DISCRIMINANT FLOW STEERING: TRAINING-FREE CAPABILITY TRANSFER IN FLOW-MATCHING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open flow-matching image models fail specific, checkable and sometimes trivial capabilities that stronger, often closed, models handle like a three-column menu, an orthographic drawing, or a diagram with exactly three boxes. We ask whether we can close such a gap with no training, no teacher weights and no teacher call at inference, and from about twenty images from each model. To answer this, we propose Discriminant Flow Steering (DFS), which encodes the teacher images with the student's VAE and uses the difference between the mean teacher latent and the mean student latent as the steering direction when generating new images. The unit vector along this difference is added to the latent during the first sampling steps; nothing else changes. We find that this vector is what guidance with a second model adds at each step, except that we compute it only once and offline. This holds exactly at the first step for any two models, and approximately at high noise if we model the latents as Gaussian. Prior steering methods fit their direction at the same noise level where they apply it. We derive the exact finite-sample law for such a fit and find that, in the first steps where steering matters most, 99% of the fitted direction is noise. This is why we estimate the direction from clean latents. We also measure how far a perturbation injected at each step travels to the final image. At the first step the mean difference is amplified 8.8 times more than a random direction, and by mid-trajectory the advantage is gone. To measure capability transfer rather than resemblance to the teacher, we build DFS-CapBench: 200 prompts in 20 capability families, each with a frozen checklist, admitted only if the teacher passes and the student fails. An open vision-language model judges every checklist item and never sees a teacher image. On the 126 strict prompts DFS passes 42.1% of images at one global strength, against 10.1% for the student on fresh seeds, and above prompt rewriting, guidance sweeps and negative-prompt guidance with the same access. Attention-space negative guidance does better than DFS when the failure can be named in a negative prompt, and worse when the missing capability is structural, so the two methods complement each other. When we combine them, they pass 63.9% of images, on par with a per-prompt LoRA trained on the same images. On structural capabilities DFS alone passes more images than that LoRA and keeps the student's seed diversity. We also test the same method on a second student, a second teacher, and a small-scale video experiment.

1 INTRODUCTION

Text-to-image models are unevenly capable. An open flow-matching model such as FLUX.1-dev (Labs, 2024) renders convincing photographs and typography. Yet it reliably fails prompts that contradict a strong prior: it draws one column when asked for a three-column menu, one tree when asked for exactly three, and a perspective view when asked for an orthographic projection. A stronger model, often a closed one behind an API, handles the same prompts. Can the weaker model be made to handle them too, and what access to the stronger model does that need?

We study the most restrictive practical setting. The teacher is a black box: no weights, no gradients, no score function and no calls at inference. The only channel is about twenty of its output images

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.

per prompt. The student is an open flow-matching model whose sampler we can modify. Fine-tuning (Hu et al., 2021; Ruiz et al., 2023) fits this access model but needs a training run per capability. Distillation (Salimans & Ho, 2022; Song et al., 2023; Poole et al., 2022) needs the teacher's weights or score. Weak-to-strong guidance (Bai et al., 2026; Karras et al., 2024) needs both models at every step. We want a method that is offline, sample-only and free at inference. The images are all a closed model can hand over. From them, we want the open model to produce the capability itself, on its own seeds, with its own diversity and controls, and without calling the teacher again.

Discriminant Flow Steering (DFS) is such a method (Figure 1). For a prompt, it encodes the N teacher images through the student's VAE and generates N student latents for the same prompt. Let $\boldsymbol{\mu}_G$ and $\boldsymbol{\mu}_F$ be the means of the two sets. The direction is the unit vector $\mathbf{v}$ along $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_G - \boldsymbol{\mu}_F$, which at the noise levels where the push acts is Fisher's discriminant of the two sets, hence the name (Section 3.1). During sampling, DFS adds $\alpha\, h(\sigma)\, \mathbf{v}$ to the latent after every Euler step, where the schedule $h(\sigma)$ concentrates the push in the first steps. Nothing else changes. We add no forward passes, edit no weights, and never query the teacher again.

The broader point is simple. We can improve an open model with no training, using only a stronger model's images. And turning these failures into successes takes only a small push, 7.6% of the latent norm in total and mostly in the first five steps.

Three results say why this simple recipe is the right one. First, why this direction. Guidance with a second model, as in autoguidance or weak-to-strong guidance, adds the difference between two velocity fields at every step. Call the field the student would have after fine-tuning on the teacher's images the *target field*. At the first sampling step, a step with the target field instead of the student's field moves the latent exactly along $\Delta\boldsymbol{\mu}$, whatever the two distributions are. Under a Gaussian model the same holds in direction while the noise is high. So $\Delta\boldsymbol{\mu}$ is what two-model guidance adds, except that we compute it once from samples (Section 3.1). It is also easy to estimate. We need no covariance, the $N = 20$ estimate is within cosine 0.95 of what unlimited images would give, and no alternative estimator steers significantly better.

Second, why clean latents. Prior steering methods fit their direction at the noise level where they apply it. With few samples, that adds the mean of the particular noise draws to the estimate. We give the exact law for this contamination, confirm it on held-out images with no free parameter, and show that at the levels where steering acts such a vector is almost entirely noise.

Third, why the first steps. A perturbation along $\Delta\boldsymbol{\mu}$ injected at the first step reaches the output at 55 times its size, against 5.1 times for a random direction of the same size, and the advantage decays to parity by mid-trajectory (Section 3.3).

To systematically measure where the student falls short of the teacher, and how much of that gap DFS and other methods close, we build **DFS-CapBench**. It has 200 prompts across 20 capability families, each with a frozen checklist of binary items. A prompt is admitted only if the teacher passes every item on all four seeds and the student fails on all four (126 strict prompts) or on all but one (74 relaxed). An open vision-language judge scores each item and never sees a teacher image. We chose this design for two reasons. A judge that sees teacher images would score resemblance to the teacher rather than the capability. And because prompts are admitted using only the teacher and the student, the prompt set cannot favour any method.

On the strict tier DFS passes 42.1% of images at one global strength, a gain of 33.1 points over the student's floor measured on fresh seeds. Attention-space negative guidance (NAG) and DFS solve nearly disjoint sets of prompts. NAG wins where the failure is an extra object that a prompt can name, such as the toppings on a hamburger ordered with only a patty and a bun. DFS wins where the capability is structural, such as layout, diagrams, counting and projection. When we combine them, they pass 63.9%. A LoRA trained per prompt on the same images passes more images than DFS alone, but it needs a training run per prompt and passes fewer images on the structural families.

Contributions. (1) We propose DFS, which transfers a capability from a black-box teacher using only about twenty of its images, with no training and no extra cost at inference. We show that its direction is what guidance with a second model adds, computed once offline (Section 3). (2) We derive the exact finite-sample law for directions fitted on noised latents, which explains why we estimate from clean latents. We also give the first measurement, for a flow-matching image model, of how far a perturbation at each step travels to the final image, which explains why we steer only

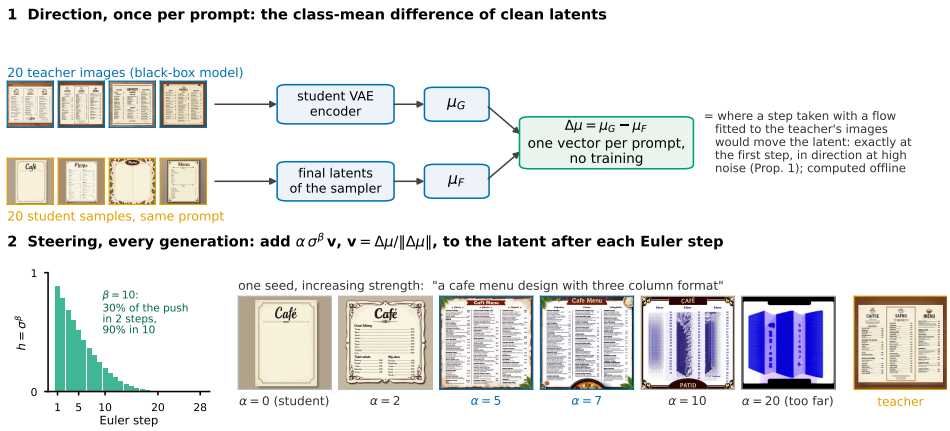


Figure 1: **DFS on a strict benchmark prompt.** The prompt is "A cafe menu design with three column format." Top: we encode the teacher images with the student's VAE, generate the same number of student samples for the same prompt, and take the difference of the two class means as the direction. This gives one vector per prompt with no training. Bottom: during sampling, we add the unit vector to the latent after each Euler step with a front-loaded schedule. One seed at increasing strength: the unsteered student ignores the three-column instruction, $\alpha \in [5, 7]$ produces the teacher's layout, and larger strengths leave the region where the sampler is reliable.

in the first steps (Sections 3.1 and 3.3). (3) We build DFS-CapBench, a benchmark in which every prompt has a verified gap between teacher and student and the judge never sees a teacher image. We release it as a testbed for steering and capability-transfer methods (Section 4).

2 Related Work

Steering directions in diffusion and flow models. Linear structure in generative representations underlies steering in language models (Zou et al., 2023; Turner et al., 2024; Li et al., 2023; Park et al., 2023a) and in diffusion models (Kwon et al., 2022; Park et al., 2023b). Linear-AcT (Rodriguez et al., 2025), SHIFT (Konovalova et al., 2026) and CountSteer (Boo et al., 2025) fit affine maps or mean-difference vectors over one model's activations between two prompt or outcome distributions. NA-RFM learns class directions from the noised activations of labelled images and notes that they lose alignment at high noise (Wang et al., 2026). Concept Sliders train low-rank adaptors between two prompts (Gandikota et al., 2024). All of these steer one model toward a concept it already represents. DFS estimates a direction *between two models* from the teacher's images alone, and asks whether it moves a capability the student lacks. Proposition 2 adds the finite-sample law for directions estimated from few noised samples.

Using a second model, or its images. Classifier-free guidance (Ho & Salimans, 2022), autoguidance (Karras et al., 2024), weak-to-strong diffusion (Bai et al., 2026) and sparse guidance (Krause et al., 2026) all add the difference between two velocity or score predictions at every step. This needs both models at inference and rules out a closed teacher. Proposition 1 shows that at the first step, and under a Gaussian model at high noise, this difference is a fixed direction that can be computed offline from samples. DFS is therefore the sample-based, zero-overhead form of two-model guidance. Task arithmetic (Ilharco et al., 2022) takes the difference between two models in weight space, DFS in sample space. Negative guidance, in the prompt or in attention space (NAG) (Chen et al., 2026), pushes away from a concept named in text. DreamBooth (Ruiz et al., 2023), LoRA (Hu et al., 2021) and textual inversion (Gal et al., 2022) learn from a few images by fitting the weights, an adapter or a token embedding, and each needs a training run per concept. IP-Adapter (Ye et al., 2023) and ControlNet (Zhang et al., 2023) condition each generation on a reference input through a trained module. Distillation (Hinton et al., 2015; Salimans & Ho, 2022; Song et al., 2023; Poole et al., 2022) needs the teacher's weights or score. We compare against negative guidance and LoRA.

Noise-level structure and evaluation. Critical windows (Li & Chen, 2024) and limited guidance intervals (Kynkäänniemi et al., 2024) show that the noise level decides what an intervention can do; our transport measurement asks how far a perturbation injected at a given level travels. VQA-based metrics (Cho et al., 2024; Lin et al., 2024) and compositional benchmarks (Ghosh et al., 2023; Huang et al., 2023) score prompt adherence with pretrained vision models. DFS-CapBench adds a verified, student-specific gap per prompt and a binary all-items pass.

3 METHOD

Setting. A *teacher* G produces images with a capability that the *student* F lacks for a prompt c. We cannot access the teacher's internals, and we only have N of its images. The student is a flow-matching model with VAE encoder $\mathcal{E}_F$, and we can modify its sampler. Both sets of images are represented in the student's latent space $\mathbb{R}^d$, $d = C \times H \times W$. The teacher latents are $\mathbf{z}_0^{G,i} = \mathcal{E}_F(\mathbf{x}_G^{(i)})$, and the student latents $\mathbf{z}_0^{F,i}$ are the final latents of N student generations for c. The student samples with rectified flow (Lipman et al., 2022; Liu et al., 2022): $\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_0 + \sigma\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ drawn independently of $\mathbf{z}_0$, integrated by Euler steps $\mathbf{z}_{\sigma_{i+1}} = \mathbf{z}_{\sigma_i} + (\sigma_{i+1} - \sigma_i)\mathbf{f}_F(\mathbf{z}_{\sigma_i}, \sigma_i)$ from $\sigma_0 = 1$ down to $\sigma_T = 0$, where $\mathbf{f}_F(\mathbf{z}, \sigma) = \mathbb{E}[\boldsymbol{\epsilon} - \mathbf{z}_0 \mid \mathbf{z}_\sigma = \mathbf{z}]$ is the student's velocity field. A closed teacher need share neither latent space nor architecture with the student, so we never refer to its network. Instead, $\mathbf{f}_G$ denotes the *target field*: the velocity of a flow whose clean endpoints are distributed as the teacher's images encoded by $\mathcal{E}_F$. This is the field the student would have after fine-tuning to convergence on the teacher's outputs. All numerical checks in this section and in Section 3.3 use our main pair, with FLUX.1-dev as the student and FLUX.2-klein-9B as the teacher.

3.1 WHY THIS DIRECTION: WHAT A STEP WITH THE TARGET FIELD WOULD DO

Guidance methods that use a second model at inference, such as classifier-free guidance (Ho & Salimans, 2022), autoguidance (Karras et al., 2024) and weak-to-strong diffusion (Bai et al., 2026), add a multiple of the *difference between two velocities* at every step. We cannot evaluate $\mathbf{f}_G$, but its difference from $\mathbf{f}_F$ has a closed form. Since $\mathbf{z}_\sigma - \mathbf{z}_0 = \sigma(\boldsymbol{\epsilon} - \mathbf{z}_0)$, any such field can be written as $\mathbf{f}(\mathbf{z}, \sigma) = (\mathbf{z} - \mathbf{m}(\mathbf{z}))/\sigma$, where $\mathbf{m}(\mathbf{z}) = \mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_\sigma = \mathbf{z}]$ is the model's current estimate of the clean latent. Two fields therefore differ only through their estimates $\mathbf{m}$. One sign convention matters here. The sampler moves toward decreasing σ, so a velocity difference and the displacement it causes have opposite signs. A step taken with $\mathbf{f}_G$ instead of $\mathbf{f}_F$ moves the latent by an extra $+\frac{|\Delta\sigma|}{\sigma}(\mathbf{m}_G - \mathbf{m}_F)$, toward the target field's estimate.

Proposition 1 (A step with the target field displaces the latent along the discriminant)**.** *Let p_F and p_G be the distributions of the student's clean latents and of the teacher's images encoded by $\mathcal{E}_F$, with means $\boldsymbol{\mu}_F, \boldsymbol{\mu}_G$ and $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_G - \boldsymbol{\mu}_F$, and let $\mathbf{f}_k(\mathbf{z}, \sigma) = \mathbb{E}[\boldsymbol{\epsilon} - \mathbf{z}_0 \mid \mathbf{z}_\sigma = \mathbf{z}]$ be the rectified-flow velocity with endpoint distribution p_k.*
(i) First step, any distributions. $\mathbf{f}_G(\mathbf{z}, 1) - \mathbf{f}_F(\mathbf{z}, 1) = -\Delta\boldsymbol{\mu}$ *for every* $\mathbf{z}$.
(ii) Gaussian classes with a shared covariance. *If $p_k = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, then for every $\mathbf{z}$ and $\sigma \in (0, 1]$,*

$$\mathbf{f}_G(\mathbf{z}, \sigma) - \mathbf{f}_F(\mathbf{z}, \sigma) = -\sigma\, \mathbf{S}_W(\sigma)^{-1}\Delta\boldsymbol{\mu}, \qquad \mathbf{S}_W(\sigma) = (1-\sigma)^2\boldsymbol{\Sigma} + \sigma^2\mathbf{I}. \tag{1}$$

(iii) Consequence. *An Euler step from σ_i to $\sigma_{i+1} < \sigma_i$ taken with $\mathbf{f}_G$ in place of $\mathbf{f}_F$ displaces the latent by an additional $|\Delta\sigma_i|\,\sigma_i\,\mathbf{S}_W(\sigma_i)^{-1}\Delta\boldsymbol{\mu}$. For $0 < \sigma < 1$ this is parallel to Fisher's discriminant of the two noised classes, and $\cos\big(\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\mu}\big) \geq 2\sqrt{1+\rho}/(2+\rho)$ with $\rho = (1-\sigma)^2\lambda_{\max}(\boldsymbol{\Sigma})/\sigma^2$, so its direction tends to $\Delta\boldsymbol{\mu}$ as $\sigma \to 1$.*

The proof is in Appendix A.1. Part (i) needs no model. At $\sigma = 1$ the latent is pure noise, independent of the image it is paired with, so each field's estimate of the clean latent is its class mean. So at the first step, two-model guidance adds the same vector for every starting noise. This is why we can compute the vector once, offline, and use it in place of a second model. Part (ii) says how this degrades at lower noise. $\mathbf{S}_W^{-1}$ suppresses the components of $\Delta\boldsymbol{\mu}$ along which images within one class vary by more than the noise does. This is Fisher's logic (Fisher, 1936) and the origin of the method's name. Part (iii) makes "high noise" a property of the spectrum, not of σ alone: the direction stays close to $\Delta\boldsymbol{\mu}$ as long as the noise variance σ^2 dominates the largest within-class variance $(1-\sigma)^2\lambda_{\max}(\boldsymbol{\Sigma})$. With unequal class covariances the best single vector is along $\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$, where $\mathbf{S}_{W,G}$ is the noised covariance of the teacher class (Proposition 3, Appendix A). On FLUX

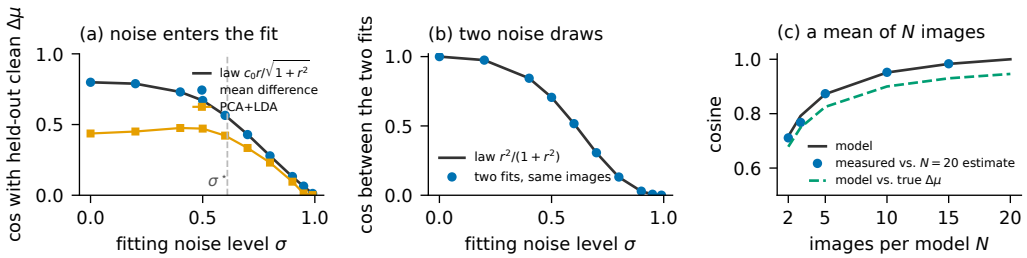


Figure 2: **The estimator** (200 prompts, medians). (a) A mean difference fit on latents noised to level σ follows Equation (3) with no free parameter; PCA+LDA on the same noised sample is worse throughout. (b) Two fits on the same images under different noise draws. (c) The clean mean difference from N images against the $N = 20$ estimate that contains them, the one-parameter sampling model, and what it implies against the population $\Delta\boldsymbol{\mu}$.

latents the part of the velocity difference that no fixed vector captures is 1% of its energy at the first step and 24% at $\sigma = 0.5$ (Table 3). That remainder is the price of not running the teacher, and it is paid mostly at low noise. This is the first reason to steer at high noise. Section 3.3 gives the second.

The estimator. Proposition 1 asks for $\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu} \propto (\boldsymbol{\Sigma} + \tau\mathbf{I})^{-1}\Delta\boldsymbol{\mu}$ with $\tau = \sigma^2/(1-\sigma)^2$. This is a Fisher discriminant whose ridge τ is set by the noise level. At the first step the ridge is 7,271, far above the largest within-class eigenvalue of either class, and it is still 90 at the seventh step, $\sigma = 0.905$ (Table 3). So at the steps that matter, the prescribed direction is the mean difference, and we do not need to estimate a covariance. We could not estimate one anyway from $N = 20$ samples in $d = 262{,}144$ dimensions (Bickel & Levina, 2004). We use

$$\mathbf{v} = \widehat{\Delta\boldsymbol{\mu}}/\|\widehat{\Delta\boldsymbol{\mu}}\|, \qquad \widehat{\Delta\boldsymbol{\mu}} = \tfrac{1}{N}\textstyle\sum_i \mathbf{z}_0^{G,i} - \tfrac{1}{N}\sum_i \mathbf{z}_0^{F,i}, \tag{2}$$

computed from *clean* latents. $N = 20$ is enough. Estimates from two disjoint halves of ten images agree at cosine 0.81, and a one-parameter sampling model fitted to Figure 2c puts the $N = 20$ estimate at cosine 0.95 with the direction unlimited images would give (Appendix A.4). No alternative estimator steers significantly better (Section 5.2).

Why clean latents. The vector is applied to noised latents, so it is natural to estimate it from latents noised to the same level. Published steering methods do this with noised activations (Wang et al., 2026; Boo et al., 2025). With finite samples, however, the noise draws enter the estimated class means.

Proposition 2 (Noise contamination of the sample mean difference)**.** *Fix the $2N$ clean latents and noise each with its own draw, $\mathbf{z}_\sigma^{k,i} = (1-\sigma)\mathbf{z}_0^{k,i} + \sigma\boldsymbol{\epsilon}_i^k$. Then $\widehat{\Delta\boldsymbol{\mu}}(\sigma) = (1-\sigma)\,\widehat{\Delta\boldsymbol{\mu}}_0 + \sigma\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \bar{\boldsymbol{\epsilon}}_G - \bar{\boldsymbol{\epsilon}}_F \sim \mathcal{N}(\mathbf{0}, \frac{2}{N}\mathbf{I})$ independent of the images and $\|\boldsymbol{\eta}\| = \sqrt{2d/N}\,(1 + O(d^{-1/2}))$. With $r(\sigma) = \frac{1-\sigma}{\sigma}\,\kappa$ and $\kappa = \|\widehat{\Delta\boldsymbol{\mu}}_0\|\sqrt{N/2d}$,*

$$\cos\big(\widehat{\Delta\boldsymbol{\mu}}(\sigma), \widehat{\Delta\boldsymbol{\mu}}_0\big) = \frac{r}{\sqrt{1+r^2}} + O(d^{-1/2}). \tag{3}$$

The two terms have equal norm at $\sigma^\star = \kappa/(1+\kappa)$, where the cosine is $1/\sqrt{2}$; as $\sigma \to 1$ the estimate tends to the noise direction $\boldsymbol{\eta}$.

The proof is in Appendix A.3. In words, $\boldsymbol{\eta}$ is pure noise that a mean over N draws does not remove and is nearly orthogonal to $\widehat{\Delta\boldsymbol{\mu}}_0$ in this dimension. The ratio r compares the lengths of the signal and noise terms. For FLUX.1-dev latents the two terms have equal norm at $\sigma^\star = 0.67$, and the damage is worst where steering acts. At $\sigma = 0.95$ the cosine is 0.11, so 99% of the fitted vector's energy is the noise draw. Figure 2a,b confirms Equation (3) on held-out images with no free parameter. Whitening does not help. PCA+LDA divides by a covariance built from the same noised latents, so $\boldsymbol{\eta}$ lies in its span and cannot be divided out (Figure 2a). Sharing or averaging the draws removes $\boldsymbol{\eta}$ but only returns the clean estimator scaled by $(1-\sigma)$ (Appendix A.3). The law matches the alignment loss at high noise that Wang et al. (2026) report.

3.2 STEERING

After each Euler step, at level σ_{i+1}, we add the direction to the latent,

$$\mathbf{z} \leftarrow \mathbf{z} + \alpha\, h(\sigma_{i+1})\, \mathbf{v}, \qquad h(\sigma) = \sigma^{\beta}, \tag{4}$$

and skip steps where $h < 10^{-4}$. Nothing else in the sampler changes (Algorithm 1, appendix). Since $\|\mathbf{v}\| = 1$, $\alpha h(\sigma)$ is the norm of the push at each step. At $\alpha = 7$ the pushes total 7.6% of the latent's norm, and with $\beta = 10$ on the 28-step FLUX schedule 62% of that falls in the first five steps. This is the range in which $\Delta\boldsymbol{\mu}$ is the right vector (Section 3.1) and in which a perturbation is transported most strongly (Section 3.3). The exponent is not critical. Every front-loaded schedule performs like $\beta = 10$, and a push applied only below $\sigma = 0.85$ has no effect (Table 18).

3.3 WHY THE FIRST STEPS: HOW A PERTURBATION REACHES THE OUTPUT

The schedule question is how a perturbation δ injected at level σ propagates through the remaining steps. We measure this directly. For 6 strict benchmark prompts and 3 seeds we run the unmodified sampler, then re-run it from the same seed with δ added after the step that reaches $\sigma \in \{0.988, 0.95, 0.89, 0.76, 0.51\}$, with $\|\delta\| \in \{1, 10\}$. We record the gain $\|\tilde{\mathbf{z}}(0) - \mathbf{z}(0)\|/\|\delta\|$, where $\mathbf{z}(0)$ and $\tilde{\mathbf{z}}(0)$ are the final latents without and with the perturbation, for three directions: the clean-latent $\Delta\boldsymbol{\mu}$, a phase-scrambled $\Delta\boldsymbol{\mu}$ with the same spatial power spectrum, and isotropic Gaussian directions (Figure 4 and Table 5, appendix).

Transport is front-loaded, direction-selective and sublinear. A random perturbation of norm 10 injected at $\sigma = 0.988$ moves the final latent by 5.1 times its norm, and $\Delta\boldsymbol{\mu}$ moves it by 55 times. By $\sigma = 0.51$ both gains are a few units, and the advantage of $\Delta\boldsymbol{\mu}$ over a random direction of equal norm, the median per-seed ratio of the two gains, has fallen from 8.8 times to 1.5 times. A phase-scrambled $\Delta\boldsymbol{\mu}$ with the same power spectrum is transported 3.5 times more than random, and a Gaussian model of the flow, which sees only second moments, predicts a first-step gain of 21 against the measured 55 (Appendix A.5). So part of $\Delta\boldsymbol{\mu}$'s advantage is linear: it lies along the directions in which the student's own images vary most, and the flow amplifies those. The remaining factor of about 2.6 has no Gaussian explanation. Fitting on clean latents and steering at high noise are therefore not in tension. The flow carries the direction farthest at exactly the steps where Proposition 1 says it is the right one. Finally, raising $\|\delta\|$ from 1 to 10 raises the output displacement by 3.9 times, not 10 times. The map is compressive at the strengths steering uses, so no first-order budget holds and α must be set by a sweep (Table 9).

4 DFS-CAPBENCH: MEASURING A CAPABILITY GAP AND ITS CLOSURE

Prompts and frozen checklists. Prompts come from 20 capability families (Appendix B). A language model wrote most of them from family templates, and we wrote 20 by hand to contradict a strong prior. For each prompt, a language model wrote a checklist of binary questions (median 7 items) from the prompt text alone, before we generated any image, and we never edited it. Exclusivity items make sure an image cannot pass by over-generating.

Judge and pass rule. A locally hosted open vision-language model (Qwen3-VL-32B-Instruct (Bai et al., 2025)) answers each item; $P(\text{yes})$ is read from the yes/no logits with an instruction to answer "No" when a property is ambiguous. An image **passes** iff every item has $P(\text{yes}) \geq 0.5$, with no partial credit. The judge never sees a teacher image, and the same frozen checklist scores every arm.

Verified-gap gate. For each candidate prompt, we score four seeds of the student (FLUX.1-dev) and the teacher (FLUX.2-klein-9B (Labs, 2026)). We admit the prompt only if the teacher passes on all four seeds. It is *strict* if the student passes on none and *relaxed* if on at most one. Of 1909 scored candidates, 200 remain after the gate, near-duplicate removal and an author veto of borderline prompts (126 strict, 74 relaxed). Gate and evaluation share their four seeds, so the student's zero on the strict tier holds by construction and says nothing about how often it fails these prompts. On four fresh seeds it passes 10.1% of the same prompts' images, and a push along an unrelated direction reaches about the same. Section 5.1 therefore reports fresh seeds and a control direction next to the gate seeds, and reads every gain against that floor.

Table 1: **Strict tier of DFS-CapBench** (126 prompts $\times$ 4 seeds). The tier is defined by the student failing on the evaluation seeds, so its first row is zero by construction; "fresh seeds" rows use four seeds never used for selection, and the control direction is another family's $\Delta\boldsymbol{\mu}$. Pass: images passing every checklist item; solved: prompts with $\geq 3/4$ passing seeds; any: ≥ 1; pick: the cell passes if the image at any of the arm's strengths passes, DFS $\alpha \in \{5, 7, 10\}$, NAG scale $\{3, 5, 8\}$, LoRA scale $\{0.5, 1, 1.5\}$, negative CFG $\{2, 4\}$, the stack's three settings (Section 4). Tch: teacher images used; fwd: extra forward passes per step. The negative prompts for CFG and NAG are written by the MLLM (Section 5). All rows and strengths: Table 7.

			at the strength in the row label			best strength
Method	tch	fwd	pass %	solved %	any %	pick %
Student (unsteered)	–	0	0.4	0.0	1.6	–
on four fresh seeds	–	0	10.1	1.6	29.4	–
CFG sweep, best of $\{1.5, 7, 10\}$	–	0	12.3	7.9	26.2	25.6
Negative-prompt CFG, best of $\{2, 4\}$	3	1$\times$	22.8	12.7	43.7	31.3
Prompt rewrite by MLLM	3	0	30.4	19.0	55.6	–
Control: another prompt's $\Delta\mu$, α=7	20	0	10.7	5.6	23.8	20.0
NAG, scale 5	20	1$\times$	35.9	26.2	60.3	51.0
LoRA per prompt, same teacher images	20	0	58.3	54.0	80.2	75.0
DFS, $\Delta\mu$, α=7	20	0	**42.1**	**33.3**	**65.9**	**59.5**
on four fresh seeds	20	0	**43.3**	**33.3**	**73.0**	**56.0**
DFS α=7 + NAG scale 5	20	1$\times$	**63.9**	**57.9**	**86.5**	**75.2**

Metrics. Each arm is scored as follows. The *image pass rate* is the fraction of prompt–seed cells whose image passes every checklist item. The *prompt solved rate* is the fraction of prompts on which at least three of the four seeds pass, and *any* the fraction with at least one passing seed. For a method with a strength setting, the *pick rate* counts a cell as passed if the image at any of the method's declared strengths passes, which is what a user who generates each strength and keeps the best image obtains. We nevertheless lead with one fixed strength per method, because choosing α per prompt on two seeds does not transfer to two other seeds (45.8% on the tuning seeds, 41.7% on fresh ones, Table 9). A difference is called significant when a paired bootstrap over prompts rejects zero at the 5% level. The gate is student-specific: re-running it for another student yields that student's own benchmark (Section 5.2).

5 Experiments

Setup. Student FLUX.1-dev, teacher FLUX.2-klein-9B treated as a black box. Per prompt, $N = 20$ teacher images and as many student generations with fixed seeds give $\Delta\boldsymbol{\mu}$. Steering uses $\beta = 10$, 28 steps, guidance 3.5 and 1024^2. Evaluation uses the four gate seeds, and the student's floor is measured on four further seeds (Section 4). DFS is reported at $\alpha = 7$, the strength that was best on the evaluation seeds when it was chosen. It is also the best of $\{5, 7, 10\}$ on four fresh seeds that played no part in that choice (Table 9). NAG and LoRA are reported at their pre-declared scale 5 and scale 1, the two guidance sweeps at their best value, labelled so, and every sweep is in Appendix D. To test a student from another model family we repeat the benchmark with Stable Diffusion 3-medium as the student (Section 5.2, Appendix D.1), and to test a teacher from another family we replace FLUX.2-klein by Qwen-Image-2512 (Wu et al., 2025) with the same FLUX student (Appendix D.2).

Alternatives with the same access. All use the same teacher images or fewer. *Prompt rewrite*: the judge model sees the prompt and three teacher images and rewrites the prompt so the weaker model produces images like them; it also writes a negative prompt naming the failure. *Negative-prompt CFG*: true classifier-free guidance away from that negative prompt. *NAG* (Chen et al., 2026): attention-space guidance away from the same negative prompt, at its default scale. *CFG sweep*: guidance $\{1.5, 7, 10\}$. *LoRA*: a rank-16 adapter trained per prompt on the same teacher images (DreamBooth-FLUX recipe). Two controls test the direction itself: another family's $\Delta\boldsymbol{\mu}$ at the same strength (Table 1) and its per-channel constant part alone (Table 7).

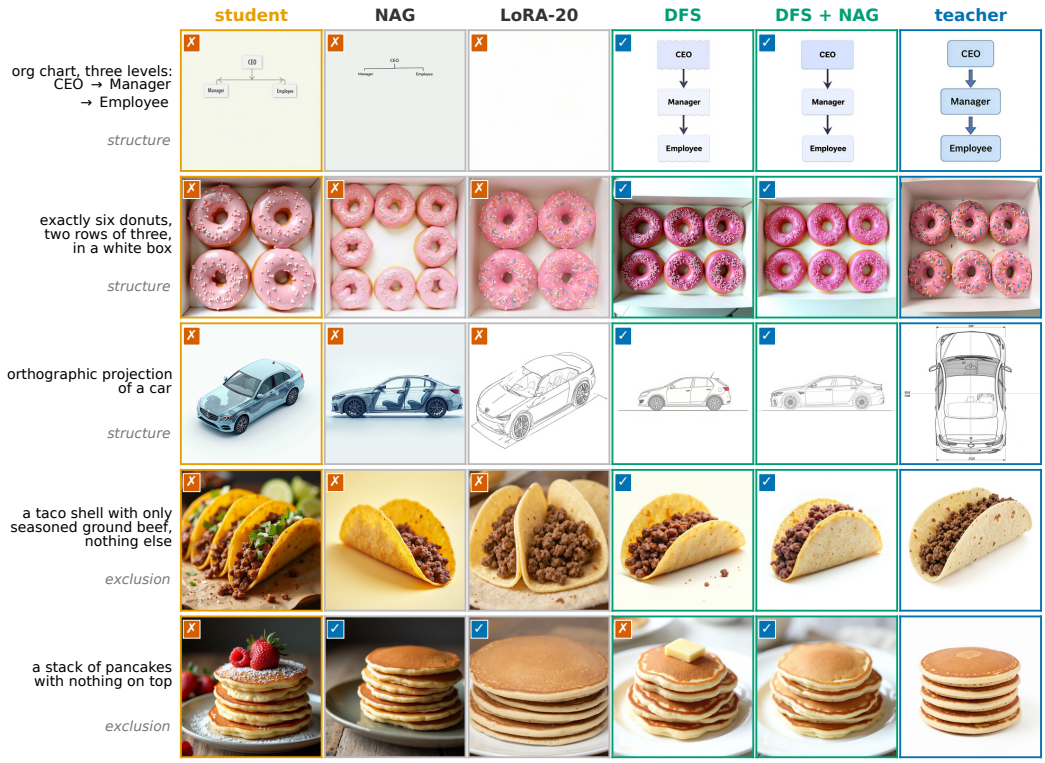


Figure 3: **The same seed under every method.** The tick or cross is the judge's verdict that enters Table 1. Rows 1–3, structure gaps from three families: only DFS and the DFS+NAG stack produce the three-level chart, the box of two rows of three and the orthographic car, and LoRA-20 returns a near-blank page for the chart. Row 4, an exclusion gap that DFS passes: the taco keeps only its filling. Row 5, an exclusion gap it fails: NAG and LoRA remove the topping, DFS alone does not, and the stack does. The rows were chosen to show the patterns of Table 2, not at random. More prompts are in Figures 7 to 10, and failures of DFS in Figure 11.

Table 2: **Strict-tier pass rate (%) by what the capability gap is** (families pre-assigned when the benchmark was built). Exclusion gaps have a nameable extra object; structure gaps cannot be phrased as an exclusion. NAG at its default scale 5 (scale 8: 76.7, 29.2, 43.9). Per family: Table 6.

gap type	n	rewrite	neg. CFG	NAG	LoRA	DFS	DFS +NAG
exclusion (negation, subtraction)	30	45.8	40.0	65.8	**79.2**	25.0	72.5
structure (layout, diagram, counting, projection, text, binding)	59	22.9	9.3	18.6	45.8	56.8	**63.6**
appearance (medium, viewpoint, spatial, other)	37	29.7	30.4	39.2	**61.5**	32.4	57.4

5.1 Main results

At one global strength DFS passes 42.1% of strict images and solves 33.3% of prompts (Table 1 and Figure 3; relaxed tier and all 200 prompts in Table 13). On fresh seeds the student passes 10.1%, and an equal push along another family's $\Delta\mu$, which carries no capability, passes 10.7% (Section 4). DFS is unaffected by the selection (43.3% on fresh seeds), so its gain is 33.1 points. No training-free alternative without a second forward pass comes close (Table 1), and the pick column leaves this ordering unchanged (Table 9).

What a latent direction can and cannot do. Table 2 splits the strict tier by the kind of gap. On the 30 exclusion prompts, whose failure is an extra object that a text prompt can name, NAG passes

65.8% and DFS 25.0%. On the 59 structure prompts the order reverses: DFS passes 56.8% and NAG 18.6%, and here DFS also passes more images than the LoRA trained per prompt on the same images (45.8%) and solves more prompts (47.5% vs. 39.0%). Across the tier the two solve nearly disjoint prompt sets: DFS solves 42 prompts and NAG 33, only 10 in common. A latent direction pulls the sample toward what the teacher does. A negative prompt pushes it away from what an assistant could name. Stacking them at their pre-declared strengths confirms that the mechanisms are complementary: DFS with NAG passes 63.9% of strict images, above either alone and above the per-prompt LoRA, with no training (Table 1 and Figure 3). On the remaining appearance families (medium, viewpoint, spatial relation) NAG and DFS tie and LoRA leads (61.5%), which is also where fine-tuning has the most to copy.

Against fine-tuning. LoRA passes 58.3% of strict images against 42.1% for DFS, at the cost of a training run per prompt (about 17 minutes on one A100) where DFS costs N encodes and a subtraction. In DINOv2 features (Oquab et al., 2023) on the same seeds, LoRA moves outputs closest to the teacher's images and loses part of the student's seed diversity, while DFS keeps it intact. No method produces near-duplicates of a teacher image (Table 12). A LoRA pooled over all prompts, the "just distil" alternative, passes 41.1%.

5.2 Estimator analysis and generality

We fit every direction estimator a reader might propose, from logistic regression and a linear SVM to the AGOP eigenvector of NA-RFM (Wang et al., 2026) and PCA+LDA on noised latents, on the same $N = 20$ latents per class of a 30-prompt strict subset, and steer each at its own best strength (Table 10). Every alternative fits a model or tunes a hyperparameter, while DFS only subtracts two means. Still, none is significantly better than $\Delta\boldsymbol{\mu}$ (45.0%). The best two, the AGOP eigenvector and logistic regression, lead by +5.0 and +3.3 points, and with the strength chosen without leakage no alternative is above $\Delta\boldsymbol{\mu}$. The noised fits fall with σ as Proposition 2 predicts. $\Delta\boldsymbol{\mu}$ also works with fewer images. Even $N = 2$ is not significantly different from $N = 20$ (Table 11).

Other students, teachers and video. On Stable Diffusion 3-medium (Esser et al., 2024), re-gated, DFS passes 24.3% of strict images at one global strength, against 42.0% for FLUX on the same prompts. SD3-medium amplifies $\Delta\boldsymbol{\mu}$ over a random direction only 4.1× at the first step, against 8.8× for FLUX, so it needs a larger push, and its best strength, $\alpha = 15$, is the largest we swept. The capability is still within reach. A user who keeps the best of five strengths per image passes 43.0%, about FLUX's fixed-strength rate (Appendix D.1). With Qwen-Image-2512, a teacher from another family, the FLUX student reaches 42.9% against 44.0% with klein on the 42 strict prompts that both teachers pass reliably, a paired tie (Appendix D.2). With SD3.5-medium as the teacher for SD3-medium, a real network plays the target field. Its displacement toward the target has median cosine 0.34 with $\Delta\boldsymbol{\mu}$ at the first step and 0.04 at $\sigma = 0.9$, the sign and decay of Proposition 1 but not its magnitude. Applied unchanged to video, $\Delta\boldsymbol{\mu}$ moves the motion of Wan2.1-T2V-1.3B toward its teacher Wan2.2-T2V-A14B on all 11 prompts where the two differ (Appendix E).

6 Limitations and conclusion

Limitations. A fixed vector moves the mean of the student's distribution and cannot change its spread, and one direction captures one axis of difference per prompt. Proposition 1 describes ideal, unguided fields. A real guided network's velocity difference aligns with $\Delta\boldsymbol{\mu}$ far less than it predicts, and the strength and schedule are empirical. On FLUX $\alpha \in [5, 10]$ has held for every prompt, and the second student, which amplifies $\Delta\boldsymbol{\mu}$ half as much, needs about twice the push. The benchmark is judged by a 32B open model (human agreement in Appendix B). A closed commercial teacher, the motivating case, is not tested, and the video evidence is a generality check on fifteen prompts.

Conclusion. We showed that a capability gap between two flow-matching models can often be closed along one direction in the student's latent space, computed from about twenty images per model and added only during the first sampling steps. On DFS-CapBench this direction passes 42.1% of strict images, up from 10.1% for the student, and 63.9% when combined with negative guidance. On structural capabilities it passes more images than a per-prompt LoRA and keeps the student's diversity. We release the benchmark as a testbed for better estimators and other students.

REPRODUCIBILITY STATEMENT

Appendix B lists the benchmark construction, the frozen checklist format, the judge prompt and the gate. Appendix C lists every hyperparameter, baseline recipe and prompt, and Appendix A gives full proofs. The core code is at `https://anonymous.4open.science/r/DFS-5366`. We will add the benchmark, the direction files and the remaining code there, together with every random seed we used for the teacher images, the student latents and the evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks that require disclosure. Language models generated the candidate benchmark prompts and per-prompt checklists (Gemini 3.1 Pro and Qwen3-VL-32B, Appendix B), which we audited and froze before any steering run. A vision-language model (Qwen3-VL-32B-Instruct) serves as the automatic judge and as the prompt-rewrite baseline (Section 5). AI coding assistants helped implement the method, the baselines and the analysis code. AI assistants also gave feedback on the experimental design, helped derive and check the mathematical claims, and helped interpret results. We did not use generative AI for translation or for qualitative or thematic data analysis. Additionally, we used AI tools to write figure code, search for literature and check references, draft parts of the paper and edit it for readability. We reviewed all AI-assisted work. We tested all AI-assisted code and checked every reference against its source. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly released generative models under their licenses and introduces no new data about people. The benchmark prompts and checklists were authored by the authors and a language model and audited by the authors. The judge-audit annotations were provided by two volunteer raters, with no crowdsourcing. The raters only judged whether model-generated images satisfied checklist items, and we recorded no information about them. We therefore assessed that this annotation did not require approval from an ethics review board, and both raters agreed to the use of their labels. The method does not modify model weights and inherits the content-safety behaviour of the underlying student model; misuse risks are those of image generation generally.

REFERENCES

Lichen Bai, Masashi Sugiyama, and Zeke Xie. Weak-to-strong diffusion with reflection. In *International Conference on Learning Representations*, volume 2026, pp. 58921–58957, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Peter J Bickel and Elizaveta Levina. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Hyemin Boo, Hyoryung Kim, Myungjin Lee, Seunghyeon Lee, Jiyoung Lee, Jang-Hwan Choi, and Hyunsoo Cho. Countsteer: Steering attention for object counting in diffusion models. *arXiv preprint arXiv:2511.11253*, 2025.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. *Advances in Neural Information Processing Systems*, 38:137412–137446, 2026.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International conference on learning representations*, volume 2024, pp. 15625–15645, 2024.

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

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Nina Konovalova, Andrey Kuznetsov, and Aibek Alanov. Shift: Steering hidden intermediates in flow transformers. *arXiv preprint arXiv:2604.09213*, 2026.

Felix Krause, Stefan Andreas Baumann, Johannes Schusterbauer, Olga Grebenkova, Ming Gui, Vincent Tao Hu, and Björn Ommer. Guiding token-sparse diffusion models. *arXiv preprint arXiv:2601.01608*, 2026.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

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

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

Black Forest Labs. Flux.2 [klein]: Towards interactive visual intelligence, January 2026. URL https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence. Accessed: 2026-03-05.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models. *arXiv preprint arXiv:2403.01633*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023a.

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

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Xavier Suau, et al. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, volume 2025, pp. 89812–89855, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. 2024.

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

Qingsong Wang, Mikhail Belkin, and Yusu Wang. General and efficient steering of unconditional diffusion. *arXiv e-prints*, pp. arXiv–2602, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Proofs and measurements for Section 3

A.1 Proof of Proposition 1

Proof. Step 1: velocity through the posterior mean. Since $\mathbf{z}_\sigma - \mathbf{z}_0 = \sigma(\boldsymbol{\epsilon} - \mathbf{z}_0)$,

$$\mathbf{f}_k(\mathbf{z}, \sigma) = \mathbb{E}[\boldsymbol{\epsilon} - \mathbf{z}_0 \mid \mathbf{z}_\sigma = \mathbf{z}, k] = \frac{\mathbf{z} - \mathbf{m}_k(\mathbf{z})}{\sigma}, \qquad \mathbf{m}_k(\mathbf{z}) = \mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_\sigma = \mathbf{z}, k].$$

Hence $\mathbf{f}_G - \mathbf{f}_F = -(\mathbf{m}_G - \mathbf{m}_F)/\sigma$ for every $\mathbf{z}$ and σ.

(i) At $\sigma = 1$, $\mathbf{z}_1 = \boldsymbol{\epsilon}$ is independent of $\mathbf{z}_0$, so $\mathbf{m}_k(\mathbf{z}) = \boldsymbol{\mu}_k$ for every $\mathbf{z}$. Therefore $\mathbf{f}_G(\mathbf{z}, 1) - \mathbf{f}_F(\mathbf{z}, 1) = -\Delta\boldsymbol{\mu}$.

(ii) For $\mathbf{z}_0 \mid k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ the pair $(\mathbf{z}_0, \mathbf{z}_\sigma)$ is jointly Gaussian with

$$\mathbb{E}[\mathbf{z}_\sigma] = (1-\sigma)\boldsymbol{\mu}_k, \qquad \mathrm{Cov}(\mathbf{z}_\sigma) = (1-\sigma)^2\boldsymbol{\Sigma} + \sigma^2\mathbf{I} = \mathbf{S}_W(\sigma), \qquad \mathrm{Cov}(\mathbf{z}_0, \mathbf{z}_\sigma) = (1-\sigma)\boldsymbol{\Sigma}.$$

Gaussian conditioning gives

$$\mathbf{m}_k(\mathbf{z}) = \boldsymbol{\mu}_k + (1-\sigma)\boldsymbol{\Sigma}\mathbf{S}_W^{-1}\big(\mathbf{z} - (1-\sigma)\boldsymbol{\mu}_k\big).$$

The $\mathbf{z}$-dependent term is the same for both classes and cancels in the difference:

$$\mathbf{m}_G - \mathbf{m}_F = \Delta\boldsymbol{\mu} - (1-\sigma)^2\boldsymbol{\Sigma}\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu} = \big[\mathbf{S}_W - (1-\sigma)^2\boldsymbol{\Sigma}\big]\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu} = \sigma^2\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu},$$

using $\mathbf{S}_W - (1-\sigma)^2\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Dividing by $-\sigma$ gives Equation (1).

(iii) An Euler step multiplies the velocity by $\sigma_{i+1} - \sigma_i = -|\Delta\sigma_i|$, so the extra displacement is $+|\Delta\sigma_i|\,\sigma_i\,\mathbf{S}_W(\sigma_i)^{-1}\Delta\boldsymbol{\mu}$. The noised classes have means $(1-\sigma)\boldsymbol{\mu}_k$ and the common covariance $\mathbf{S}_W$, so their Fisher direction is $\mathbf{S}_W^{-1}(1-\sigma)\Delta\boldsymbol{\mu}$, parallel to the displacement for $\sigma < 1$. (At $\sigma = 1$ the two noised classes coincide and the Fisher direction vanishes; the displacement is then given by (i).) For the bound, write $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, $a_j = \mathbf{u}_j^\top\Delta\boldsymbol{\mu}$, $\rho_j = (1-\sigma)^2\lambda_j/\sigma^2$ and $w_j = (1+\rho_j)^{-1}$. Then

$$\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu} = \sigma^{-2}\sum_j w_j a_j \mathbf{u}_j,$$

$$\cos\big(\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\mu}\big) = \frac{\sum_j w_j a_j^2}{\sqrt{\sum_j w_j^2 a_j^2}\,\sqrt{\sum_j a_j^2}} = \frac{\mathbb{E}_p[w]}{\sqrt{\mathbb{E}_p[w^2]}}, \qquad p_j = \frac{a_j^2}{\|\Delta\boldsymbol{\mu}\|^2}.$$

Every w_j lies in $[m, M] = [(1+\rho)^{-1}, 1]$ with $\rho = (1-\sigma)^2\lambda_{\max}(\boldsymbol{\Sigma})/\sigma^2$. Kantorovich's inequality, $\mathbb{E}[w^2] \le \frac{(m+M)^2}{4mM}\,\mathbb{E}[w]^2$, gives

$$\cos\big(\mathbf{S}_W^{-1}\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\mu}\big) \ge \frac{2\sqrt{mM}}{m+M} = \frac{2\sqrt{1+\rho}}{2+\rho},$$

which tends to 1 as $\sigma \to 1$. □

A.2 Best fixed vector under unequal covariances

Proposition 3 (Best fixed vector under unequal covariances)**.** *Let $p_k = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\mathbf{S}_{W,k} = (1-\sigma)^2\boldsymbol{\Sigma}_k + \sigma^2\mathbf{I}$, $\mathbf{A} = \boldsymbol{\Sigma}_G\mathbf{S}_{W,G}^{-1} - \boldsymbol{\Sigma}_F\mathbf{S}_{W,F}^{-1}$ and $\tilde{\mathbf{z}} = \mathbf{z} - (1-\sigma)\boldsymbol{\mu}_F$. For $\sigma \in (0, 1]$,*

$$\mathbf{f}_G(\mathbf{z}, \sigma) - \mathbf{f}_F(\mathbf{z}, \sigma) = -\sigma\,\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu} - \tfrac{1-\sigma}{\sigma}\,\mathbf{A}\,\tilde{\mathbf{z}}. \tag{5}$$

Let $\mathbf{z}$ be distributed as the student's latent at level σ, $\mathbf{z} \sim \mathcal{N}((1-\sigma)\boldsymbol{\mu}_F, \mathbf{S}_{W,F})$. The fixed vector that minimises $\mathbb{E}\|\mathbf{f}_G - \mathbf{f}_F - \mathbf{u}\|^2$ is $\mathbf{u}^\star = -\sigma\,\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$. Equivalently, the best fixed displacement for a decreasing-σ step is along $+\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$. This direction obeys the bound of Proposition 1(iii) with $\lambda_{\max}(\boldsymbol{\Sigma}_G)$, and tends to $\Delta\boldsymbol{\mu}$ as $\sigma \to 1$. The remainder $-\frac{1-\sigma}{\sigma}\mathbf{A}\tilde{\mathbf{z}}$ has mean zero and

$$\mathbb{E}\big\|\tfrac{1-\sigma}{\sigma}\mathbf{A}\tilde{\mathbf{z}}\big\|^2 = \tfrac{(1-\sigma)^2}{\sigma^2}\,\mathrm{tr}(\mathbf{A}\mathbf{S}_{W,F}\mathbf{A}^\top) \;\to\; (1-\sigma)^2\|\boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_F\|_F^2 \quad \text{as } \sigma \to 1.$$

Proof. Step 1: conditional means. With $\mathbf{B}_k = (1-\sigma)\boldsymbol{\Sigma}_k \mathbf{S}_{W,k}^{-1}$, Gaussian conditioning gives, as in the proof of Proposition 1,

$$\mathbf{m}_k(\mathbf{z}) = \boldsymbol{\mu}_k + \mathbf{B}_k\big(\mathbf{z} - (1-\sigma)\boldsymbol{\mu}_k\big).$$

Step 2: the difference. Substituting $\mathbf{z} - (1-\sigma)\boldsymbol{\mu}_G = \tilde{\mathbf{z}} - (1-\sigma)\Delta\boldsymbol{\mu}$,

$$\begin{aligned}\mathbf{m}_G - \mathbf{m}_F &= \big[\mathbf{I} - (1-\sigma)\mathbf{B}_G\big]\Delta\boldsymbol{\mu} + (\mathbf{B}_G - \mathbf{B}_F)\,\tilde{\mathbf{z}} \\ &= \sigma^2 \mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu} + (1-\sigma)\,\mathbf{A}\,\tilde{\mathbf{z}},\end{aligned}$$

using $\mathbf{I} - (1-\sigma)^2\boldsymbol{\Sigma}_G\mathbf{S}_{W,G}^{-1} = \sigma^2\mathbf{S}_{W,G}^{-1}$ and $\mathbf{B}_G - \mathbf{B}_F = (1-\sigma)\mathbf{A}$. Dividing by $-\sigma$ gives Equation (5).

Step 3: the best constant. For any random vector $\mathbf{D}$ and constant $\mathbf{u}$,

$$\mathbb{E}\|\mathbf{D} - \mathbf{u}\|^2 = \mathbb{E}\|\mathbf{D} - \mathbb{E}\mathbf{D}\|^2 + \|\mathbb{E}\mathbf{D} - \mathbf{u}\|^2,$$

so the minimiser is $\mathbf{u}^\star = \mathbb{E}\mathbf{D}$. Under the student's latent distribution $\mathbb{E}\tilde{\mathbf{z}} = \mathbf{0}$, hence $\mathbf{u}^\star = -\sigma\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$. Only the teacher's covariance enters, because at the centre of the student's class the student's own estimate is $\boldsymbol{\mu}_F$ whatever $\boldsymbol{\Sigma}_F$ is. As $\sigma \to 1$, $\mathbf{S}_{W,G} \to \mathbf{I}$ and $\mathbf{u}^\star \to -\Delta\boldsymbol{\mu}$.

Step 4: the remainder. $\mathrm{Cov}(\tilde{\mathbf{z}}) = \mathbf{S}_{W,F}$ gives the trace formula. As $\sigma \to 1$, $\mathbf{S}_{W,k} \to \mathbf{I}$, so $\mathbf{A} \to \boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_F$ and $\mathrm{tr}(\mathbf{A}\mathbf{S}_{W,F}\mathbf{A}^\top) \to \|\boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_F\|_F^2$. □

A fixed vector shifts every latent by the same amount. It can move the centre of the student's class, but it cannot make the class narrower or wider. The remainder is exactly this missing part. In step form it is $+|\Delta\sigma|\frac{1-\sigma}{\sigma}\mathbf{A}\tilde{\mathbf{z}}$, proportional to the latent's offset from the centre. Along a direction in which the teacher's class is the tighter one, where $\mathbf{A}$ is negative, it draws every latent toward the centre, so the target field narrows the student's class. Where the teacher's class is wider, it widens the student's class. As a numerical check on both propositions, we also evaluate $\mathbf{f}_G - \mathbf{f}_F$ directly from its definition, by importance-weighting six million samples of $\mathbf{z}_0$ per class in three dimensions, with no Gaussian conditioning formula. The estimate matches Equation (1) to within 0.3% and Equation (5) to within 0.1%. At $\sigma = 1$ it returns $-\Delta\boldsymbol{\mu}$ exactly for a bimodal student against a skewed teacher, as Proposition 1(i) requires of any pair of distributions.

On real latents. Proposition 3 makes two asymptotic promises. As $\sigma \to 1$, the best fixed displacement $\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$ aligns with the $\Delta\boldsymbol{\mu}$ we steer with, and the energy of the remainder, the part no fixed vector captures, falls as $(1-\sigma)^2$. Both rates carry prefactors that grow with the covariance eigenvalues, and on image latents these reach the thousands. The largest belong to the per-channel constants, the 16 directions that shift one channel uniformly over space and change only the image's overall colour and brightness. Their top eigenvalue has median 1884 for the student and 1107 for the teacher. So the asymptotics alone do not say whether $\Delta\boldsymbol{\mu}$ is a good fixed vector at $\sigma = 0.95$. Table 3 therefore evaluates both statements with the covariances measured on the 200 prompts, treating each channel as spatially stationary. The two classes are measurably unequal (the teacher's within-class variance is 0.73 of the student's in total and 0.63 in the per-channel constants), so the unequal-covariance form is the one to evaluate. The table supports three claims about the first steps, which receive most of the push (last column). First, $\Delta\boldsymbol{\mu}$ is nearly the best fixed vector there. Its cosine with $\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$ stays at 0.95 or above down to $\sigma = 0.95$. Second, the same cosine computed with the pooled covariance is lower, so the shared-covariance model of Proposition 1 is the pessimistic reading, not a favourable idealisation. Third, the remainder is small where steering acts. It holds 1% of the field difference's energy at the first step and 24% at $\sigma = 0.5$. Of the first-step remainder, 81% lies in the per-channel constants. That part carries no capability. Steering with it alone reaches at most 14.9% (Table 7).

Table 3: **Proposition 3 evaluated with measured FLUX spectra** (200 prompts). Columns 2–4: cosine between the best fixed displacement $\mathbf{S}_{W,G}^{-1}\Delta\boldsymbol{\mu}$ and the $\Delta\boldsymbol{\mu}$ we steer with, computed with the teacher's measured covariance (median over prompts, and the worst prompt) and with the pooled covariance of the two classes, the shared-covariance model of Proposition 1. Columns 5–6: share of $\mathbb{E}\|\mathbf{f}_G - \mathbf{f}_F\|^2$ that no fixed vector captures (the remainder of Equation (5); spectra estimated on disjoint halves of the images to debias it), and the part of that remainder in the per-channel constants. Last column: cumulative share of the $\beta = 10$ push injected by the time the sampler reaches σ.

	best fixed displacement: cosine with $\Delta\boldsymbol{\mu}$			remainder		schedule
σ	median	worst prompt	pooled $\boldsymbol{\Sigma}$	share	constant (%)	injected (%)
0.988	0.999	0.86	0.998	0.01	81	16
0.976	0.993	0.68	0.982	0.04	59	30
0.963	0.976	0.55	0.948	0.05	38	42
0.950	0.949	0.48	0.908	0.08	23	53
0.936	0.919	0.44	0.866	0.09	16	62
0.905	0.843	0.39	0.766	0.12	8	77
0.850	0.744	0.34	0.655	0.16	3	90
0.700	0.544	0.22	0.457	0.21	0	–
0.500	0.382	0.14	0.333	0.24	0	–

A.3 Proof of Proposition 2

Proof. Step 1: the decomposition. Averaging $\mathbf{z}_\sigma^{k,i} = (1-\sigma)\mathbf{z}_0^{k,i} + \sigma\boldsymbol{\epsilon}_i^k$ over i within each class and subtracting the two class means gives

$$\widehat{\Delta\boldsymbol{\mu}}(\sigma) = (1-\sigma)\,\widehat{\Delta\boldsymbol{\mu}}_0 + \sigma\,\boldsymbol{\eta}, \qquad \boldsymbol{\eta} = \bar{\boldsymbol{\epsilon}}_G - \bar{\boldsymbol{\epsilon}}_F.$$

Step 2: the noise term. A mean of N standard Gaussian vectors is $\mathcal{N}(\mathbf{0}, \mathbf{I}/N)$, so $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \frac{2}{N}\mathbf{I})$, independent of the images. $\|\boldsymbol{\eta}\|^2$ is $\frac{2}{N}$ times a χ^2_d variable, with mean $2d/N$ and relative spread $\sqrt{2/d}$, so $\|\boldsymbol{\eta}\| = \sqrt{2d/N}\,(1 + O(d^{-1/2}))$. The direction of $\boldsymbol{\eta}$ is uniform on the sphere.

Step 3: orthogonality. The component of $\boldsymbol{\eta}$ along the fixed unit vector $\widehat{\Delta\boldsymbol{\mu}}_0/\|\widehat{\Delta\boldsymbol{\mu}}_0\|$ is $\mathcal{N}(0, 2/N)$, which is $O(d^{-1/2})$ relative to $\|\boldsymbol{\eta}\|$. The two terms of Step 1 are therefore orthogonal up to $O(d^{-1/2})$.

Step 4: the cosine. By Step 3 the estimate is the hypotenuse of a right triangle with a signal leg of length $a = (1-\sigma)\|\widehat{\Delta\boldsymbol{\mu}}_0\|$ and a noise leg of length $b = \sigma\|\boldsymbol{\eta}\|$, so Pythagoras gives

$$\cos\big(\widehat{\Delta\boldsymbol{\mu}}(\sigma), \widehat{\Delta\boldsymbol{\mu}}_0\big) = \frac{a}{\sqrt{a^2+b^2}} + O(d^{-1/2}) = \frac{r}{\sqrt{1+r^2}} + O(d^{-1/2}),$$
$$r = \frac{a}{b} = \frac{1-\sigma}{\sigma}\,\|\widehat{\Delta\boldsymbol{\mu}}_0\|\sqrt{\frac{N}{2d}} = \frac{1-\sigma}{\sigma}\,\kappa,$$

which is Equation (3). The two terms have equal norm when $r = 1$, that is at $\sigma^\star = \kappa/(1+\kappa)$, where the cosine is $1/\sqrt{2}$. As $\sigma \to 1$, $r \to 0$ and the estimate tends to $\boldsymbol{\eta}$. Two fits on the same images under independent draws $\boldsymbol{\eta}, \boldsymbol{\eta}'$ share only the first term, so their cosine is $a^2/(a^2+b^2) = r^2/(1+r^2)$. □

Three consequences follow from Step 1.

- Draws shared across the two classes give $\boldsymbol{\eta} = \mathbf{0}$. The estimate is then $(1-\sigma)\widehat{\Delta\boldsymbol{\mu}}_0$, the clean estimator scaled down.
- Averaging M draws per image replaces $\boldsymbol{\eta}$ by $\boldsymbol{\eta}/\sqrt{M}$.
- For a fixed linear map $\mathbf{B}$ applied to the estimate, the ratio r becomes $(1-\sigma)\|\mathbf{B}\widehat{\Delta\boldsymbol{\mu}}_0\|/(\sigma\sqrt{2/N}\,\|\mathbf{B}\|_F)$. The proposition therefore makes no claim about every whitened estimator. For PCA fitted on the noised sample itself, the subspace is the span of the centred sample and contains $\boldsymbol{\eta}$ whole; that is the case measured in Table 4.

The sweep behind Figure 2. For each prompt, the mean difference is fit on ten of the twenty images per model, noised to level σ with noise drawn independently for every prompt. It is compared by signed cosine with the clean mean difference of the other ten. Even at $\sigma = 0$ this comparison gives only $c_0 = 0.80$ (Table 4). Ten images carry their own sampling error, and c_0 is the split-half ceiling of Appendix A.4. Write $\mathbf{m}$ for the held-out clean estimate. The unit vector along $\widehat{\Delta\boldsymbol{\mu}}_0$ is then $c_0\,\mathbf{m}/\|\mathbf{m}\| + \sqrt{1-c_0^2}\,\mathbf{e}$, where $\mathbf{e} \perp \mathbf{m}$ is the fitting half's own sampling error. The noise direction is orthogonal to both $\mathbf{m}$ and $\mathbf{e}$. The inner product of the noised fit with $\mathbf{m}$ is therefore $a\,c_0$, and its norm is $\sqrt{a^2+b^2}$. The predicted cosine is the product $c_0\,r/\sqrt{1+r^2}$. Each error tilts the vector into a fresh orthogonal direction, so the cosines multiply. The crossover is $\sigma^\star = 0.61$ here ($N = 10$ per fit) and 0.67 (range 0.49–0.86) at the method's $N = 20$. It grows with N through κ and does not depend on c_0. PCA+LDA uses all $2N - 1 = 19$ components the twenty centred samples span, whitened, with the sign chosen from the noised sample. It trails the raw fit already at $\sigma = 0$. Dividing by a within-class scatter estimated from twenty samples overfits before any noise is added. The zoo rows of Table 10 show the noised fits steering progressively worse.

Table 4: **Noise contamination on FLUX latents** (200 prompts, medians). The mean difference is fit on ten images per class noised to the row's σ. Every entry is a cosine. *Vs. held-out clean* $\Delta\boldsymbol{\mu}$: the noised fit against the clean fit of the other ten images. The same comparison without noise gives $c_0 = 0.80$. The law is $c_0\,r/\sqrt{1+r^2}$ with no fitted parameter. Here r is the signal-to-noise length ratio of Proposition 2. *Two fits, independent draws*: the same ten images noised twice, one fit against the other. The law is $r^2/(1+r^2)$. *16 draws averaged*: as the first pair, with 16 noise draws averaged per image. *PCA+LDA*: as the first pair, for the whitened fit. With draws shared between the classes the first column equals c_0 at every σ.

	vs. held-out clean $\Delta\boldsymbol{\mu}$		two fits, independent draws		16 draws	PCA+LDA
fitting σ	measured	law	measured	law	averaged	
0	0.799	0.799	1.000	1.000	0.799	0.437
0.2	0.789	0.789	0.975	0.975	0.798	0.451
0.4	0.730	0.730	0.844	0.844	0.795	0.476
0.5	0.670	0.670	0.706	0.706	0.789	0.471
0.6	0.565	0.565	0.516	0.517	0.778	0.422
0.7	0.429	0.430	0.307	0.306	0.745	0.334
0.8	0.278	0.279	0.132	0.131	0.670	0.231
0.9	0.132	0.131	0.029	0.029	0.441	0.096
0.95	0.064	0.063	0.007	0.007	0.240	0.013
0.99	0.013	0.012	0.000	0.000	0.049	0.001

A.4 How good is a mean of N images

Write $\widehat{\Delta\boldsymbol{\mu}}_N = \Delta\boldsymbol{\mu} + \mathbf{n}_N$, where $\mathbf{n}_N$ is the sampling error of the two class means, with $\mathbb{E}\|\mathbf{n}_N\|^2 = (\mathrm{tr}\boldsymbol{\Sigma}_G + \mathrm{tr}\boldsymbol{\Sigma}_F)/N$. Let $\nu_1 = (\mathrm{tr}\boldsymbol{\Sigma}_G + \mathrm{tr}\boldsymbol{\Sigma}_F)/\|\Delta\boldsymbol{\mu}\|^2$ be the noise-to-signal ratio of a single image pair. In high dimension $\mathbf{n}_N$ is nearly orthogonal to $\Delta\boldsymbol{\mu}$ and to independent errors, which gives three cosines:

$$\cos(\widehat{\Delta\boldsymbol{\mu}}_N, \Delta\boldsymbol{\mu}) = (1+\nu_1/N)^{-1/2} \qquad \text{estimate against the population direction,}$$

$$\cos(\widehat{\Delta\boldsymbol{\mu}}_M, \widehat{\Delta\boldsymbol{\mu}}'_M) = (1+\nu_1/M)^{-1} \qquad \text{two estimates from disjoint sets of } M \text{ images,}$$

$$\cos(\widehat{\Delta\boldsymbol{\mu}}_N, \widehat{\Delta\boldsymbol{\mu}}_{20}) = \sqrt{\frac{1+\nu_1/20}{1+\nu_1/N}} \qquad N \text{ of the twenty images against all twenty.}$$

The first is the right triangle of the proof of Proposition 2, with signal leg $\|\Delta\boldsymbol{\mu}\|$ and noise leg $\|\mathbf{n}_N\|$. The second has an independent noise leg on each estimate. Only the signal survives the inner product, so the cosine is the first formula squared with M in place of N. The third follows because the shared images contribute $\langle \mathbf{n}_N, \mathbf{n}_{20}\rangle = \frac{\nu_1}{20}\|\Delta\boldsymbol{\mu}\|^2$. The measured split-half cosine 0.81 gives $\nu_1 = 2.3$ by the second formula (a direct estimate from the sample variances gives 2.6). The third formula then reproduces every measured point of Figure 2c. The first gives cosines with the population direction $\Delta\boldsymbol{\mu}$ of 0.68, 0.83, 0.90 and 0.95 for $N = 2, 5, 10, 20$.

A.5 What a Gaussian flow predicts for transport

For $p = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the sampling ODE is linear. Consider an eigen-direction of $\boldsymbol{\Sigma}$ with eigenvalue λ. Write $y = z - (1-\sigma)\mu$ for the coordinate along it and $S(\sigma) = (1-\sigma)^2\lambda + \sigma^2$. Substituting $\mathbf{m}(\mathbf{z})$ from the proof of Proposition 1 into $dz/d\sigma = (z-m)/\sigma$ gives

$$\frac{dy}{d\sigma} = y\,\frac{\sigma - (1-\sigma)\lambda}{S(\sigma)} = y\,\frac{d}{d\sigma}\,\tfrac{1}{2}\ln S(\sigma), \qquad \text{so} \qquad y(\sigma) \propto \sqrt{S(\sigma)}.$$

A perturbation injected at level σ therefore reaches the image, where $S(0) = \lambda$, multiplied by $\sqrt{\lambda/S(\sigma)}$. The flow stretches unit-variance noise into image variance. With the student's measured spectrum this predicts a gain of 21 at the first step and 2.0 at $\sigma = 0.51$ for a push with $\Delta\boldsymbol{\mu}$'s spectrum, against the measured 55 and 2.4 (Table 5). That is the right order and the right decay, because $\Delta\boldsymbol{\mu}$ lies along the directions in which the student's own images vary most. For an isotropic push the model predicts a contraction (0.7) where the real flow expands (5.1), and it assigns a phase-scrambled $\Delta\boldsymbol{\mu}$ the same gain as $\Delta\boldsymbol{\mu}$. A second-moment model is a guide to the direction, not to transport.

Table 5: **Transport gain relative to isotropic random directions**, $\|\delta\| = 10$, medians over 6 prompts $\times$ 3 seeds; last rows give the absolute gain.

direction	σ=0.99	σ=0.95	σ=0.89	σ=0.76	σ=0.51
DFS direction $\Delta\mu$ (clean latents)	8.80	5.50	4.00	2.89	1.52
phase-scrambled $\Delta\mu$ (same spectrum)	3.45	3.01	3.12	2.38	1.64
isotropic Gaussian	0.90	0.97	0.97	1.04	1.00
absolute gain $\|\Phi\delta\|/\|\delta\|$, $\Delta\mu$, $\|\delta\|$=10	55.4	21.6	11.7	6.0	2.4
absolute gain, random, $\|\delta\|$=10	5.1	3.8	2.8	1.9	1.5

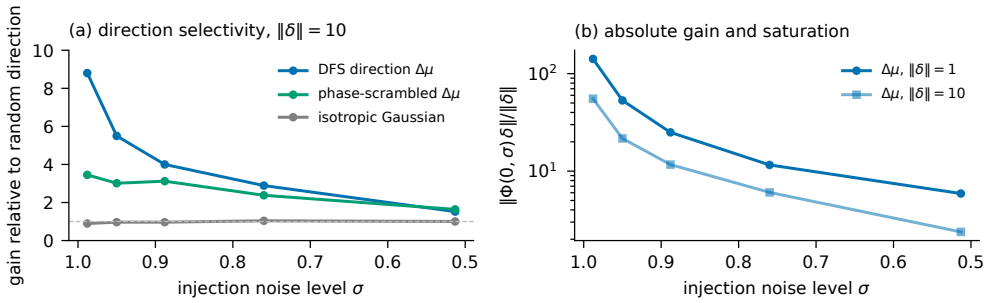


Figure 4: **Perturbation transport in FLUX.1-dev** (medians over 6 prompts $\times$ 3 seeds). (a) Gain relative to random directions at the same σ, $\|\delta\| = 10$: $\Delta\boldsymbol{\mu}$ is amplified 8.8$\times$ more than random at $\sigma = 0.988$, a spectrum-matched control 3.5$\times$; the advantage decays toward parity by $\sigma \approx 0.5$. (b) Absolute gain for $\Delta\boldsymbol{\mu}$ falls from 142 ($\|\delta\| = 1$) and 55 ($\|\delta\| = 10$) at the first step to a few units mid-trajectory; the gap between the curves is the saturation.

B DFS-CapBench details

Families and sources. The 200 prompts span 20 families; Table 6 lists the strict-tier families with at least four prompts. 171 prompts were authored by a language model from family templates, with a difficulty calibration that favours few, sharp constraints (stacked conjunctions make the teacher fail the all-seed gate). 20 were hand-written to contradict a strong prior with a single short clause. The remaining 9 survive from the 80 hand-curated prompts of the pilot described below. They were checklisted and gated like every other candidate.

Checklists. Each prompt's checklist was authored once from the prompt text by a language model (Gemini 3.1 Pro for the first batches, Qwen3-VL-32B in text-only mode for the last) with the in-

struction to add exclusivity items so that over-generation cannot pass; item kinds are attribute, exclusion, count, spatial, relation, presence, structure, hand-written and text (1401 items, 262 of them exclusions and 219 counts). Checklists were frozen at authoring and are identical for the gate, the baselines and every steered arm.

Judge. Qwen3-VL-32B-Instruct, NF4 double quantisation, one forward pass per (image, item), system instruction "*You are a precise visual verification tool. Answer the question about the image with exactly one word: Yes or No. Judge only what is visible. If a property is ambiguous or not visible, answer No.*", $P(\text{yes})$ from the softmax over the Yes and No continuations. The judge was chosen over an API model for reproducibility (a pinned revision can be re-run), graded scores and cost.

Gate. Four fixed seeds per candidate for student and teacher; the teacher must pass all four; strict = student passes none, relaxed = at most one; near-duplicate prompts removed at embedding similarity 0.55; 28 borderline prompts flagged and vetoed by an author. Of the 1909 scored candidates, teacher unreliability removes about 60%. The failing teacher items have median $P(\text{yes})$ of 0.09, so the gate is not judge noise. The evaluation uses the same four seeds, so the strict tier also contains prompts the student sometimes gets right. A prompt with true student pass rate p enters it with probability $(1-p)^4$. On four fresh seeds the student passes 10.1% of strict images. DFS is unaffected (43.3% against 42.1%, paired difference 1.2 $[-3.0, 5.4]$). Another family's $\Delta\boldsymbol{\mu}$ at $\alpha = 7$ passes 10.7%, the fresh-seed floor. An isotropic random direction barely changes the image and passes 2.8%. Every arm that alters the image carries a floor of about ten points.

Judge validation. Two raters independently audited the same 200 random (image, item) judgements, sampled balanced between judge yes and judge no verdicts. Their agreement with the judge is 82.5% (Cohen's $\kappa = 0.65$) and 70.0% ($\kappa = 0.40$). The raters agree with each other on 72.5% ($\kappa = 0.43$). On the 145 items where they agree, the judge matches their consensus on 86.2% ($\kappa = 0.71$).

Why not a preference study. A pilot with 80 hand-curated prompts used a two-alternative preference between steered and unsteered images, with teacher references shown. The automatic judge preferred the steered image 83.8% of the time when it was shown on the left and 68.3% on the right. That protocol measures teacher-likeness on prompts selected as student failures, with a position bias. It motivated the verified-gap, teacher-free design and is not used anywhere in this paper.

Table 6: **Strict-tier pass rate (%) by family** for families with at least four prompts; best method per row in bold. NAG at scale 5, LoRA at scale 1, DFS at $\alpha = 7$.

family	prompts	student	rewrite	neg. CFG	NAG	LoRA	DFS
negation-exclusion	24	1	45	33	57	**75**	23
document-layout	18	0	14	6	12	26	**50**
diagram-structure	13	0	15	2	6	33	**77**
medium-fidelity	12	0	19	40	33	**50**	33
spatial-relation	12	2	21	21	42	**60**	31
projection-convention	11	0	5	5	36	**66**	36
viewpoint	8	0	47	25	19	**78**	34
counting	6	0	38	29	25	54	**67**
subtraction	6	0	50	67	**100**	96	33
text-rendering	4	0	**75**	19	19	75	38
attribute-binding	4	0	25	19	25	**75**	50

C SETUP AND BASELINE DETAILS

Hyperparameters. Student FLUX.1-dev, 28 Euler steps with the dynamic-shift schedule, guidance 3.5, 1024×1024, max sequence length 512. Direction: $N = 20$ teacher images (FLUX.2-klein-9B, same prompt, seeds fixed), $N = 20$ student latents with seeds drawn once; teacher images

encoded with the student VAE at 1024^2 and mapped to the scaled latent space. Steering: $\beta = 10$, $h_{\min} = 10^{-4}$, $\alpha \in \{2, 5, 7, 10, 20\}$, evaluation seeds $\{42, 123, 456, 789\}$, direction applied in the transformer's packed latent layout, arithmetic in float32. Algorithm 1 gives the whole method.

Algorithm 1 Discriminant Flow Steering

Require: N teacher images $\{\mathbf{x}_G^{(i)}\}$, prompt c, strength α, schedule exponent β, threshold $h_{\min}$

1: $\mathbf{z}_0^{G,i} \leftarrow \mathcal{E}_F(\mathbf{x}_G^{(i)})$; $\mathbf{z}_0^{F,i} \leftarrow$ final latent of student sampling on c, seed i ▷ $2N$ latents, once per prompt
2: $\mathbf{v} \leftarrow \widehat{\Delta\boldsymbol{\mu}}/\|\widehat{\Delta\boldsymbol{\mu}}\|$ with $\widehat{\Delta\boldsymbol{\mu}}$ from Equation (2)
3: **for** each generation: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; $i = 0, \ldots, T-1$ **do**
4: $\quad \mathbf{z} \leftarrow \mathbf{z} + (\sigma_{i+1} - \sigma_i)\, \mathbf{f}_F(\mathbf{z}, \sigma_i)$; $\quad h \leftarrow \sigma_{i+1}^{\beta}$
5: $\quad$ **if** $h \geq h_{\min}$ **then** $\mathbf{z} \leftarrow \mathbf{z} + \alpha h \mathbf{v}$
6: $\quad$ **end if**
7: **end for**
8: **return** decoded $\mathbf{z}$

Prompt rewrite and negative prompts. Qwen3-VL-32B-Instruct receives the prompt and three of the twenty teacher images (resized to 512^2) with the instruction: *"The first line is a text-to-image prompt. The images that follow were produced by a stronger model and all satisfy it correctly. A weaker model fails this prompt. Output exactly two lines: REWRITE: <the prompt rewritten so the weaker model produces images like these—be explicit about whatever the images make true that a naive reading would miss, under 60 words>. NEGATIVE: <a short comma-separated list of the wrong content the weaker model produces instead, under 20 words>."* The rewritten prompt is sampled at guidance 3.5; the negative prompt feeds the true-CFG baseline (`true_cfg_scale` $\in \{2, 4\}$) and NAG.

NAG. Normalized attention guidance (Chen et al., 2026) with the upstream implementation for FLUX, the same MLLM-written negative prompt, $\tau = 2.5$, $\alpha_{\text{NAG}} = 0.25$, scale $\in \{3, 5, 8\}$; scale 5 is the implementation's default for FLUX and was declared as the reported strength before the run. NAG runs a second attention branch at every step, the extra forward pass in Table 1.

LoRA. Per prompt, a rank-16 LoRA on the attention and feed-forward projections of the FLUX.1-dev transformer, trained for 500 steps at learning rate 10^{-4} on the same twenty teacher latents used for $\Delta\boldsymbol{\mu}$, with the benchmark prompt as caption, logit-normal timestep sampling through the FLUX shift, guidance embedding 3.5 and no prior preservation (the diffusers DreamBooth-FLUX recipe); inference with the original prompt at adapter scale $\in \{0.5, 1, 1.5\}$, scale 1 declared in advance. About 17 minutes per prompt on one A100-40GB. The pooled variant trains one rank-16 adapter for 3000 steps on all strict prompts' teacher latents at once.

Concept Sliders. A rank-4 slider on the attention projections trained for 300 steps at 2×10^{-4} with the teacher latents at $+1$ and the student latents at -1, applied at scales $\{1, \ldots, 5\}$ and, because every scale above 1 produced noise images on every prompt (pass rate $\leq 3\%$), also at $\{0.25, 0.5, 0.75\}$. Table 1 reports the best scale of the whole sweep (1); the scale declared in advance, 3, gives 0.6%.

Controls. A decode–encode round trip of the student's latents, which makes them pass through the VAE exactly as the teacher images do, changes $\Delta\boldsymbol{\mu}$ by at most 4% (cosine 0.96 or better on 8 prompts). The direction is therefore not an artifact of encoding one class and not the other. Shrinkage LDA uses Ledoit–Wolf covariance (Ledoit & Wolf, 2004) in the $2N - 1$ dimensional span of the centred latents.

D Additional results

Table 7 gives every arm at every strength. Table 8 gives the paired prompt-level overlap of DFS with each alternative. Table 13 covers the relaxed tier and all 200 prompts. Table 9 gives strength selec-

tion, Tables 10 and 11 the estimator zoo and Table 12 output diversity. The transport measurement of Section 3.3 is in Appendix A.5.

Table 7: **Strict tier, all arms and strengths** (columns of Table 1, plus the 95% bootstrap CI of the pass rate over prompts). The pick column takes the best image over the arm's sweep: $\alpha \in \{5, 7, 10\}$ for $\Delta\boldsymbol{\mu}$ and $\{5, 10\}$ for the other latent directions (strengths above 10 pass mostly blurred images), the three stack settings, NAG $\{3, 5, 8\}$, LoRA $\{0.5, 1, 1.5\}$, CFG $\{1.5, 7, 10\}$, negative CFG $\{2, 4\}$. Concept Sliders is shown at its best scale (1) because scales above 1 produce noise images with this recipe.

			at the strength in the row label				best strength
Method	tch	fwd	pass %	95% CI	solved %	any %	pick %
Student (unsteered)	–	0	0.4	[0.0, 1.0]	0.0	1.6	–
on four fresh seeds	–	0	10.1	[6.9, 13.5]	1.6	29.4	–
CFG sweep, best of $\{1.5, 7, 10\}$	–	0	12.3	[8.3, 16.9]	7.9	26.2	25.6
Negative-prompt CFG, best of $\{2, 4\}$	3	1×	22.8	[17.5, 28.6]	12.7	43.7	31.3
Prompt rewrite by MLLM	3	0	30.4	[24.4, 36.7]	19.0	55.6	–
Control: another prompt's $\Delta\mu$, α=7	20	0	10.7	[7.1, 14.9]	5.6	23.8	20.0
Control: random direction, α=7	–	0	2.8	[1.2, 4.8]	0.8	8.7	–
Per-channel constant of $\Delta\mu$ only, α=10	20	0	14.9	[10.1, 19.4]	7.9	31.7	24.2
Concept Sliders, best scale in [0.25–5]	20	0	25.6	[19.6, 31.2]	17.5	46.8	36.9
NAG, scale 5	20	1×	35.9	[29.2, 42.9]	26.2	60.3	51.0
NAG, scale 8	20	1×	44.8	[37.9, 51.4]	35.7	67.5	51.0
LoRA per prompt, same teacher images	20	0	58.3	[51.4, 65.3]	54.0	80.2	75.0
LoRA pooled over all prompts	20	0	41.1	[34.3, 47.2]	31.0	69.8	58.7
DFS, shrinkage-LDA direction, α=5	20	0	39.9	[33.9, 46.2]	31.7	67.5	55.2
DFS, $\Delta\mu$ constant removed, α=10	20	0	40.5	[33.9, 47.2]	34.1	63.5	54.6
DFS, $\Delta\mu$, α=5	20	0	37.1	[30.6, 43.3]	28.6	61.9	59.5
DFS, $\Delta\mu$, α=7	20	0	**42.1**	[35.1, 48.6]	**33.3**	**65.9**	**59.5**
on four fresh seeds	20	0	**43.3**	[37.1, 49.6]	**33.3**	**73.0**	**56.0**
DFS, $\Delta\mu$, α=10	20	0	42.1	[35.5, 48.4]	34.9	66.7	59.5
DFS α=7 + NAG scale 5	20	1×	**63.9**	[57.3, 70.0]	**57.9**	**86.5**	**75.2**
DFS α=5 + NAG scale 5	20	1×	61.7	[54.8, 68.5]	55.6	82.5	75.2
DFS α=7 + NAG scale 8	20	1×	62.7	[56.2, 69.4]	57.1	84.1	75.2

Table 8: **Paired prompt-level comparison, strict tier** (prompt solved = $\geq 3/4$ seeds pass). Columns: solved by both, by DFS ($\alpha = 7$) only, by the other arm only, by neither.

other arm	both	DFS only	other only	neither
Prompt rewrite	13	29	11	73
Negative-prompt CFG	7	35	9	75
CFG 1.5	4	38	6	78

Table 9: **Strength selection, strict tier.** Selection rows use $\alpha \in \{5, 7, 10\}$; the leak-free row picks α per prompt on two seeds and scores the other two, and its gain does not replicate on fresh seeds. Oracle rows are upper bounds at three samples per image; they are what a user who inspects several strengths obtains, not a tuning technique. The control rows give the floor of each rule. The same rows for the baselines: NAG 35.9 fixed, 38.7 leak-free, 51.0 per-image pick; LoRA 58.3, 62.1, 75.0; the DFS+NAG stack 63.9 fixed and 75.2 per-image pick.

rule	pass %	cost
fixed global α=2	15.9	1×
fixed global α=5	37.1	1×
fixed global α=7	42.1	1×
fixed global α=10	42.1	1×
fixed global α=20	19.8	1×
per-prompt α: chosen on 2 seeds, scored on 2	45.8	1×
oracle α per prompt, same seeds	55.2	3×
user picks best α per image	59.5	3×
fresh seeds: fixed α=7	43.3	1×
fixed α=5	35.3	1×
fixed α=10	37.7	1×
per-prompt α, 2 seeds / 2 seeds	41.7	1×
oracle α per prompt	52.4	3×
best α per image	56.0	3×
control direction: fixed α=7	10.7	1×
per-prompt α, 2 seeds / 2 seeds	11.3	1×
oracle α per prompt	18.1	3×
best α per image	20.0	3×

Table 10: **Estimator zoo**: every direction is fit from the same 20+20 clean latents (or their noised versions) on a 30-prompt strict subset, unit-normalised, steered with the same schedule and seeds, and reported at its own best global α. cos: median cosine with $\Delta\boldsymbol{\mu}$ (the top principal component is below 0.8 on 10 prompts); Δ: paired difference from $\Delta\boldsymbol{\mu}$ in points with a 95% bootstrap interval over prompts; Δ_{lf}: the same with α chosen per prompt on two seeds and scored on the other two; pick: a user keeps the best of the five strengths per image. PCA uses every component the sample allows (39). The fresh-seed student passes 10.0% on this subset. Every alternative fits a model or tunes a hyperparameter (shrinkage target, regulariser, margin, kernel iterations) where $\Delta\boldsymbol{\mu}$ is a subtraction. At this sample size shrinkage LDA returns almost the same vector as $\Delta\boldsymbol{\mu}$ (cos column). Under leak-free strength selection (Δ_{lf}) no alternative is above $\Delta\boldsymbol{\mu}$.

direction estimator	cos	$\alpha^\star$	pass %	Δ vs. $\Delta\boldsymbol{\mu}$	Δ_{lf}	pick %
Class-mean difference $\Delta\boldsymbol{\mu}$ (DFS)	1.00	10	45.0	–	–	61.7
Shrinkage LDA	0.99	10	45.0	+0.0 [−5.0, +5.0]	−4.2	67.5
Diagonal whitening $\mathbf{D}^{-1}\Delta\boldsymbol{\mu}$	0.73	10	24.2	−20.8 [−31.7, −10.0]	−25.0	42.5
Logistic regression weights	0.84	7	48.3	+3.3 [−6.7, +13.3]	−1.7	67.5
Linear SVM weights	0.78	10	45.0	+0.0 [−9.2, +10.0]	−3.3	69.2
RFM / AGOP top eigenvector	0.95	7	50.0	+5.0 [−4.2, +15.0]	−3.3	68.3
Top principal component	0.97	10	28.3	−16.7 [−30.0, −4.2]	−18.3	41.7
PCA+LDA, latents noised to σ=0.5	0.51	7	32.5	−12.5 [−24.2, −0.8]	−15.0	46.7
PCA+LDA, latents noised to σ=0.9	0.17	15	21.7	−23.3 [−34.2, −13.3]	−28.3	33.3
PCA+LDA, latents noised to σ=0.99	0.00	15	6.7	−38.3 [−53.3, −23.3]	−40.8	9.2

Table 11: **Pass rate against the number of images per model**, 30-prompt strict subset; $\Delta\boldsymbol{\mu}$ from N student and N teacher latents (fixed random subsets of the twenty), each at its own best α. The cosine is with the $N = 20$ estimate that contains the N images; against the population direction the model of Appendix A.4 gives 0.68, 0.83, 0.90 and 0.95. Paired against $N = 20$: -8.3 $[-20.8, +3.3]$ points for $N = 2$ and $+1.7$ $[-6.7, +10.0]$ for $N = 10$.

N per class	cos with N=20	$\alpha^\star$	pass %	95% CI	solved %	pick %
2	0.68	10	36.7	[24.2, 49.2]	26.7	45.8
5	0.87	10	41.7	[29.2, 53.3]	26.7	51.7
10	0.95	10	46.7	[34.2, 59.2]	30.0	59.2
20	1.00	10	45.0	[32.5, 59.2]	36.7	61.7

Table 12: **Seed diversity and teacher proximity** on the strict tier, DINOv2 and LPIPS features. Diversity: mean pairwise distance across the four evaluation seeds of a prompt (higher is more diverse). Proximity: mean over outputs of the cosine to the nearest of the twenty teacher images; the teacher row gives the leave-one-out value among the teacher images themselves. At most 0.2% of any arm's outputs are closer to a teacher image than the teacher images are to each other.

arm	DINO diversity	LPIPS diversity	nearest-teacher DINO cosine
Teacher (4 of its 20 images)	0.134	0.375	0.930
Student	0.243	0.452	0.713
Prompt rewrite	0.235	0.387	0.750
NAG, scale 5	0.257	0.413	0.722
LoRA pooled	0.258	0.477	0.778
LoRA per prompt	0.223	0.393	0.816
DFS $\Delta\mu$, α=7	0.242	0.414	0.782

Table 13: **Relaxed tier (74 prompts) and all 200 prompts.** GapClosed uses the student's own pass rate as the floor and the gated teacher (1.0) as the ceiling. GapClosed $= (S_A - S_{\text{student}})/(S_{\text{teacher}} - S_{\text{student}})$.

	relaxed			all 200		
arm	pass %	solved %	GapClosed	pass %	solved %	GapClosed
Student	25.0	0.0	0.00	9.5	0.0	0.00
DFS $\Delta\mu$, α=5	56.4	44.6	0.42	44.2	34.5	0.38
DFS $\Delta\mu$, α=7	60.1	50.0	0.47	48.8	39.5	0.43
DFS $\Delta\mu$, α=10	47.6	37.8	0.30	44.1	36.0	0.38
DFS shrinkage-LDA, α=7	54.4	44.6	0.39	54.4	44.6	0.50

D.1 SECOND STUDENT

DFS-CapBench is re-gated for Stable Diffusion 3-medium (Esser et al., 2024). A prompt enters the new tier if the klein teacher passes all four seeds and SD3-medium fails all four. After the gate, 103 strict prompts with teacher image sets remain. On them DFS passes 24.3% at its best strength, $\alpha = 15$, the largest swept (Table 15); Figure 6 shows solved examples. On the 74 where the rewrite was also run it passes 25.3% against 17.9%, a difference of +7.4 $[-2.0, +16.6]$ points that is not significant, where FLUX on the same prompts gives 34.5% against 25.7%. The effect is about 0.6 of FLUX's. The paragraph below traces it to the flow, not to the direction or the VAE (Table 16).

Same-latent-space oracle. SD3.5-medium shares SD3-medium's VAE and text encoders, so with it as teacher a real network plays the target field and $\mathbf{f}_G - \mathbf{f}_F$ can be evaluated along the student's trajectories. The target-ward displacement $-(\mathbf{f}_G - \mathbf{f}_F)$ has median cosine 0.34 with the precomputed $\Delta\boldsymbol{\mu}$ at the first step, is positive on 91% of seed-steps above $\sigma = 0.98$, and decays to 0.04 at $\sigma = 0.9$ (Figure 5). This is the sign and the decay of Proposition 1, not its magnitude, which assumes ideal unguided fields.

Table 14: **SD3-medium student**, klein teacher. The student row covers all 184 gated prompts, DFS the 103 with teacher image sets, and the rewrite 74 of those, on which DFS passes 25.3%.

arm	teacher	pass %	95% CI	solved %
Student SD3-medium (unsteered)	–	0.0	[0.0, 0.0]	0.0
Prompt rewrite by MLLM	klein imgs	17.9	[11.1, 26.0]	13.5
DFS $\Delta\boldsymbol{\mu}$, α=15	klein imgs	24.3	[18.0, 31.1]	15.5

Table 15: **Strength selection for the SD3-medium student** (same protocol as Table 9).

rule	pass %	cost
klein → SD3-medium (103 prompts): fixed global α=15	24.3	1×
per-prompt α chosen on 2 seeds, scored on 2	25.7	1×
oracle α per prompt, same seeds	36.4	5×
user picks best α per image	43.0	5×

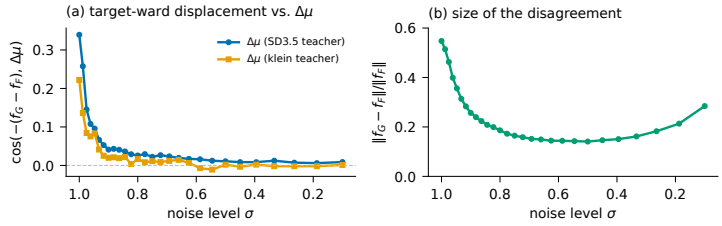


Figure 5: **Alignment of the target-ward displacement $-(\mathbf{f}_G - \mathbf{f}_F)$ of a real second network with the precomputed $\Delta\boldsymbol{\mu}$** against noise level (SD3-medium student, SD3.5-medium teacher, 184 prompts × 4 seeds; medians). The sign and the decay follow Propositions 1 and 3. For ideal fields the first-step value would be about 0.95; guided networks, and a $\Delta\boldsymbol{\mu}$ computed from finished samples, fall well short of that.

Why the effect is smaller on SD3-medium. On the 103 prompts shared by both students, DFS passes 42.0% of images on FLUX.1-dev and 24.3% on SD3-medium. Three explanations were tested (Table 16). The direction is not noisier: split-half cosine 0.84 on SD3 against 0.82 on FLUX from the same images. The VAE does not lose the capability: klein teacher images pass the judge at 87.5% as generated, 88.0% after an SD3-VAE round trip and 87.9% after a FLUX-VAE round trip. What differs is the flow: with the same transport protocol as Section 3.3, SD3-medium amplifies $\Delta\boldsymbol{\mu}$ over a random direction 4.1× at $\sigma \approx 0.99$ where FLUX amplifies it 8.8×, because it amplifies

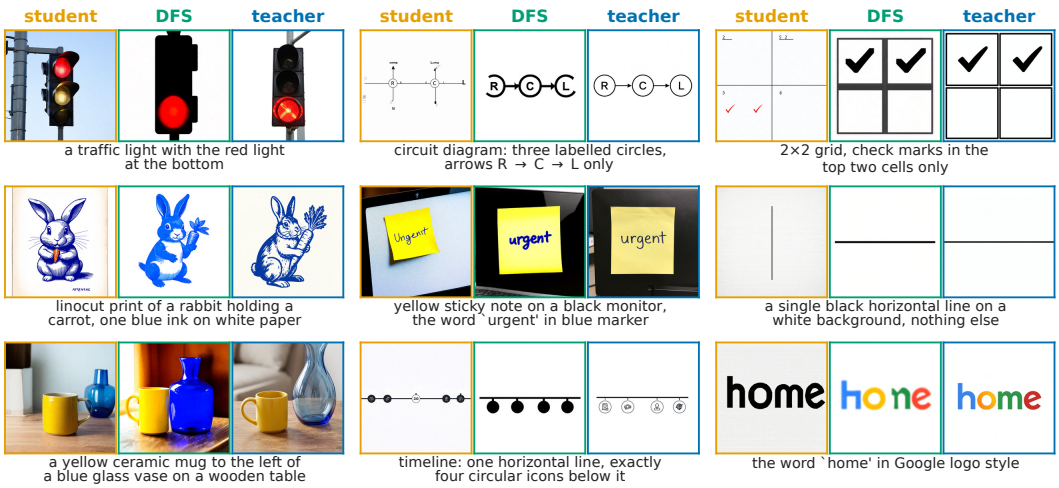


Figure 6: **SD3-medium student**, klein teacher, one seed per prompt at the strength reported in Table 14 ($\alpha = 15$). The unsteered image fails and the steered image passes the checklist in every triplet; examples were picked by hand among the prompts that DFS solves on this student.

random directions about twice as much (11.1 vs. 5.1). Direction variants that might compensate (constant removed, low-pass filtered, logistic-regression weights, other schedule exponents) change the SD3 pass rate by at most a few points, inside one confidence interval. The bottleneck is therefore the student's own trained flow, not the estimate and not the data path. SD3-medium gives $\Delta\boldsymbol{\mu}$ only half the advantage over an arbitrary direction that FLUX gives it, and the best strength doubles ($\alpha = 15$ against 7). One fixed strength is what falls short. With the best of five strengths per image, still with no teacher at inference, SD3-medium reaches the rate FLUX gets at a single fixed strength on the same prompts (pick column of Table 16). Selectivity is measured from a few samplings, so it can be checked for a new student before any benchmark run.

Table 16: **Shared prompts and transport on SD3-medium.** Upper: pass rates on the 103 prompts common to both students; pick = user picks the best of five strengths per image. Lower: transport gain relative to random directions for both students at five noise levels.

arm	teacher	pass %	95% CI	solved %	pick %
Prompt rewrite by MLLM	klein imgs	16.3	[10.7, 22.3]	10.7	
DFS $\Delta\boldsymbol{\mu}$, α=15	klein imgs	24.3	[18.0, 31.3]	15.5	43.0
FLUX.1-dev, same prompts: DFS $\Delta\boldsymbol{\mu}$, α=10	klein imgs	42.0	[34.7, 49.3]	33.0	

direction, student	σ=0.99	σ=0.95	σ=0.88	σ=0.75	σ=0.50
$\Delta\boldsymbol{\mu}$, SD3-medium	4.12	3.16	2.33	1.57	1.23
phase-scrambled $\Delta\boldsymbol{\mu}$, SD3-medium	2.69	2.17	1.76	1.37	1.16
absolute gain $\Delta\boldsymbol{\mu}$, SD3-medium	49.6	20.4	9.5	4.9	3.3
absolute gain random, SD3-medium	11.1	6.1	4.0	3.0	2.7
$\Delta\boldsymbol{\mu}$, FLUX.1-dev	8.80	5.50	4.00	2.89	1.52
phase-scrambled $\Delta\boldsymbol{\mu}$, FLUX.1-dev	3.45	3.01	3.12	2.38	1.64
absolute gain $\Delta\boldsymbol{\mu}$, FLUX.1-dev	55.4	21.6	11.7	6.0	2.4

D.2 A TEACHER FROM ANOTHER MODEL FAMILY

The main experiments pair FLUX.1-dev with a teacher from its own family. To test whether that matters, we keep the student, its cached latents, the evaluation seeds, the checklists and the judge, and replace the teacher by Qwen-Image-2512 (Apache-2.0; run with fp8 weight storage and bf16

compute), so that only the twenty images that define $\Delta\boldsymbol{\mu}$ change. Its outputs go through the same gate: Qwen-Image passes all four gate seeds on 70 of the 200 prompts (42 strict, 28 relaxed) and 58% of its gate images, weakest on projection, counting and negation. Its direction agrees with the klein direction for the same prompt at median cosine 0.56 (IQR 0.20–0.75), while two halves of Qwen's own images agree at 0.81, so the two teachers pull toward measurably different means.

Table 17: **Cross-family teacher, strict tier**: the 42 strict prompts that Qwen-Image-2512 passes on all four gate seeds, FLUX.1-dev student, $\alpha = 7$ pre-declared for both teachers. The klein rows are recomputed on the same prompts. The last three columns are the strength-selection rows of Table 9 over $\alpha \in \{2, 5, 7, 10, 20\}$ (pick at five samples per image).

arm	pass %	95% CI	solved %	leak-free	oracle α	pick
Student (unsteered)	0.6	[0.0, 1.8]	0.0	–	–	–
DFS, Qwen-Image-2512 teacher	42.9	[32.1, 54.8]	38.1	46.4	58.9	65.5
DFS, FLUX.2-klein teacher, same prompts	44.0	[32.7, 56.0]	33.3	50.0	60.7	68.5

On these prompts the two teachers are interchangeable (Table 17). Two caveats remain. The judge and this teacher share a model family, although the judge sees Qwen-Image's own images only in the gate and never when scoring the steered FLUX images. The benchmark was gated on klein, so its prompts are pre-selected for klein's reliability, not Qwen's.

D.3 SCHEDULE

Table 18: **Schedule ablation** on the 30-prompt subset (one confidence interval is about ± 13 points; the fresh-seed student passes 10.0% here). Active: steps that receive a push; matched budget: α chosen so that $\alpha \sum_i h$ equals that of the reported schedule at $\alpha = 7$, with a sign test against it over the 120 cells. Over all cells of the table, the pass rate has Spearman correlation 0.77 with the push delivered on the first four steps ($\sigma \geq 0.95$) and 0.12 with the total push.

schedule h	active	$\sum_i h$	pass % at α=7	best pass % (α)	matched budget
β=10 (reported)	23	5.57	39.2	45.0 (10)	–
first two steps only	2	1.68	29.2	43.3 (10)	35.8 (p=0.67)
β=50	10	1.15	27.5	42.5 (33.8)	42.5 (p=0.64)
β=5	26	9.03	38.3	40.0 (6.17)	33.3 (p=0.12)
β=2	27	14.55	42.5	42.5 (7)	27.5 (p=0.0066)
uniform over all steps	27	27.00	36.7	36.7 (7)	21.7 (p<0.001)
only $\sigma < 0.85$	17	17.00	11.7	11.7 (7)	7.5 (p<0.001)

A literal reading of Proposition 1(iii) would not front-load: the displacement of a step with the target field is proportional to $|\Delta\sigma_i|/\sigma_i$, which grows along the shifted FLUX schedule. Front-loading rests instead on the shrinking validity of a fixed vector (Table 3), on transport (Section 3.3) and on this ablation: the effect is carried by the first few steps, any front-loaded schedule performs like $\beta = 10$, and a push applied only below $\sigma = 0.85$ stays at the floor at every strength (at most 11.7%).

D.4 QUALITATIVE RESULTS

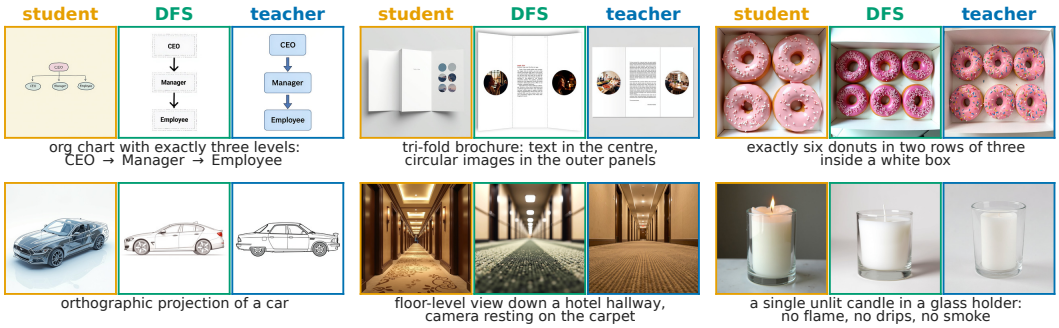


Figure 7: **One seed with and without steering** at the fixed $\alpha = 7$, with one of the teacher images the direction was computed from. Prompts were picked by hand to cover families, and the judge passes every steered image shown.

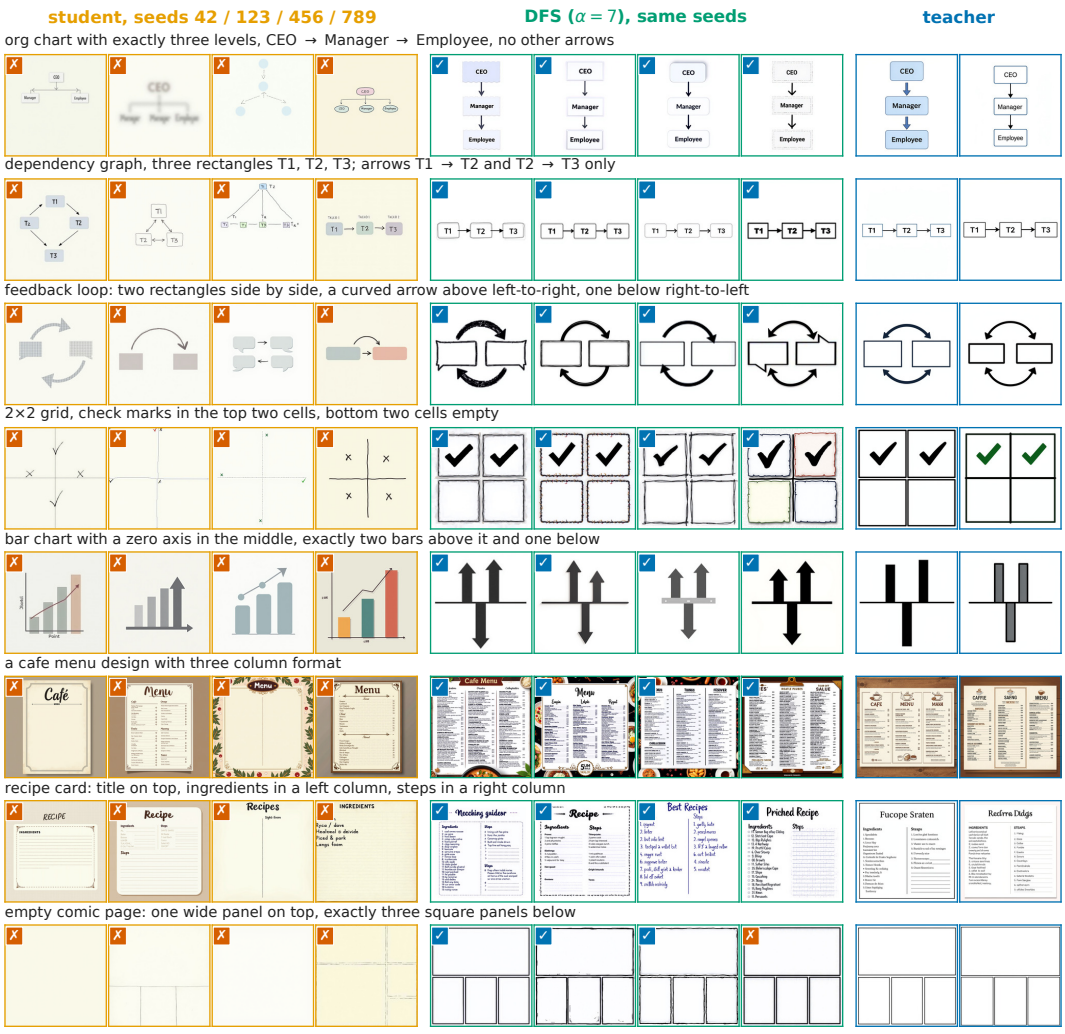


Figure 8: **All four evaluation seeds, diagram structure and document layout.** Left: the unsteered student; middle: the same seeds steered at the fixed $\alpha = 7$; right: two of the twenty teacher images. The tick or cross on each image is the judge's verdict (every checklist item satisfied). No seed is omitted.



Figure 9: **All four evaluation seeds, counting, attribute binding, projection and text** (layout as in Figure 8).

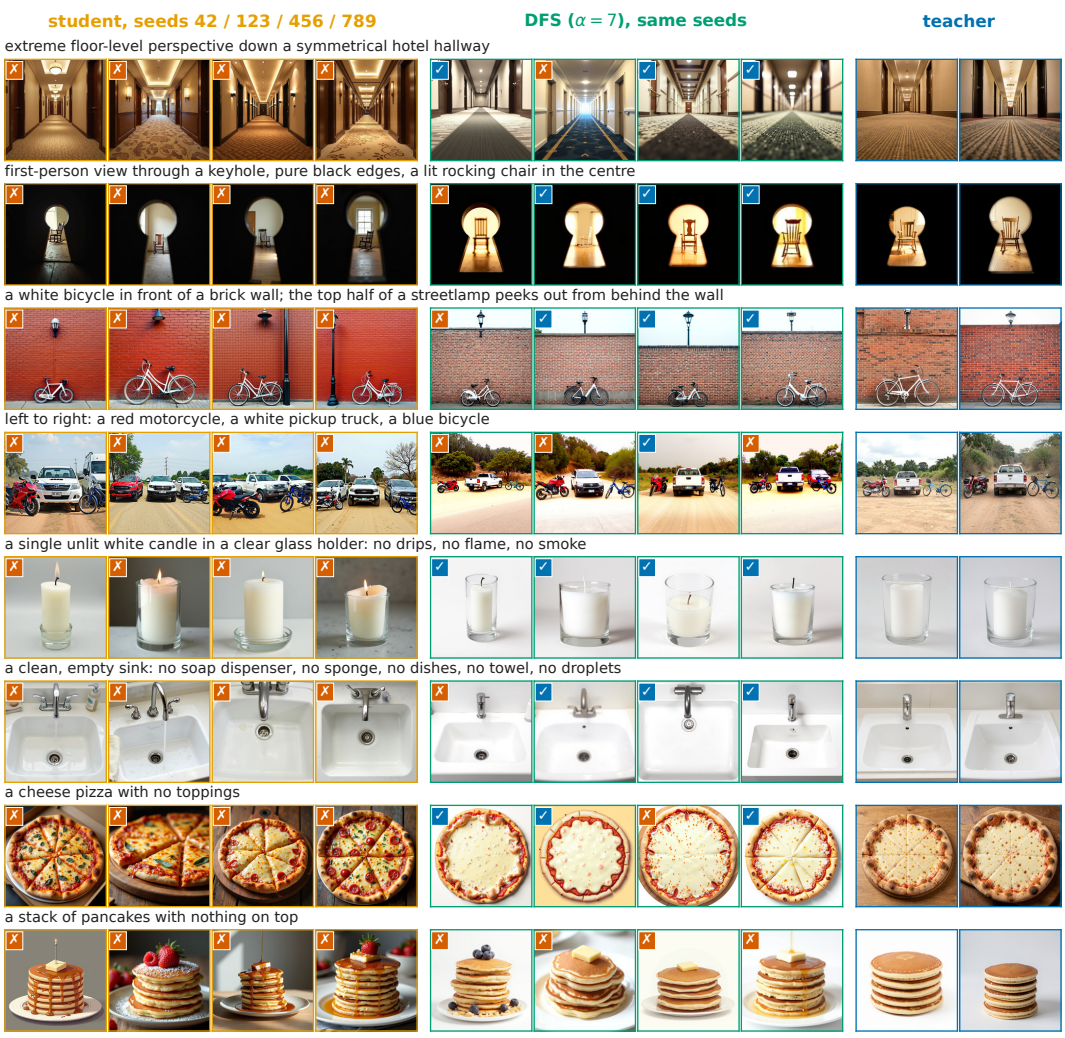


Figure 10: **All four evaluation seeds, viewpoint, spatial relation and exclusion** (layout as in Figure 8). The last row is an exclusion prompt on which DFS fails on every seed; NAG and LoRA pass it (Figure 3).

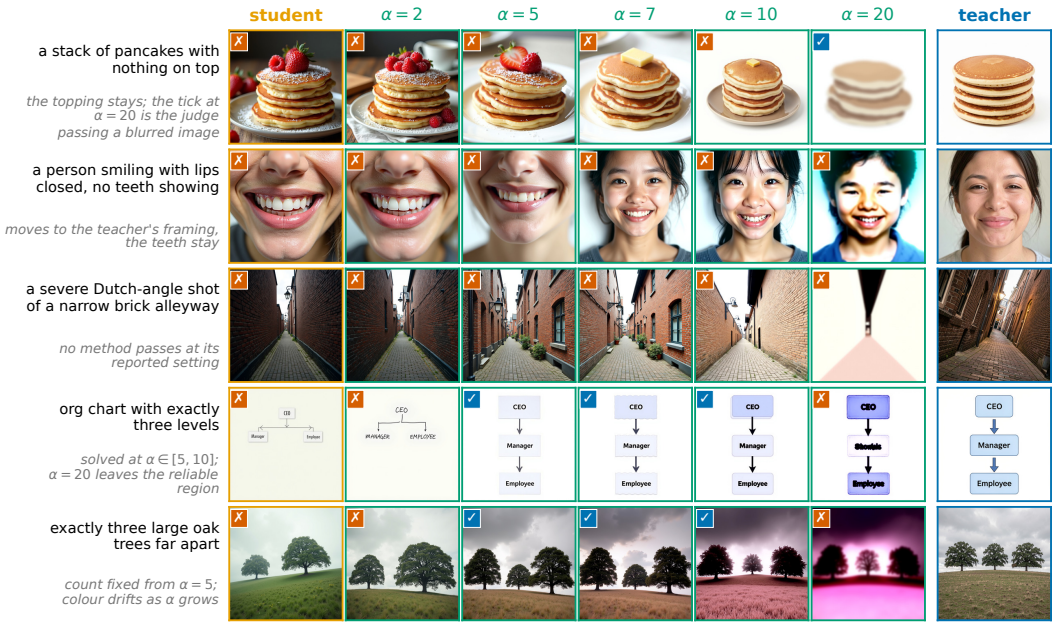


Figure 11: **Where DFS fails, and what too large a strength does** (one seed per row, increasing α; marks are the judge's verdict). Rows 1–2: gaps that require removing something the student adds; the mean difference moves the image toward the teacher set's framing and the unwanted content stays. The tick at $\alpha = 20$ in row 1 is the judge accepting a blurred image; headline results use the fixed $\alpha = 7$. Row 3: a viewpoint prompt no method passes at its reported setting. Rows 4–5: prompts that are solved from $\alpha = 5$; at $\alpha = 20$ the latent has left the region where the sampler is reliable (Figure 12), and colour drifts before structure breaks.

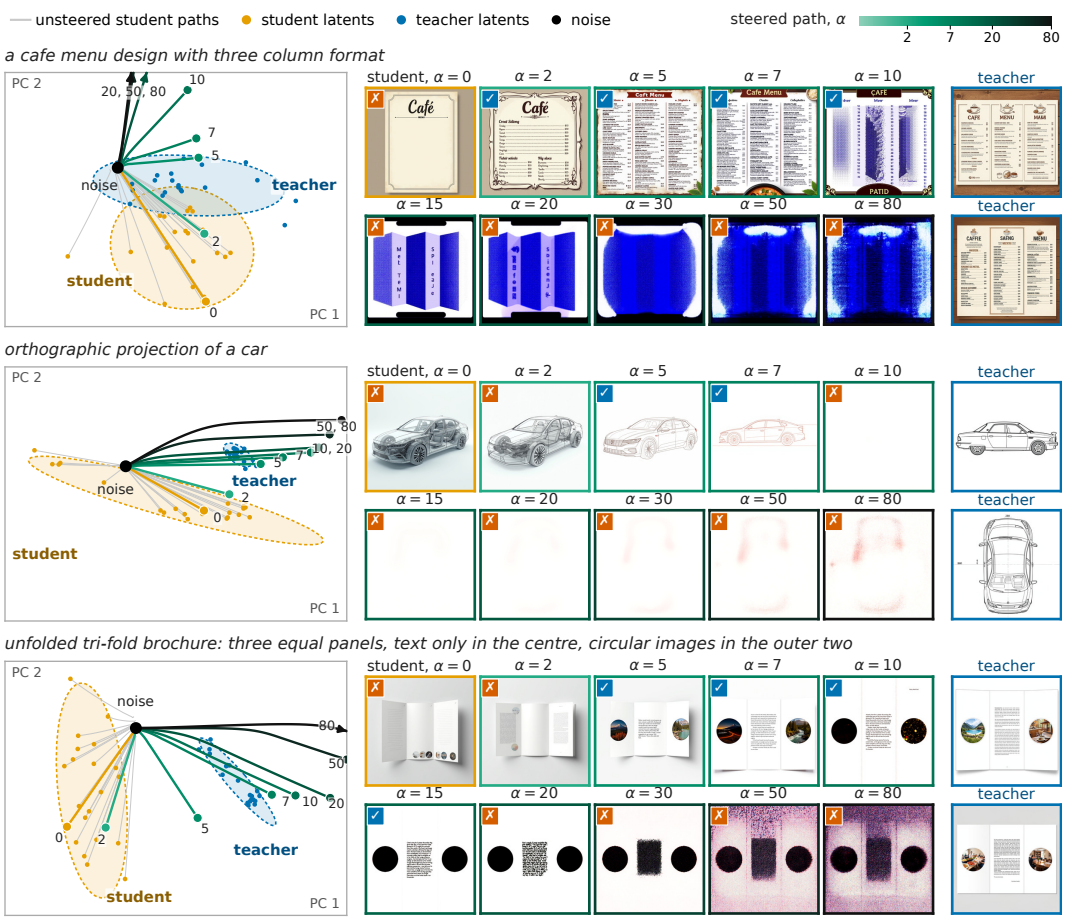


Figure 12: **Sampling paths under increasing strength** (one seed per row). Left: every latent along the sampling path, projected on the top two principal components of the 40 clean latents the direction is computed from. The components are computed without class labels, and both axes share one scale. Grey lines are the 20 student samples of that set, traced from noise to their final latents (orange dots). Blue dots are the 20 teacher images, encoded by the student's VAE. Ellipses cover two standard deviations. The orange line is the shown seed without steering. The green to black lines start from the same noise and are steered with increasing α. Numbers give α at each end point, and arrows mark paths that leave the frame. Right: the decoded images with the judge's verdict, and two teacher images. The other three seeds look alike. The plane holds 26–40% of an unsteered latent's deviation from the mean, so distances in it are only indicative.

D.5 Sampling paths under increasing strength

Figure 12 traces every latent along the FLUX.1-dev sampler for 6 strict-tier prompts at increasing strength, on prompts chosen so that $\alpha = 7$ passes and $\alpha = 20$ fails on all four seeds. Without steering, the path ends among the student's latents. At the working strengths it ends among the teacher's encodings, and not only in the projected plane: in the full latent space the nearest of the 40 fitting latents is a teacher latent on all 24 of 24 paths at every $\alpha \geq 5$. Pushed further, the path leaves the teacher's region and moves away from all real latents, and the image degrades into artifacts. At $\alpha \geq 30$ the judge accepts 1 of 72 images, a faint outline on a blank page.

E VIDEO

Applied unchanged to Wan2.1-T2V-1.3B with Wan2.2-T2V-A14B as teacher on 15 motion prompts, $\Delta\boldsymbol{\mu}$ moves the student's motion toward the teacher's on all 11 of the 11 prompts where the two flows differ by at least 0.3 pixels per frame, and reverses with the sign of α.

Student Wan2.1-T2V-1.3B, teacher Wan2.2-T2V-A14B (Wan et al., 2025) (same VAE family), 81 frames at 480×832, $N = 20$ videos per model per prompt, $\Delta\boldsymbol{\mu}$ over the clean video latent ($16 \times 21 \times 60 \times 104$), three seeds per steered arm, the same steerer as the image runs with $\beta = 10$ and $\alpha \in \{5, 10, 20\}$. Motion is summarised by the mean Farnebäck optical-flow magnitude over the clip (pixels per frame); the teacher moves more than the student on 11 of the 15 prompts and less on the rest. The direction is noisier than on images, as expected for twenty samples in a latent forty times larger (split-half cosine 0.59, held-out Fisher 2.9).

All prompts. Table 19 lists the 15 prompts with the measured flows and the motion GapClosed at each strength. The four prompts without a checkmark have no motion gap to close, and their ratios are noise. Table 20 gives the reference-free dimensions of VBench (Huang et al., 2024). Dynamic degree rises from 0.44 to 0.67 against a teacher value of 0.64. Imaging quality moves part of the way toward the teacher. Motion smoothness and the consistency dimensions stay within 0.03 of the student, with a small subject-consistency dip at $\alpha = 20$.

Table 19: **Video, per prompt.** The flow columns give the mean optical-flow magnitude of teacher and student in pixels per frame. The GapClosed columns give (steered − student)/(teacher − student): 1 lands on the teacher, 0 is no change, above 1 overshoots, below 0 moves away. Checkmark: the teacher–student gap is at least 0.3 pixels per frame.

	flow (px/frame)		GapClosed			
prompt	teacher	student	α=5	α=10	α=20	gap
a dog running happily	6.3	3.8	1.10	2.21	-0.04	✓
a horse galloping across an open field	8.1	6.3	0.66	0.08	-2.31	✓
a motorcycle turning a corner	7.6	7.0	0.91	-1.00	-4.14	✓
a person picks up a cup from the table t…	2.0	0.6	0.32	0.57	1.48	✓
a sailboat sailing to the left while a b…	1.5	0.5	0.71	1.30	1.11	✓
a tranquil tableau of an ornate victoria…	0.2	0.1	-0.36	-0.27	1.24	
a white flower gradually turns pink as i…	0.5	0.1	0.25	1.28	2.22	✓
aerial flyover of a coastal city at suns…	0.4	0.5	0.01	-1.58	-0.77	
an inflated basketball dribbling on the …	3.6	0.9	1.18	1.94	1.09	✓
cinematic dolly shot moving through a mi…	2.2	0.6	0.56	1.63	2.22	✓
cream swirling dreamily into coffee	1.2	0.2	0.06	0.90	3.57	✓
in a still frame within the desolate des…	0.2	0.2	5.74	8.26	12.12	
slow tracking shot following a person wa…	7.8	11.7	1.04	1.80	2.51	✓
timelapse of an ice cube melting on a so…	2.0	0.8	0.22	0.41	0.99	✓
wine pouring from a bottle into a glass	0.5	0.6	-0.51	-1.31	1.82	

Table 20: **Video, VBench reference-free dimensions** (15 prompts, 45 videos per arm, 300 for the teacher), computed in custom-input mode on our prompts and therefore not comparable to the VBench leaderboard.

arm	dyn. degree	img. quality	motion smooth.	subject cons.	backgr. cons.
Student	0.444	0.587	0.983	0.941	0.953
Teacher	0.643	0.681	0.981	0.938	0.962
DFS $\Delta\mu$, α=5	0.533	0.600	0.983	0.933	0.948
DFS $\Delta\mu$, α=10	0.667	0.608	0.981	0.923	0.943
DFS $\Delta\mu$, α=20	0.667	0.629	0.982	0.913	0.944

F COMPUTE

Direction estimation costs N student samplings and N VAE encodes per prompt (about five minutes per prompt on one A100-40GB for $N = 20$). Steering adds one vector addition and one scalar normalisation per step (under 0.1% of a step). The benchmark run generated 9424 steered images for the four estimator arms, 3024 for the training-free baselines, plus 13752 gate images, and issued 108000 judge calls, on six A100-40GB GPUs over about two days; the NAG, LoRA and Concept Sliders arms took a further day, most of it LoRA training. The transport measurement is 1260 samplings.