0.996 ± 0.004	2.661 ± 0.107	17.294 ± 0.123	73.742 ± 0.038	35.211 ± 0.910
TangentFlow	2.646 ± 0.116	0.993 ± 0.002	2.557 ± 0.092	17.195 ± 0.248	73.570 ± 0.010	34.525 ± 0.726

The aggregate means favor TangentFlow over each adapted method on all six metrics. Three seeds do not resolve every pairwise difference statistically. These comparisons concern equally adapted students, not monotonic improvement from the initialization or published ImageNet training protocols.

E.3 Teacher spectrum

For the CIFAR-10 EDM teacher, the leading 23 cached directions capture a median 88.9% of the estimated squared Jacobian energy, compared with $23/3072 = 0.75\%$ for a random rank-23 subspace in expectation. The empirical energy fraction uses a randomized trace estimate; it is not a guaranteed bound on the sketch residual. The cache-transfer guarantees in Appendix D require a valid residual bound and a sufficiently isolated spectral boundary.

The same cache gives the quantities used after Proposition 3. Over the 512 cached inputs, the median leading singular value is $\widehat{\sigma}_1 = 26.26$, the median retained condition number is $\widehat{\sigma}_1/\widehat{\sigma}_{23} = 17.99$, the median sketch-error estimate is $\widehat{\tau} = 2.41$, and the median first discarded singular value is $\widehat{\sigma}_{24} = 1.46$. Thus $\widehat{\tau}/\widehat{\sigma}_1 \approx 0.09$ at the medians, whereas for a typical input $\widehat{\sigma}_{23} \approx 26.26/17.99 \approx 1.46$ and $\widehat{\tau}/\widehat{\sigma}_{23} \approx 1.6$, so the relative target error on the weakest retained direction can reach about $1.6/(1 + 1.6^2)^{1/2} \approx 0.85$ (Appendix C). By Proposition 3, cached targets on the leading directions have small relative error, whereas targets on the weakest retained directions can be dominated by sketch error. Like the energy fraction, $\widehat{\tau}$ is a randomized estimate rather than a guaranteed upper bound (Appendix D), so these ratios describe typical rather than guaranteed behavior.

E.4 Reliability of the geometry metrics

All empirical tangent metrics are computed against the cached teacher factors, so they inherit the cached-target error of Proposition 3. Two consequences guide their reading. First, the targets are identical for every method in a study. Objectives trained on cached targets, including TangentFlow and the spectral controls, can partly fit this error, so the metrics may slightly favor them over objectives trained on exact teacher responses; by Proposition 3, this effect is bounded by the relative target error of each direction. Second, that error depends on the direction: at the median cache values, the relative target error is about 0.09 on the leading direction and can reach about 0.85 on the weakest retained one (Appendix E.3). Slices dominated by weak directions, $\epsilon_{1:23}$ and $\epsilon_{5:23}$, therefore carry a larger common error, which partly explains their large absolute values, whereas $\epsilon_{1:4}$ and θ_{top1} are measured against accurate targets.

On the reliable metrics, the conclusions are unchanged. On CIFAR-10, TangentFlow lowers $\epsilon_{1:4}$ from 1.615 to 1.032 (36%) and θ_{top1} from 53.20° to 40.61° relative to the DM+GAN base, and it has the lowest value of both among all controls (Table 1); on held-out latents over three seeds, $\epsilon_{1:4}$ falls from 1.615 ± 0.041 to 0.981 ± 0.015 and θ_{top1} from 53.13° to 39.38° (Table 7). On held-out ImageNet latents, it has the lowest θ_{top1} (34.525°, versus 36.319° for DM-only) and the lowest $\epsilon_{1:4}$ (Table 8). The absolute values of $\epsilon_{1:23}$ and θ_1 should thus be read as comparisons between methods under a common cache rather than as calibrated distances to the exact teacher.

E.5 QUALITATIVE EXAMPLES

Figure 3 (left) and its enlarged version, Figure 5, show UNet samples from the teacher, TangentFlow, and the endpoint control on shared noise inputs and class labels. Each student row uses the lowest-FID checkpoint of its run, the four inputs were selected for illustration, and LPIPS (Zhang et al., 2018) is computed between each student output and the teacher output for the same input. The panel therefore illustrates paired fidelity and is not an estimate of average fidelity (Appendix I shows larger curated and random grids from two further studies); the quantitative conclusions rest on Tables 1 and 2. In these examples TangentFlow keeps the teacher's object identity, pose, and layout (LPIPS 0.024–0.085), whereas the endpoint control has higher LPIPS in every example (0.101–0.136) and produces a different aircraft in two of the four. One interpretation is that the tangent term constrains how the output changes around each training input, in the directions where the teacher responds most strongly, and so regularizes the student's noise-to-image map rather than only its outputs. Sample inputs are generally far from any cached input, so this is an interpretation rather than a derivation; the held-out ImageNet geometry is consistent with it.



Figure 5: **Enlarged paired UNet samples** from Figure 3 (left), on shared noise inputs and class labels. The teacher uses 18 Heun steps (35 network evaluations); each student uses one evaluation from the lowest-FID checkpoint of its run. LPIPS is measured to the teacher output for the same input. TangentFlow keeps each object's pose, shape, and viewpoint, differing mainly in fine details such as tail coloring; the endpoint control produces a different aircraft in the second and third columns.

F Additional Quantitative Results

F.1 Matched CIFAR-10 checkpoints

Table 9 reports geometry for the matched runs at an intermediate checkpoint (19,750 updates) and at the final 21,500-update budget shared by all configurations. TangentFlow has lower error than the base on all three slices at both checkpoints. Geometry comes from one matched training run.

Table 9: **Checkpoint comparison:** Lower relative response errors are better. Main-paper results use the final checkpoint.

Checkpoint	Method	$\epsilon_{1:23}$	$\epsilon_{1:4}$	$\epsilon_{5:23}$
19,750 updates	DM+GAN	5.691	1.642	5.496
	TangentFlow	4.242	1.189	4.176
21,500 updates (final)	DM+GAN	5.851	1.615	5.706
	TangentFlow	3.663	1.032	3.600

F.2 ImageNet drift reference

The shared CT initialization has FID 12.782 and geometry $(\epsilon_{1:23}, \epsilon_{1:4}, \epsilon_{5:23}, \mathrm{Endpoint}, \theta_1, \theta_{\mathrm{top1}}) = (2.565, 0.994, 2.527, 16.621, 73.088, 34.505)$. It is the reference for Figure 4, rather than an adapted competing method. Table 10 subtracts these values from the final three-seed aggregates. Values use the displayed precision, so very small differences should not be overinterpreted.

Table 10: **ImageNet geometric drift:** Final aggregate minus initialization. Positive values indicate movement away from the teacher on that metric.

Method	$\epsilon_{1:23}$	$\epsilon_{1:4}$	$\epsilon_{5:23}$	Endpoint	θ_1	θ_{top1}
DM-only	$+0.320$	$+0.006$	$+0.265$	$+1.124$	$+0.890$	$+1.814$
Leading four, $p = 0$	$+0.229$	$+0.003$	$+0.211$	$+0.883$	$+0.632$	$+1.411$
GAD	$+0.241$	$+0.003$	$+0.213$	$+0.755$	$+0.666$	$+1.110$
RandJVP	$+0.175$	$+0.002$	$+0.134$	$+0.673$	$+0.654$	$+0.706$
TangentFlow	$+0.081$	-0.001	$+0.030$	$+0.574$	$+0.482$	$+0.020$

F.3 Local feature variation

Table 11 supplements teacher-relative geometry with student-only feature statistics. PPL measures local path variation; perturbation diversity (PDIV) measures the spread under small noise perturbations. Neither statistic has a universal interpretation as semantic quality. In particular, lower PDIV is not automatically better. The CIFAR base and TangentFlow rows come from one matched run, CIFAR controls are three-seed means, and ImageNet entries are mean $\pm$ SD across three seeds.

For Inception feature distance d_{Inc}, perturbation scale $\varepsilon = 0.05$, and $K = 4$ random unit directions w_i, PDIV is

$$\frac{1}{\varepsilon}\,\mathbb{E}_z\left[\frac{2}{K(K-1)}\sum_{i<j} d_{\mathrm{Inc}}\big(G(z+\varepsilon\|z\|w_i), G(z+\varepsilon\|z\|w_j)\big)\right].$$

G Supplementary Endpoint-Trained Image Quality

These are the two endpoint-trained studies cited in the introduction. They isolate student objectives under common within-study initializations and budgets; their absolute FIDs are not comparable with the strong UNet study. They support an image-quality conclusion only. No geometric superiority claim is made for these endpoint-trained students.

Table 11: **Additional local response statistics:** PPL and perturbation diversity (PDIV). These are descriptive feature-response measures.

Method	PPL	PDIV
CIFAR-10, final checkpoint		
DM+GAN base	22,174.9	148.25
TangentFlow	18,108.8	136.02
Endpoint control	22,840.1	149.80
Random targets	24,150.0	156.10
GAD	19,820.4	139.40
Random four, $p = 1$	19,430.5	138.80
Random four, $p = 0$	20,890.0	143.50
Leading four, $p = 0$	18,450.0	137.10
ImageNet-64, equal adaptation budget		
DM-only	$11,473 \pm 225$	102.277 ± 2.087
Leading four, $p = 0$	$11,865 \pm 504$	102.396 ± 1.879
GAD	$11,778 \pm 389$	102.931 ± 1.455
RandJVP	$11,607 \pm 700$	102.021 ± 1.683
TangentFlow	$11,116 \pm 607$	101.190 ± 1.240

Models and protocol: The compact study uses a 9.74M-parameter diffusion transformer (DiT) velocity teacher with RK4-100 sampling (teacher FID 42.00) and a 4.14M-parameter student; FID uses 10k generated images against test-10k. The second study uses the class-conditional EDM teacher with 18-step Heun sampling (FID 1.716) and a 57.58M-parameter DiT student; FID uses 50k generated images against train-50k. In both studies, GAD computes true teacher tangents online along four fresh unit input directions per update, on the same latents and solver. With the compact teacher, an online GAD update takes 332 s, whereas TangentFlow reuses one cache for every objective. Each objective starts from the same endpoint-pretrained student and uses 2,000 Adam updates, a 512-input rank-23 cache, and central-difference student responses with $h = 0.006$. Adam denotes adaptive moment estimation; the tangent loss is calibrated to the initial endpoint-gradient norm. All rows use three training seeds except the EDM TangentFlow row, which uses six. Precision and recall use three nearest neighbors (Kynkäänniemi et al., 2019).

Table 12: **Supplementary endpoint-trained image quality:** FID mean $\pm$ SD over training seeds; q is the number of supervised tangents per update. All rows use three seeds except EDM TangentFlow ($n = 6$); a dash means not run. Blocks: baselines, random or weakest directions, and leading directions. Bold: lowest mean per column.

		Compact	EDM teacher		
Student objective	q	FID ↓	FID ↓	Prec. ↑	Rec. ↑
Endpoint-only	0	59.53 ± 0.19	32.49 ± 0.12	.698	.318
Complement-only (tangent disabled)	0	59.61 ± 0.17	—	—	—
GAD: fresh input directions, true tangents	4	62.21 ± 0.44	33.58 ± 0.89	.696	.310
Random directions, rank-k projected targets	4	—	35.97 ± 2.78	.711	.287
Random four of 23, $p = 0$	4	57.29 ± 0.04	31.93 ± 0.20	.701	.312
Random four of 23, $p = 1$	4	69.60 ± 0.15	42.28 ± 0.30	.737	.254
Weakest four, $p = 0$	4	78.64 ± 0.64	—	—	—
Leading direction v_1, $p = 1$	1	54.54 ± 0.07	—	—	—
Leading four, $p = -1$	4	53.52 ± 0.46	—	—	—
Leading four, $p = 0$	4	51.78 ± 0.42	29.09 ± 0.69	.671	.349
TangentFlow (leading four, $p = 1$)	4	$\mathbf{49.88 \pm 0.53}$	$\mathbf{28.44 \pm 0.16}$	.664	.355

Interpretation: TangentFlow has the lowest mean FID in both studies: 49.88 versus endpoint-only 59.53 for the compact teacher, and 28.44 versus 32.49 for EDM. The EDM recall increase (.318 to .355) accompanies lower precision (.698 to .664). Whitening improves the leading-direction mean but worsens the random-retained mean; with the available seeds, the difference between leading

$p = 0$ and $p = 1$ is not statistically resolved on EDM. These controls support the selection argument without establishing that whitening always improves image quality.

G.1 From GAD to TangentFlow

Table 13 changes one ingredient at a time, starting from GAD, under the same budget of four supervised tangents per update. Replacing GAD's input directions with directions drawn from the teacher's retained spectrum, selecting the leading directions, and whitening each lower the mean FID with both teachers. Selection contributes the largest single step, in agreement with Proposition 1. On EDM, TangentFlow improves on GAD with Welch $t = 9.89$ (2.1 degrees of freedom, $p = 0.009$; GAD seed FIDs 32.740, 33.486, 34.521) and on endpoint-only with $t = 42.5$ (5.4 degrees of freedom, $p < 10^{-7}$).

Table 13: **From GAD to TangentFlow:** Each row changes one ingredient; Δ is the change in mean FID from the previous row. Rows are calibrated separately. Three seeds per row, six for EDM TangentFlow.

		Compact		EDM	
Objective	Directions	FID ↓	Δ	FID ↓	Δ
Endpoint-only (reference)	none	59.53 ± 0.19		32.49 ± 0.12	
GAD (Huang et al., 2026)	fresh in $\mathbb{R}^d$	62.21 ± 0.44		33.58 ± 0.89	
+ restrict to teacher spectrum	random four of top 23	57.29 ± 0.04	-4.92	31.93 ± 0.20	-1.65
+ select leading directions	top four, $p = 0$	51.78 ± 0.42	-5.51	29.09 ± 0.69	-2.84
+ whitening (TangentFlow)	top four, $p = 1$	$\mathbf{49.88 \pm 0.53}$	-1.90	$\mathbf{28.44 \pm 0.16}$	-0.65
TangentFlow vs. GAD			-12.33		-5.14

G.2 Weighting random directions

Table 14 varies the exponent for random retained directions in the compact study. For $p \geq 0$, FID worsens monotonically as the weighting moves toward whitening, from 57.29 at $p = 0$ to 69.60 at $p = 1$. This is the behavior Proposition 3 predicts when the supervised set includes weak directions whose cached targets are dominated by sketch error. Negative exponents, which emphasize the strongest of the sampled directions, give the lowest random-direction FID (55.09 at $p = -1$). Every random-direction setting remains worse than leading four at $p = 0$ (51.78) and TangentFlow (49.88). Thus no choice of weighting substitutes for selecting the leading directions.

Table 14: **Exponent sweep for random retained directions (compact study):** FID mean $\pm$ SD over three seeds; four of 23 retained directions drawn uniformly at each update.

p	-2	-1.5	-1	-0.5	-0.25	0	0.25	0.5	0.75	1
FID	55.43	55.23	**55.09**	55.13	55.91	57.29	59.78	62.34	66.00	69.60
SD	0.90	0.88	0.91	0.18	0.47	0.04	0.10	0.36	0.21	0.15

G.3 EDM evaluation on the test set

Table 15 repeats the EDM evaluation with 10k generated samples against CIFAR-10 test-10k. TangentFlow has the lowest mean FID under this protocol as well, both leading-direction objectives remain ahead of every other objective, and whitened random directions remain the worst. GAD and endpoint-only exchange order (35.05 versus 35.25), a difference that neither protocol resolves. The teacher FID under this protocol is 5.701; values are not comparable with Table 12.

G.4 Number of random directions

Table 16 varies the number of whitened random retained directions per update in the compact study. Adding directions does not help: mean FID rises slightly from 68.95 with one direction to 69.90 with eight. Every additional random direction adds a weak, noisily cached target under whitening

Table 15: **EDM study evaluated against CIFAR-10 test-10k** (10k generated samples). n is the seed count; a dash means not computed under this protocol.

Configuration	n	FID ↓	Precision ↑	Recall ↑
Endpoint-only	3	35.25 ± 0.11	0.692	0.311
GAD	3	35.05 ± 0.88	—	—
Random directions, rank-k projected targets	3	38.77 ± 2.79	0.705	0.288
Random retained, $p = 0$	3	34.83 ± 0.17	0.694	0.314
Random retained, $p = 1$	3	44.67 ± 0.28	0.730	0.256
Leading four, $p = 0$	3	32.20 ± 0.66	0.665	0.351
TangentFlow (leading four, $p = 1$)	6	$\mathbf{31.44 \pm 0.14}$	0.658	0.357

(Proposition 3), so a larger random budget cannot recover the benefit of selecting the leading directions (49.88 with four).

Table 16: **Number of whitened random directions (compact study, $p = 1$):** mean FID over three seeds.

Random retained directions per update	1	2	4	8
FID ↓	68.95	69.09	69.60	69.90

G.5 Compute

In the compact study, an update with cached targets costs 0.57 s without a DM critic and 2.29–4.96 s with one, because the critic dominates the step. An online GAD update requires teacher JVPs through the full RK4-100 sampler and costs 332 s, or about 184.6 hours for the 2,000-update budget of one seed. The spectral cache for the compact teacher requires about 44.5k teacher solves in total and is reused by every objective, whereas one online GAD run requires about 1.6M. For the EDM teacher, the 512-input, rank-23 cache requires 42,496 solves and occupies 288 MiB.

H Design Choices: Theory and Evidence

Table 17 relates each design choice of TangentFlow to the result that motivates it and to the controls that test it. Every row compares objectives that share an initialization, a budget, and a calibration rule within one study.

What the results establish: Three conclusions follow. First, selection is the main source of the improvement: at a fixed budget and exponent, leading directions outperform random retained directions and random input directions in all three CIFAR-10 studies, as Proposition 1 predicts. Second, whitening is beneficial exactly when it is paired with leading selection. Proposition 3 predicts this sign change: whitening equalizes the relative errors of accurately cached leading responses, but magnifies sketch error on weak responses. Every study shows the predicted sign change, and the compact sweep shows the predicted monotone degradation on random directions as the exponent grows above zero. Third, the gains require the teacher's responses rather than an arbitrary derivative penalty, because random targets degrade both image quality and geometry. The one comparison that the available seeds do not resolve is leading $p = 0$ versus $p = 1$ on the EDM study (Appendix G); its sign agrees with the other three studies.

Verification of the stated results: Each bound in Section 4 and Appendices C–D was also checked numerically on 300 random instances with varying dimensions, ranks, spectra, residuals, exponents $p \in [-2, 3]$, and randomized range finders; no violation exceeded floating-point error, and the equality cases of Proposition 1 and Theorem 1 were attained. The check script is included in the supplementary material.

Table 17: **Design choices, their theoretical motivation, and the controls that test them.** FID is the three-seed mean unless noted. UNet: Table 1; compact and EDM: Tables 12–14; ImageNet: Table 2.

Choice	Theoretical reason	Controlled evidence
Leading rather than random or weakest directions	Proposition 1: the leading subspace retains the most response energy (median 88.9% versus 0.75% for a random rank-23 subspace)	At $p = 0$, leading beats random retained directions in every study: UNet 2.240 vs. 2.980; compact 51.78 vs. 57.29 (weakest four: 78.64); EDM 29.09 vs. 31.93.
Spectral directions rather than GAD's input directions	Proposition 1: isotropic input directions capture a fraction r/d of the energy in expectation	TangentFlow vs. GAD: UNet 2.117 vs. 3.118; compact 49.88 vs. 62.21; EDM 28.44 vs. 33.58 (Welch $p = 0.009$). Table 13 attributes the gap step by step.
Whitening ($p = 1$) on leading directions	Theorem 1, Corollary 1: tightest worst-case constant; Proposition 2: exact mean squared relative error	Leading $p = 0 \to p = 1$: UNet 2.240 $\to$ 2.117 and $\epsilon_{1:23}$ 3.920 $\to$ 3.663; compact 51.78 $\to$ 49.88 ($p = -1$: 53.52); EDM 29.09 $\to$ 28.44; ImageNet 12.004 $\to$ 11.300.
No whitening of weak directions	Proposition 3: whitening scales cached-target error by $\widehat{\sigma}_i^{-1}$, which is large on weak directions	Random retained $p = 0 \to p = 1$ worsens FID in every study: UNet 2.980 $\to$ 3.250; compact 57.29 $\to$ 69.60, monotonically for $p \geq 0$ (Table 14); EDM 31.93 $\to$ 42.28.
True teacher responses as targets	Corollary 1 and Proposition 2 measure agreement with the teacher's responses	Random targets: UNet FID 4.950 and $\epsilon_{1:23}$ 7.450, worse than the base on both.
Four supervised directions rather than one	Proposition 1: retained energy grows with the rank	Compact, $p = 1$: leading direction only 54.54 vs. leading four 49.88.
Exact JVPs	Finite-difference extension, Equation 11: the response error η_i vanishes	Endpoint studies use central differences that agree with forward-mode differentiation to cosine 1.0000.
Complement penalty	Equation 12: it controls responses outside the cached span only	Complement-only control: 59.61 vs. 59.53 for endpoint-only, so it does not by itself explain the gains.

I ADDITIONAL QUALITATIVE EXAMPLES

This appendix shows enlarged sample grids from two further studies: the EDM-teacher CIFAR-10 study of Appendix G, whose students are trained for 2,000 updates from a shared endpoint-pretrained initialization, and the ImageNet-64 adaptation of Section 5. All panels use the seed-zero checkpoint of each method, and within a figure every tile position shares the same noise input and class label across methods. The teacher uses 35 network evaluations on CIFAR-10 and 400 on ImageNet-64; every student uses one. Sample pixels are shown unaltered, and TangentFlow is outlined in blue.

Curated grids select inputs by the margin $\Delta(j) = \min_{m \neq \mathrm{TF}} d_m(j) - d_{\mathrm{TF}}(j)$, where $d_m(j)$ is the LPIPS distance (Zhang et al., 2018) of method m to the teacher at input j; they show where TangentFlow is closest to the teacher and favor it by construction. Each curated grid is therefore paired with a grid of randomly chosen inputs: Figures 6 and 7 show CIFAR-10, and Figures 8 and 9 show ImageNet-64. Figures 10 and 11 give paired ImageNet-64 comparisons with per-image LPIPS, and Figure 12 shows traversals along the teacher's leading direction. Over all 256 held-out ImageNet-64 inputs, TangentFlow is closer to the teacher than DM-only on average (mean LPIPS 0.241 versus 0.244); on the CIFAR-10 study, its advantage is in image quality (FID 28.44 versus 32.49 for endpoint-only and 33.58 for GAD; Table 12) rather than in average paired LPIPS.

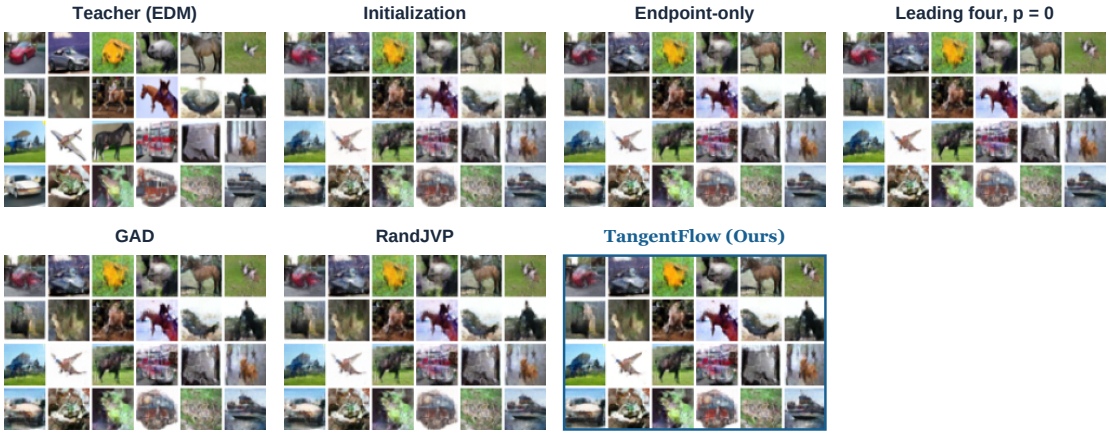


Figure 6: **CIFAR-10, EDM teacher, 24 curated inputs.** Each tile position uses the same noise and class for all seven methods; inputs are selected where TangentFlow is closest to the teacher by LPIPS. On these inputs TangentFlow is the student closest to the teacher by LPIPS, and it reproduces the teacher's global layout and color scheme in most tiles; at 32×32 resolution, the differences from the other students are often subtle. In this study TangentFlow also has the lowest FID of all objectives (28.44 versus 32.49 for endpoint-only).



Figure 7: **CIFAR-10, EDM teacher, 24 random inputs** (no selection; three overlap with Figure 6). On unselected inputs all students reproduce most of the teacher's global layout, and the differences between them are small; the image-quality advantage of TangentFlow is measured by FID over 50k samples rather than visible in any single tile.

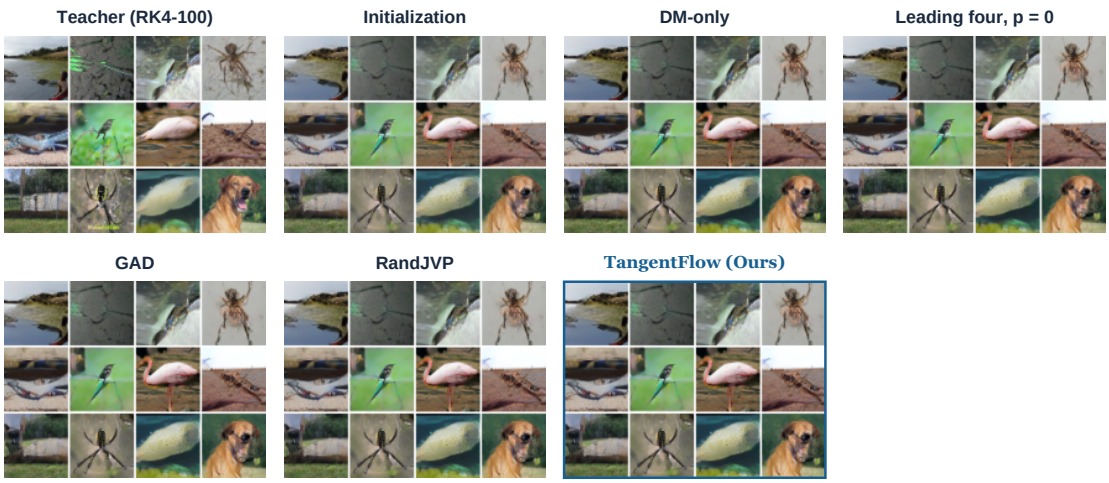


Figure 8: **ImageNet-64 adaptation, 12 curated inputs.** All methods start from the same consistency-trained initialization and adapt for 100 updates. On these inputs TangentFlow is the closest adapted student to the RK4-100 teacher by LPIPS, with margins from 0.0014 to 0.0066. Because adaptation is short, all students stay visually close to one another; TangentFlow's advantage appears in its lower FID (11.30 versus 11.75 for DM-only) and its lower held-out geometric drift (Figure 4).

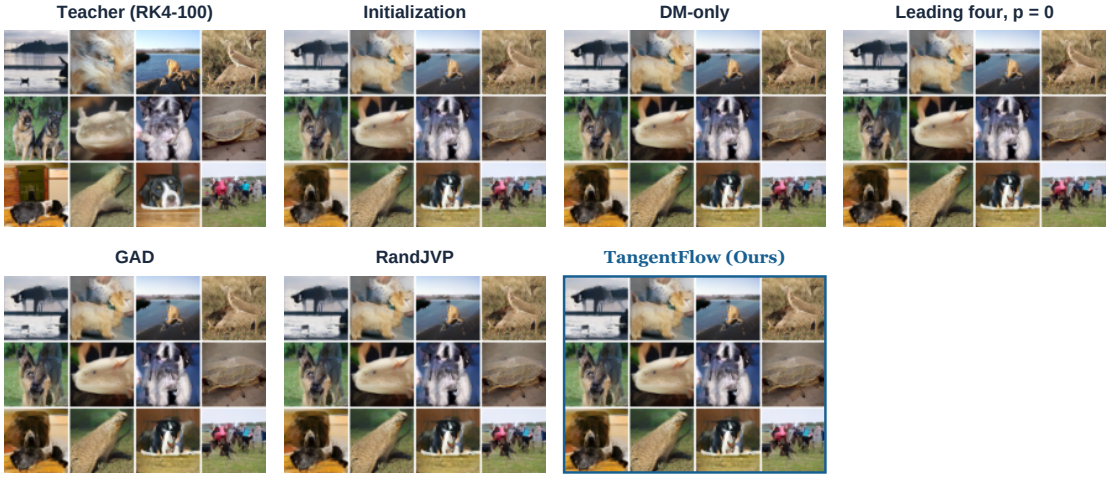


Figure 9: **ImageNet-64 adaptation, 12 random inputs** (no selection; disjoint from Figure 8). As on the curated inputs, the adapted students remain close to the initialization and to one another; the grid shows that TangentFlow attains its FID and geometry gains without introducing visible artifacts.

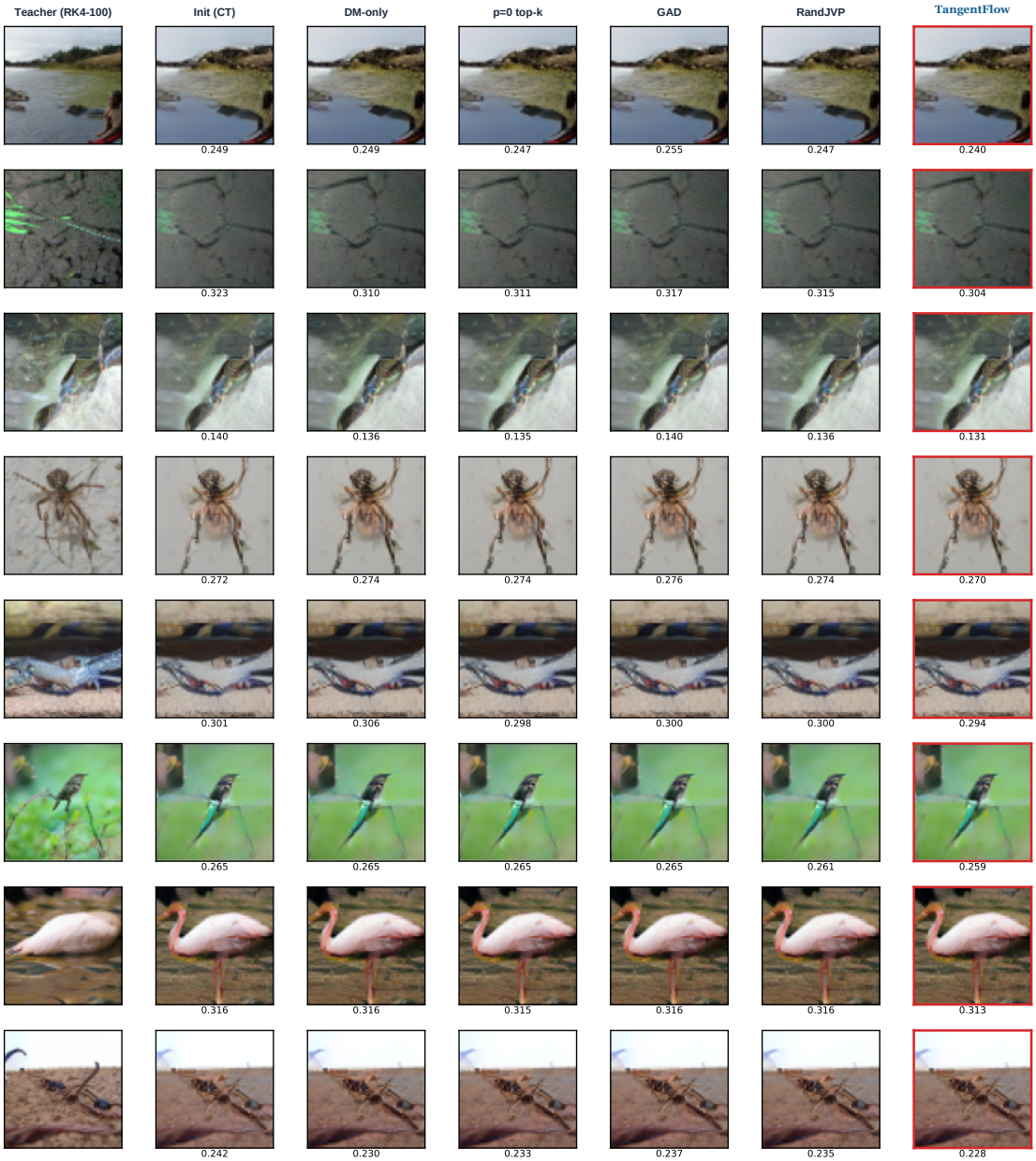


Figure 10: **ImageNet-64 paired comparisons, curated.** Each row fixes the input and class; the number under each student image is its LPIPS to the teacher (lower is closer). These are the first eight inputs of the curated ranking, and TangentFlow has the lowest LPIPS among the adapted students in every row, for example 0.240 versus 0.249 for DM-only in the first row and 0.228 versus 0.230 in the last. "p=0 top-k" denotes leading four, $p = 0$.

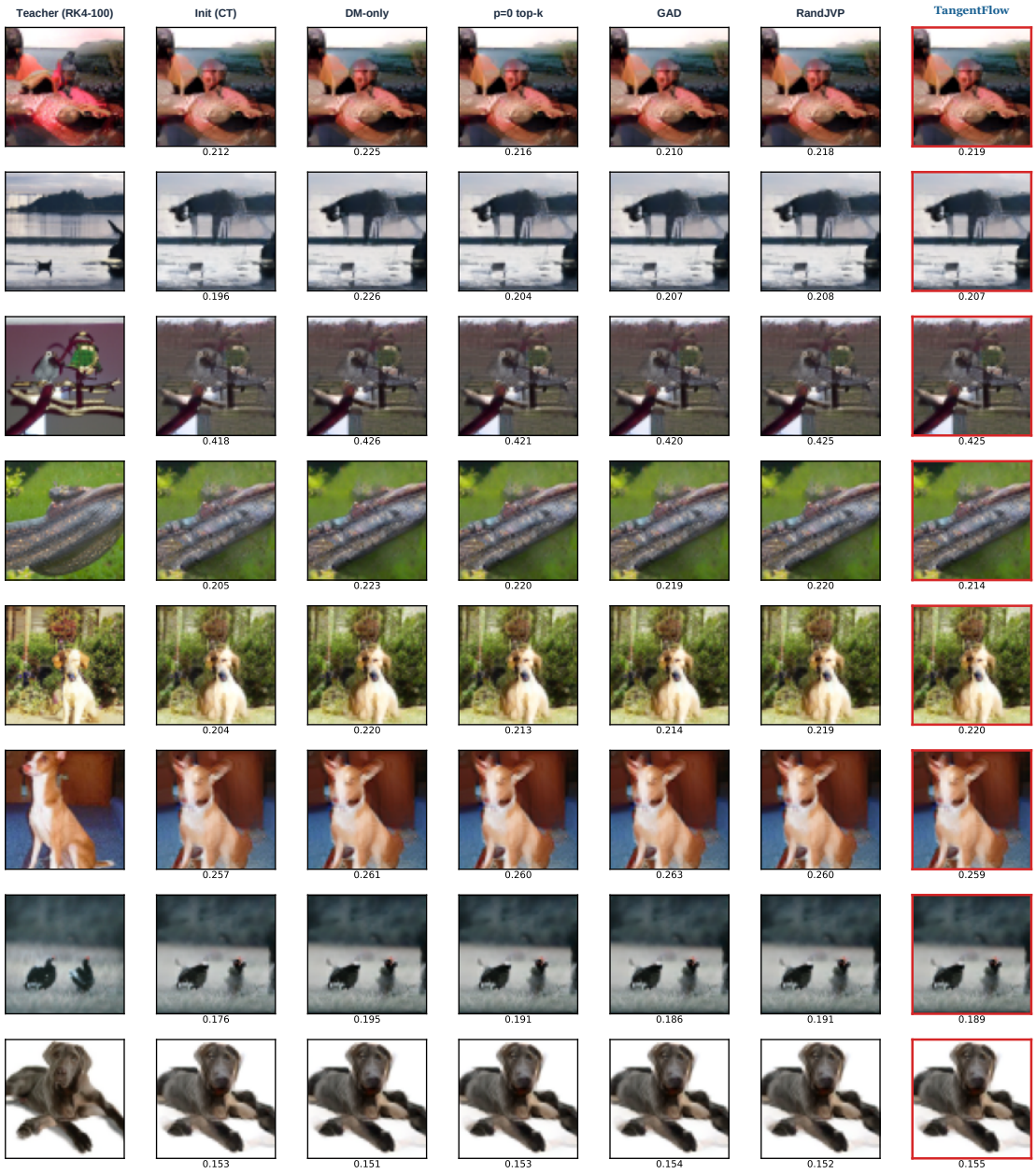


Figure 11: **ImageNet-64 paired comparisons, uncurated** (the first eight held-out inputs, no selection). TangentFlow is closer to the teacher than DM-only in six of the eight rows (for example 0.207 versus 0.226 in the second row and 0.189 versus 0.195 in the seventh), ties in one, and is farther in one. This matches the full-pool result: over all 256 held-out inputs, TangentFlow has lower mean LPIPS than DM-only (0.241 versus 0.244). Which student is closest varies from input to input.

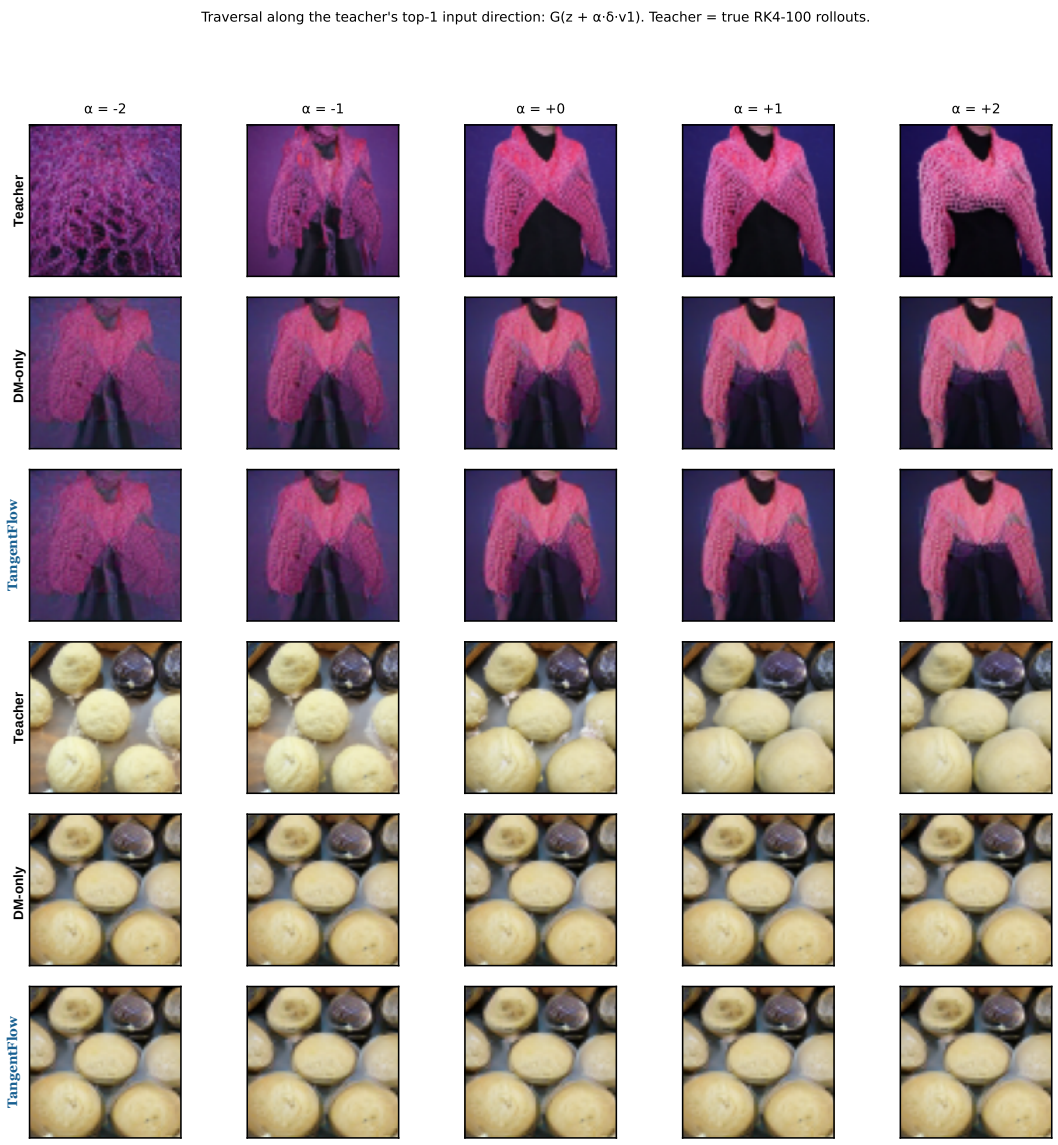


Figure 12: **Traversals along the teacher's leading input direction on ImageNet-64**, $G(z + \alpha\delta v_1)$ for $\alpha \in \{-2, \ldots, 2\}$, where v_1 is the teacher's leading right singular vector at z and δ is a fixed input step size shared by all rows, compared with true teacher rollouts. This is the direction that TangentFlow supervises most strongly. Both students vary smoothly along it while preserving the object in both examples; the teacher's own response is larger, consistent with the small, preservation-level geometric changes that a 100-update adaptation allows (Section 5.3).

J Protocols of Published CIFAR-10 Results

Table 18 gives the published protocol behind each entry of Table 3. Differences in teachers, architectures, budgets, and evaluators prevent a matched causal comparison, and the selection is neither exhaustive nor a state-of-the-art claim. Published ImageNet results are omitted because our ImageNet experiment studies a short adaptation. Teacher-relative geometry claims concern locally evaluated methods only.

Table 18: **Published one-step CIFAR-10 context:** Class-conditional one-step results; full method names are in Appendix A.1. Values follow each source's published protocol, not matched budgets or evaluators, and the selection is neither exhaustive nor a state-of-the-art claim. The Easy Consistency Tuning model (ECM) uses consistency fine-tuning; the other blocks use distillation or diffusion-guided generator training.

Method / configuration	Published protocol	FID↓
Guided distillation (Meng et al., 2023)	Multi-w; deterministic, $w = 0.3$	7.34
GET-Base (Geng et al., 2023)	Generative Equilibrium Transformer; offline EDM pairs	6.25
Guided distillation (Meng et al., 2023)	Separately trained single-$w = 0.3$	5.98
Diff-Instruct (Luo et al., 2023)	EDM teacher to UNet	4.19
KDM (EDM) (Berman et al., 2025)	Koopman Distillation Model; GET protocol	3.56
KDM-F (EDM) (Berman et al., 2025)	Factorized Koopman model; GET protocol	3.24
DMD (Yin et al., 2024b)	Distribution-matching distillation	2.66
FGM (Huang et al., 2024)	Flow Generator Matching; flow teacher	2.58
StyleGAN2-ADA+Tune+DI (Luo et al., 2023)	Style-based GAN, adaptive discriminator augmentation; Diff-Instruct (DI)	2.27
ECM (Geng et al., 2025)	Consistency fine-tuning; 100k iterations	3.81
Our DM+GAN base	EDM-initialized UNet; 50k samples	2.690 ± 0.028
TangentFlow (ours)	Same base with spectral tangents	2.117 ± 0.026