ANALYZING AND STABILIZING CLASSIFIER SCORE DISTILLATION FOR TEXT-TO-IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution Matching Distillation (DMD) accelerates diffusion and flow-matching models by compressing their iterative sampling process into a few steps. Recent work shows that its gradient can be decomposed into a DM term and a CFG-augmentation (CA) term, with CA providing a particularly strong distillation signal in text-to-image generation. Yet why CA is effective, and what constraints are needed to stabilize it, remain underexplored. To analyze its effectiveness, we first conduct synthetic experiments illustrating that finer condition groupings yield more discriminative classifier-score directions. We then probe the corresponding score-space mechanism in text-to-image models through interpolation, where retaining more fine-condition score information prolongs the effective distillation phase. These findings further support repositioning classifier score as the central distillation signal, motivating a reconsideration of stabilization within the classifier score distillation (CSD) framework. Within this framing, we ask whether sustaining CA necessarily requires online estimation of the evolving student distribution. We compare distribution matching, including fake and student side tracking strategies, with transition anchoring, which instead constrains student transitions using frozen teacher dynamics without maintaining an online fake-score model. We further equip transition-anchoring variants with a final-step perceptual loss to alleviate CA-induced oversaturation and local-detail degradation. Our empirical results show that both families improve training stability. On FLUX.2, transition anchoring variants achieve competitive performance with a 2.7-4.0x training speedup over vanilla DM, while revealing trade-offs between preference-oriented quality and alignment with the teacher distribution.

1 INTRODUCTION

Diffusion (Ho et al., 2020; Song et al., 2021) and flow-matching models (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025) achieve strong image-generation performance, but their iterative sampling requires repeated network evaluations and incurs substantial inference cost. Few-step distillation (Yin et al., 2024a;b; Song et al., 2023; Sauer et al., 2024b) seeks to compress this process into a few sampling steps. Among existing approaches, Distribution Matching Distillation (DMD) (Luo et al., 2023; Yin et al., 2024a;b) has become a widely used framework and has been scaled to modern text-conditioned generative models. Its practical training, however, remains sensitive to hyperparameter choices due to the coupling of student-fake optimization (Wu et al., 2026b).

Recently, Decoupled DMD (Liu et al., 2026b) shows that the DMD gradient can be decomposed into a distribution-matching (DM) term and a CFG augmentation (CA) term. In text-to-image generation, CA alone provides a strong distillation signal, yet eventually collapses without additional regularization. These observations raise two natural questions: why is classifier-score distillation particularly effective in text-to-image generation, and can it be stabilized using lightweight regularization without maintaining or frequently updating an online fake-score model?

In this work, we investigate these questions from mechanistic and optimization perspectives. For the former, we study how condition granularity (Yuan et al., 2026) affects the informativeness of classifier-score directions (Figure 1 (a)). Intuitively, finer conditions narrow the set of compatible samples, making the classifier-score signal more discriminative among individual modes. We first test this hypothesis in a controlled Gaussian-mixture setting that isolates condition granularity while

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

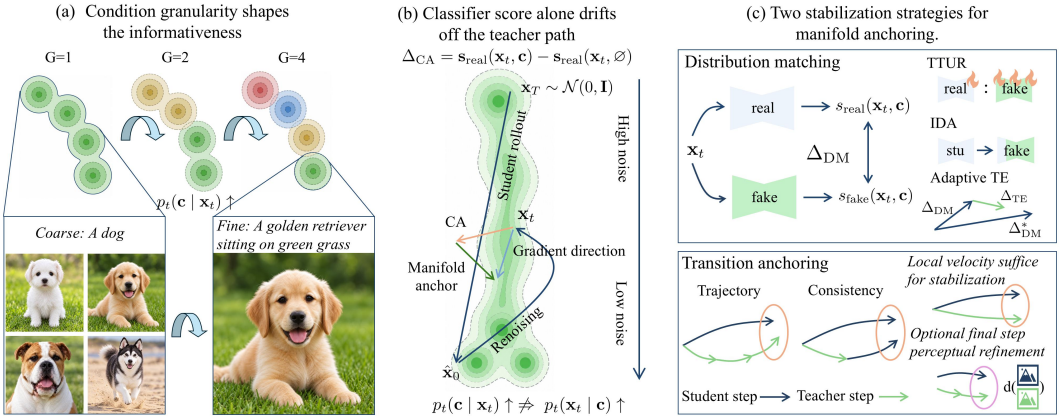


Figure 1: For analysis, we study how condition granularity affects the informativeness of classifier-score directions and probe the mechanism in text-to-image settings through score-space interpolation. For stabilization, we compare distribution matching with transition anchoring.

keeping the underlying data distribution fixed. We find that finer condition groupings produce more mode-discriminative classifier-score directions and improve the effectiveness of CA. We then probe the corresponding score-space effect in practical text-to-image models by interpolating the fine-condition component of the classifier score. Despite ultimate collapse, retaining more fine-condition score information consistently prolongs the effective optimization phase. Together, these findings further support the central role of classifier score as a distillation signal and motivate reconsidering how it should be stabilized within the classifier score distillation (CSD) framework.

Within this framing, we ask whether sustaining CA necessarily requires online estimation of the evolving student distribution (Figure 1 (c)). We first revisit distribution matching, whose practical update relies on an online approximation of the fake score, and examine tracking strategies from both fake and student-side. Despite progress, stabilization still depends on an online fake-score model that only approximates the evolving student score. Moving beyond the DM framework, we investigate transition anchoring, which constrains student transitions with frozen teacher dynamics. Transition-anchoring variants achieve competitive stabilization and provide a 2.7–4.0$\times$ training speedup over vanilla DM on FLUX.2 series. We further equip these variants with a final-step perceptual loss to alleviate CA-induced oversaturation and local-detail degradation.

We systematically evaluate these objectives across SD3.5-M,L (Esser et al., 2024) and FLUX.2-klein-base-4B,9B (Labs, 2025). Empirically, both stabilizer families improve training stability while exhibiting trade-offs between preference-oriented quality and alignment with teacher distribution.

Our contributions are summarized as follows:

- We analyze how condition granularity affects CSD using controlled Gaussian mixtures and score interpolation in text-to-image models, showing that fine-condition score information improves distillation but does not prevent collapse.
- Motivated by the central role of classifier score signal, we compare teacher-based transition anchoring with distribution matching variants for stabilization, and study final-step perceptual supervision to reduce oversaturation and preserve local details.
- We validate the stabilization framework across four text-conditioned models, demonstrating the competitive stabilization of transition anchoring variants to traditional DM.

2 Related Work

Trajectory distillation. Early efforts to compress the multi-step sampling process of diffusion models mainly focus on trajectory-level distillation (Salimans & Ho, 2022; Ren et al., 2024; Song et al., 2023). Progressive distillation (Salimans & Ho, 2022) performs this compression gradually by training a student sampler to reproduce multiple teacher steps and reducing the number of sampling

steps by a factor of two at each stage. Consistency distillation (Song et al., 2023) instead learns a consistency function whose predictions from different points along the probability-flow ODE trajectory correspond to the same endpoint, enabling one or few-step generation. Recent progress in flow-based generative modeling further extends this trajectory view, either by modeling average velocities over time intervals (Geng et al., 2025) or by learning flow maps between pairs of time points along the transport path (Boffi et al., 2024).

Distribution alignment distillation. Another line of work focuses on aligning the student distribution with either the teacher or the underlying data distribution (Sauer et al., 2024b; Yin et al., 2024a; Wang et al., 2023; Sauer et al., 2024a; Luo et al., 2023; Huang et al., 2026b). Score distillation was first explored in 3D generation (Wang et al., 2023), where the idea is to provide matching at the distribution level. Diff-Instruct (Luo et al., 2023) extends this idea to image generation. DMD and DMD2 (Yin et al., 2024a;b) further improve this formulation and demonstrates its effectiveness at scale. Adversarial approaches such as ADD (Sauer et al., 2024b) and LADD (Sauer et al., 2024a) introduce an additional discriminator to better align the student with the reference distribution, defined by the teacher or the data. Other distribution-matching variants have also been proposed to improve training stability (Huang et al., 2024; Zhou et al., 2024), and overall performance (Liu et al., 2026b; Wu et al., 2026a; Ge et al., 2026; Fan et al., 2026; Li et al., 2026; Luo et al., 2025; Chadebec et al., 2025; Liu et al., 2026a). More recently, hybrid methods have started to combine trajectory-based and distribution-based distillation (Zheng et al., 2026c).

Classifier-Free Guidance. Initial approaches to conditional diffusion generation employed an auxiliary classifier to an unconditional diffusion model (Song et al., 2021). By Bayes' rule, the conditional score can be decomposed as $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{c}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$, where $\nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t)$ serves as the classifier score. With an optimal conditional score model and exact numerical integration, sampling recovers the target conditional distribution $p(\mathbf{x}_0 \mid \mathbf{c})$ after the multi-step sampling process (Gu et al., 2025). However, due to training errors (Gu et al., 2025) and the mismatch between the MSE objective and final image in the perceptual space (Lin et al., 2026; Gladstone et al., 2026), unguided conditional generation can lead to perceptually unsatisfactory results. Introducing a scaling factor $w > 1$ in front of the classifier score gradient improves sample quality (Dhariwal & Nichol, 2021), $w \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t)$, which also gives room for the Bayes-rule decomposition behind Classifier-Free Guidance (CFG) (Ho & Salimans, 2021). Specifically, CFG replaces the explicit classifier score by the difference between conditional and unconditional diffusion scores: $\nabla_{\mathbf{x}_t} \log p_w(\mathbf{x}_t \mid \mathbf{c}) = w \cdot (\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{c}) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. The role of guidance becomes particularly pronounced in large text-conditioned (Esser et al., 2024; Labs, 2025) models, where unguided conditional sampling often underperforms unless additional post-training (Zheng et al., 2026b) is applied.

3 PRELIMINARIES

For diffusion models (Song et al., 2021; Ho et al., 2020; Sohl-Dickstein et al., 2015), given a clean sample $\mathbf{x}_0$ and condition $\mathbf{c}$, the forward perturbation process defines a Gaussian probability path

$$p_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t; \alpha_t \mathbf{x}_0,\, \sigma_t^2 \mathbf{I}\big)\,, \qquad \mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3.1}$$

and the diffusion model learns a score function $\mathbf{s}(\mathbf{x}_t, t, \mathbf{c})$ that approximates the conditional score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{c})$. In practice, it is trained with a denoising score matching objective $\mathbb{E}_{t,\mathbf{x}_0,\mathbf{c},\boldsymbol{\epsilon}}\Big[w(t)\,\|\mathbf{s}(\mathbf{x}_t, t, \mathbf{c}) + \boldsymbol{\epsilon}/\sigma_t\|_2^2\Big]$. Alternative parameterizations, such as data or velocity prediction, can be converted to the corresponding score representation. (Gao et al., 2024; Kingma & Gao, 2023). For example, given the flow matching OT (Liu et al., 2023) transport path $\alpha_t = 1 - t, \sigma_t = t$, the equivalence between the score prediction and the velocity prediction is

$$\mathbf{v}(\mathbf{x}_t, t, \mathbf{c}) = -\frac{\mathbf{x}_t + t \cdot \mathbf{s}(\mathbf{x}_t, t, \mathbf{c})}{1 - t}. \qquad t \in [0, 1) \tag{3.2}$$

In Distribution Matching Distillation (DMD) (Yin et al., 2024b;a), the generator G_θ is updated through the score discrepancy between the pretrained real score model and the auxiliary fake score model, derived from a reverse KL divergence objective along the probability path:

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}\left[-\big(\mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, \varnothing) + \alpha\big(\mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, \mathbf{c}) - \mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, \varnothing)\big) - \mathbf{s}_{\mathrm{fake}}(\mathbf{x}_t, \mathbf{c})\big)\,\frac{\partial G_\theta(\mathbf{z})}{\partial \theta}\right]. \tag{3.3}$$

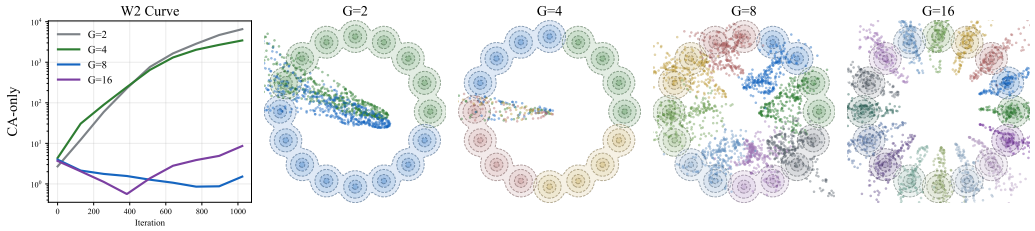


Figure 2: **Experiments on synthetic Gaussian mixtures**. CA-only distillation with varying condition grouping $G \in \{2, 4, 8, 16\}$. Full teacher and CA+DM comparison is provided in Figure 8.

With the rearrangement in decoupled-DMD (Liu et al., 2026b), the objective can be decomposed into a distribution-matching term and a CFG augmentation term:

$$\mathbb{E}\left[-\left(\underbrace{\left(\mathbf{s}_{\text{real}}(\mathbf{x}_t, \mathbf{c}) - \mathbf{s}_{\text{fake}}(\mathbf{x}_t, \mathbf{c})\right)}_{\text{DM}} + (\alpha - 1)\underbrace{\left(\mathbf{s}_{\text{real}}(\mathbf{x}_t, \mathbf{c}) - \mathbf{s}_{\text{real}}(\mathbf{x}_t, \varnothing)\right)}_{\text{CA}}\right)\frac{\partial G_\theta(\mathbf{z})}{\partial \theta}\right]. \quad (3.4)$$

Here, α denotes the original CFG (Ho & Salimans, 2021) scale, which is usually set to 8.0 for SDXL (Yin et al., 2024a; Rombach et al., 2022), and 4.0 for SD3/3.5-medium (Ge et al., 2026; Esser et al., 2024). Because the CA term enjoys the same classifier-score form as Classifier Score Distillation (CSD) (Yu et al., 2024), here we use CSD as the naming for the general distillation framework built upon CA throughout the paper.

4 Analyzing Classifier Score Distillation

4.1 On the Effectiveness of CA

Recent work (Liu et al., 2026b) suggests that the classifier-score component (CA) of DMD plays an increasingly prominent role when scaling to text-conditioned generation. In particular, the CA term introduces a condition-aligned driving force related to $-\nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t)$. While Liu et al. (2026b) explains its effectiveness through an analogy to CFG from the perspective of large language models, we investigate a complementary factor: *condition granularity* (Yuan et al., 2026). We hypothesize that finer conditions induce more specific classifier-score directions, whose increased informativeness strengthens CA.

Toy example. We begin with a 16-mode ring Gaussian mixture to illustrate the proposed intuition. We keep the marginal data distribution fixed while varying how the same modes are grouped into conditions, with $G \in \{2, 4, 8, 16\}$. For each G, we train a conditional diffusion teacher and use its conditional–unconditional score gap for CA-only distillation. Owing to the simplicity of the task, the multi-step teachers reliably recover the target distribution across all condition groupings. Full settings and additional comparisons are provided in Section A.1.

As shown in Figure 2, CA-only distillation is highly sensitive to the condition grouping. Under coarse conditions, the conditional–unconditional score gap provides broad group-level guidance and fails to preserve individual modes. As G increases, each condition identifies a narrower subset of outcomes, yielding more mode-specific CA directions and better recovery of the target structure. Such result offers intuition for why CA may be particularly effective in text-to-image generation. Class labels in datasets such as CIFAR-10 (Krizhevsky, 2009) and ImageNet-1K (Russakovsky et al., 2015) group many visually diverse samples under a shared condition, whereas text prompts can impose substantially finer semantic constraints and thereby narrow the set of compatible modes.

Score interpolation in text-to-image setting. Directly isolating condition granularity is challenging in T2I, as the effective conditioning signal is entangled with prompt semantics, text-encoder behavior, and model-specific conditioning mechanisms (Xiaofeng et al., 2025; Chen et al., 2026a). Building on the toy observation that finer conditions induce more discriminative classifier-score directions, we next directly probe this score-space mechanism in practical T2I models. Specifically,

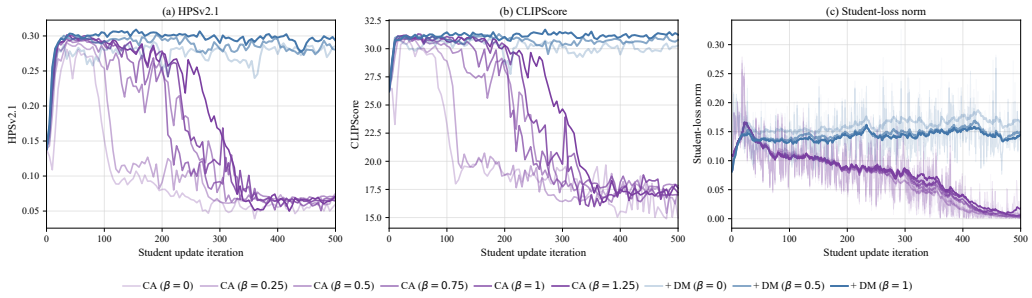


Figure 3: Score interpolation experiments on FLUX.2 Klein 4B for CA-only and CA+DM (TTUR=5). As β increases, retaining more fine-grained score information prolongs the effective distillation phase, although CA-only training still collapses without stabilization.

for each prompt P in a JourneyDB subset (Sun et al., 2023), we use Qwen3.8-27B (Qwen Team, 2026) to construct a coarsened version P_c whose semantic information is contained within P, *i.e.*, $I(P_c) \subseteq I(P)$. We then interpolate between the coarse and fine conditioned teacher scores, and test extrapolation beyond the fine-conditioned score, while keeping the underlying student generation task fixed. Specifically,

$$\mathbf{g}_{\mathrm{CA}}^{(\beta)} = -\Big(\underbrace{\mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, t, P_c) - \mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, t, \varnothing)}_{\text{coarse-condition component}} + \beta \underbrace{\big(\mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, t, P) - \mathbf{s}_{\mathrm{real}}(\mathbf{x}_t, t, P_c)\big)}_{\text{fine-condition component}} \Big), \tag{4.1}$$

We also match the magnitude of the interpolated $\mathbf{g}_{\mathrm{CA}}^{(\beta)}$ to the original signal $\mathbf{g}_{\mathrm{CA}}^{(1)}$ to exclude the effect of signal strength. Thus, increasing β increases the contribution of the fine-condition score component. As shown in Figure 3, CA-only training is sensitive to this score-space intervention: increasing β generally prolongs the effective optimization phase, while all CA-only variants eventually deteriorate. When combined with DM, the distillation process is stabilized, and the influence of score interpolation persists.

Does the benefit persist with increasing prompt specificity? The score-space intervention above shows that retaining additional fine-condition score information benefits CA. A natural follow-up question is whether this benefit can be realized simply by increasing prompt specificity in natural language. Through empirical investigation, we find the answer to be negative. The benefit eventually saturates or reverses. Moreover, textual specificity alone does not provide a direct measure of condition granularity, as additional content may be redundant, stylistic, or insufficiently represented by the teacher (Chen et al., 2026b). We discuss both aspects in Section A.2.

4.2 On the limitations of CA

Without proper stabilization, all CA-only runs eventually collapse (Figures 2 and 3). Such instability is rooted in the CA objective itself, which promotes states that are increasingly discriminative of the condition. A high conditional posterior, however, does not necessarily imply high likelihood under the teacher distribution (Figure 1 (b)). Such mismatch can manifest as over-guidance artifacts in high-dimensional image spaces, including oversaturation and simplified local structure (Sadat et al., 2025). Consequently, optimizing CA alone can increase the discriminability $p(\mathbf{c} \mid \mathbf{x}_t)$ without necessarily preserving alignment with the teacher conditional probability path $p_t(\mathbf{x}_t \mid \mathbf{c})$.

5 Stabilizing classifier score distillation

Despite being unstable, the prominent role of CA term (Liu et al., 2026b) in text to image setting motivates reconsidering how the stabilization burden is allocated within the distillation system. We keep the CA objective, and investigate two families of stabilizing constraints. We first re-examine existing fake-side tracking strategies alongside a student-side regularizer, which we find to improve efficiency with sample-level adaptation under the reverse-KL formulation. As a step further beyond the DMD framework, we investigate frozen teacher dynamics as a lightweight stabilizing anchor,

by studying a suite of transition-anchoring objectives based on trajectory, velocity, and consistency matching, and a perceptual supervision strategy in the final step.

5.1 DISTRIBUTION MATCHING

A central challenge in DM is keeping the online fake-score model aligned with the evolving student distribution. Let $\mathbf{s}_p$ denote the teacher score and $\mathbf{s}_q$ the exact score of the noisy student distribution. Because $\mathbf{s}_q$ is generally intractable, DM approximates it with an online fake-score model $\mathbf{s}_\psi$. Up to a time-dependent factor, the DM gradient decomposes as

$$\nabla_\theta \mathcal{L}_{\mathrm{DM}} = -\mathbb{E}\left[(\mathbf{s}_p - \mathbf{s}_\psi)\frac{\partial G_\theta(\mathbf{z})}{\partial \theta}\right] = \underbrace{-\mathbb{E}\left[(\mathbf{s}_p - \mathbf{s}_q)\frac{\partial G_\theta(\mathbf{z})}{\partial \theta}\right]}_{\text{ideal reverse-KL gradient}} + \underbrace{\mathbb{E}\left[(\mathbf{s}_\psi - \mathbf{s}_q)\frac{\partial G_\theta(\mathbf{z})}{\partial \theta}\right]}_{\text{fake-score tracking discrepancy}}. \quad (5.1)$$

This decomposition reveals two complementary interventions: improving fake-score estimation or seeking to compensate for its error in the student update. On the fake-model side, prevalent TTUR (Yin et al., 2024a) performs multiple fake-model updates per student update, allocating additional computation to tracking the evolving student distribution. IDA (Ge et al., 2026) couples the fake-model parameters to the student through an EMA update, $\psi \leftarrow (1-\mu)\psi + \mu\theta$.

On the student side, drawing on score identities (Zhou et al., 2024), we incorporate the negative-residual in SGMD's Fisher-divergence formulation (Wu et al., 2026b) as a regularizer. In an initial fixed-weight ablation, direct incorporation fails to stabilize at $\mathrm{TTUR} = 1$. We then analyze the NR gradient under Equation (5.1) formulation, which reveals the mismatch between the global coefficient against state-dependent behavior, thus leading to a natural sample-level adaptation.

Specifically, let $\widehat{\boldsymbol{\delta}} = \mathrm{sg}[\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi]$ and $\widehat{\mathbf{u}} = \widehat{\boldsymbol{\delta}}(I + t^2 J_\psi)$, where $J_\psi = \frac{\partial \mathbf{s}_\psi}{\partial \mathbf{x}_t}$ and $\widetilde{\mathbf{s}}_{t|0}$ is the sampled perturbation score. Here, $\widehat{\boldsymbol{\delta}}$ is a sampled proxy for the compensating direction $\mathbf{s}_q - \mathbf{s}_\psi$. For each sample, we fit $\frac{2\lambda}{t^2}\widehat{\mathbf{u}}$ to $\widehat{\boldsymbol{\delta}}$ using a regularized least-squares criterion. We refer to the resulting sample-wise calibration as *Adaptive Tracking Error (TE)*:

$$\widehat{\lambda}_{\mathrm{TE}} = \mathrm{sg}\left[\frac{t^2}{2}\frac{\max\left(\langle \widehat{\mathbf{u}}, \widehat{\boldsymbol{\delta}}\rangle, 0\right)}{\|\widehat{\mathbf{u}}\|_2^2 + \varepsilon}\right], \qquad \widehat{\lambda}_{\mathrm{TE}}\ell_{\mathrm{TE}} = -\widehat{\lambda}_{\mathrm{TE}}\left\|\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi\right\|_2^2. \quad (5.2)$$

Details are provided in Section B.1. Although it does not guarantee full recovery of the ideal marginal DM gradient, we observe sustained distillation at $\mathrm{TTUR} = 1$. Together, these improvements on both sides nevertheless retain an online estimate of the evolving student score. For a fixed class-conditional dataset, the score of its perturbed empirical distribution can be evaluated explicitly from its samples (Gu et al., 2025). Here, however, DM requires the score of the current student distribution under various text conditions, which a fixed reference dataset does not directly provide.

5.2 TRANSITION ANCHORING

The effectiveness of CA motivates asking whether its stabilization requires an online estimate of the student distribution. We investigate an alternative that directly anchors student transitions to the frozen dynamics of the pretrained teacher. Rather than estimating a teacher–student score discrepancy, these objectives penalize deviations from teacher-induced transitions at states visited by the student. We refer to this family as *transition anchoring* and study finite-interval trajectory matching, its single-step velocity form, and a compositional consistency variant.

Building on the interval formulation of TDM (Luo et al., 2025), we replace online distribution matching within each interval with deterministic teacher ODEs. Let $G_\theta^{t\to s}$ denote the student transition from the discretized steps $t, s \in \{t_i\}^S$ and $\Phi_{\phi,K}^{t\to s}$ a frozen teacher trajectory computed using K ODE steps, where $t > s$. Given a state $\mathbf{x}_t$ sampled from the student trajectory, we define

$$\mathcal{L}_{\mathrm{ODE}}^{(K)} = \mathbb{E}_{t,\mathbf{x}_t}\left[\left\|G_\theta^{t\to s}(\mathbf{x}_t) - \mathrm{sg}\left[\Phi_{\phi,K}^{t\to s}(\mathbf{x}_t)\right]\right\|_2^2\right]. \quad (5.3)$$

To examine how much teacher integration is necessary for stabilization, we reduce the teacher solver to a single Euler step. Let $h = t - s > 0$ be the step size, which gives the single-step velocity variant

$$\mathcal{L}_{\mathrm{ODE}}^{(1)} = \mathbb{E}_{t,\mathbf{x}_t}\left[h^2\left\|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}_\phi(\mathbf{x}_t, t)\right\|_2^2\right] \propto \mathcal{L}_{\mathrm{vel}}. \quad (5.4)$$

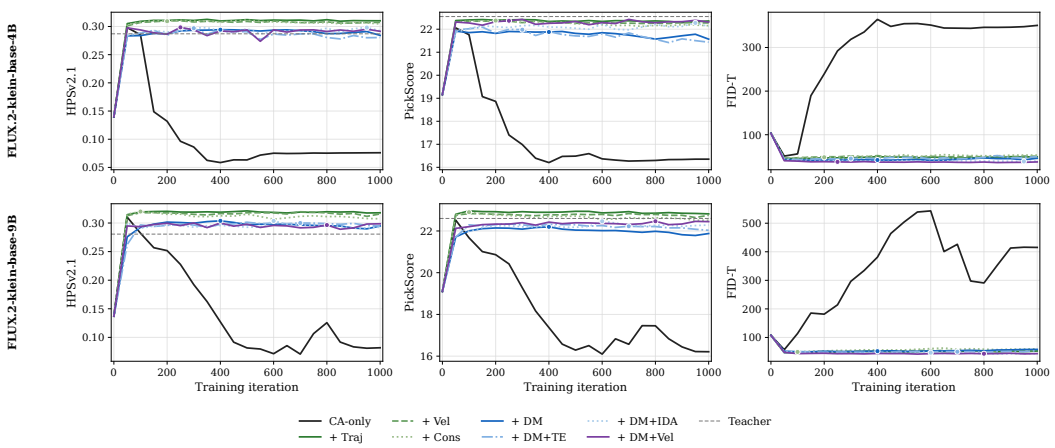


Figure 4: Training curves for FLUX.2-klein-base-4B and FLUX.2-klein-base-9B. Evaluation is conducted on 1K MSCOCO prompts. Gray dashed lines indicate teacher performance.

Although this local constraint does not guarantee agreement with the teacher's full trajectories, we find it sufficient for stabilization in our experiments. This admits an interpretation analogous to methods such as Flow-GRPO (Liu et al., 2025), except that the reward is provided internally by the teacher conditional posterior, whose gradient corresponds to the classifier score.

$$\max_{\theta} \mathbb{E}[r_\phi(\mathbf{x}_t, \mathbf{c})] - \beta \left\| \mathbf{v}_\theta - \mathbf{v}_\phi \right\|_2^2, \qquad r_\phi(\mathbf{x}_t, \mathbf{c}) = \log p_\phi(\mathbf{c} \mid \mathbf{x}_t). \tag{5.5}$$

The same principle can be applied to consistency variants. Following CTM (Kim et al., 2024) and ISG in SenseFlow (Ge et al., 2026), we also include the consistency trajectory loss (Cons). For $t > u > s$,

$$\mathcal{L}_{\text{Cons}} = \mathbb{E}_{t,\mathbf{x}_t}\left[\left\| G_\theta^{t\to s}(\mathbf{x}_t) - \text{sg}\left[G_\theta^{u\to s} \circ \Phi_{\phi,K}^{t\to u}(\mathbf{x}_t) \right] \right\|_2^2 \right]. \tag{5.6}$$

Final-step perceptual loss. During distillation driven by CA, we find oversaturation and simplified local details resembling over-guidance behaviors (Sadat et al., 2025) in some cases. Such local-detail degradation can persist along vanilla distribution matching or MSE-based transition anchoring. Perceptual losses (Zhao et al., 2026; Song et al., 2023) provide a feature-space objective that better reflects perceptual similarity than vanilla MSE. We therefore provide an option to apply perceptual supervision at the final student step in image space, where the teacher endpoint supports a reliable perceptual reference. Intermediate steps retain the original MSE anchoring objective to preserve stable transition. (Design details are in Section B.2)

6 Experiments

Setup. We evaluate SD3.5-M,L (Esser et al., 2024), and FLUX.2-klein-base-4B,9B (Labs, 2025). All students use four sampling steps, following (Luo et al., 2025; Ge et al., 2026). We compare transition anchoring (Traj, Vel, Cons), distribution matching (DM, DM+IDA (Ge et al., 2026), DM+TE), and hybrid regularization (DM+Vel). We use TDM (Luo et al., 2025) as the distribution-matching backbone and denote it by DM for brevity. For stabilization experiments, transition anchoring uses MSE as the endpoint loss by default. We evaluate on 10K MSCOCO prompts (COCO10K) (Lin et al., 2014). For evaluation metrics, we include Aesthetic Score (Aes.) (Schuhmann, 2022), HPSv2.1 (Wu et al., 2023), CLIPScore (Hessel et al., 2021), PickScore (Pick) (Kirstain et al., 2023), ImageReward (IR) (Xu et al., 2023), and FID against teacher-generated samples (FID-T) (Heusel et al., 2017). Implementation details are provided in Appendix C.1.

Main comparison. Figure 4 examines whether regularization sustains the initial benefits of CA-only distillation. CA-only initially improves but rapidly deteriorates, whereas both transition anchoring and distribution-matching variants stabilize CA in the evaluated interval.

Table 1: Main COCO10K results across four models; full six-metric results are in Tables 7 and 8. Transition anchoring, distribution matching, and hybrid denote the method families.

Method	SD3.5-M			SD3.5-L			FLUX.2-klein-base-4B			FLUX.2-klein-base-9B		
	HPS↑	Pick↑	FID-T↓	HPS↑	Pick↑	FID-T↓	HPS↑	Pick↑	FID-T↓	HPS↑	Pick↑	FID-T↓
Teacher	0.2809	22.37	–	0.2950	22.78	–	0.2870	**22.55**	–	0.2806	22.60	–
Stabilizers used during Classifier Score Distillation												
+ DM	0.3025	22.70	8.79	0.2885	22.66	13.95	0.2932	21.85	13.97	0.3006	22.11	23.37
+ DM+IDA	0.2984	22.68	11.98	0.2999	22.89	8.44	0.2963	22.22	10.76	0.3009	22.38	17.34
+ DM+TE	0.2889	22.37	7.87	0.2869	22.71	11.09	0.2954	21.95	16.74	0.2977	22.17	21.91
+ Traj	0.3102	**22.75**	11.18	0.3141	22.87	9.97	**0.3101**	22.30	20.37	**0.3200**	**22.89**	21.12
+ Vel	0.3099	22.67	12.96	0.3104	22.64	15.04	0.3085	22.29	22.65	0.3185	22.79	24.49
+ Cons	**0.3174**	22.57	19.19	**0.3179**	22.57	15.29	0.3084	22.33	21.18	0.3182	22.84	21.12
+ DM+Vel	0.2999	22.68	**7.09**	0.3068	**23.03**	**7.69**	0.2964	22.30	**9.76**	0.2944	22.39	**13.15**

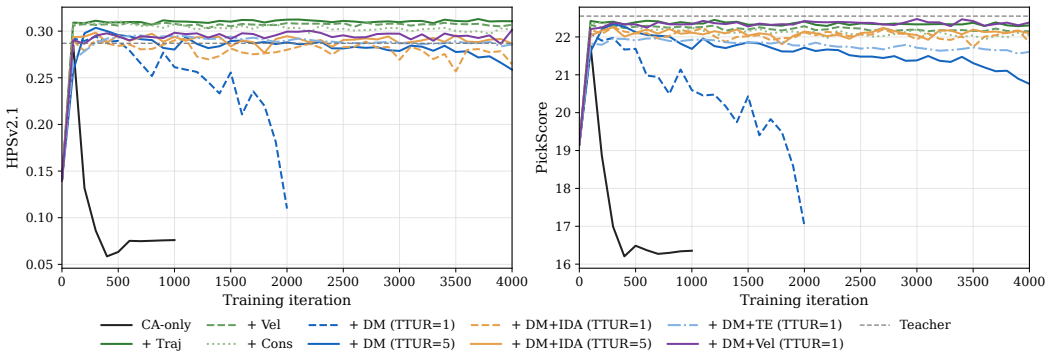


Figure 5: Long-horizon stability test on FLUX.2-klein-base-4B.

Table 1 compares quantitative metrics on MS-COCO10K across all four models. Transition anchoring tends to obtain stronger preference-oriented scores, whereas distribution-matching variants generally achieve lower FID-T, indicating closer agreement with the teacher distribution. For example, on FLUX.2-klein-base-4B, Traj achieves HPS 0.3101 and FID-T 20.37, compared with 0.2932 and 13.97 for DM. Figure 6 complements these metrics with qualitative comparisons. Full six-metric results (Tables 7 and 8) and training curves are in Appendix C.2.

Efficiency and long-horizon stability. Table 2 measures wall-clock time per student update on 4xH200. Relative to CA-only, Vel costs 1.19× and Traj 1.75× across both FLUX scales, compared

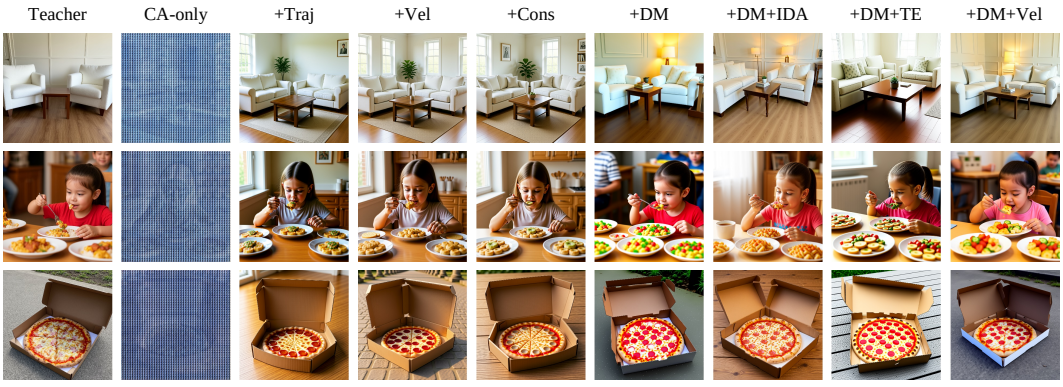


Figure 6: Qualitative comparison on FLUX.2-klein-base-4B across transition anchoring, distribution matching, and hybrid regularization.

Table 2: Mean wall-clock time in seconds per student update.

Model	CA-only	+ Traj	+ Vel	+ Cons	+ DM@5	+ DM+TE@1	+ DM+IDA@5	+ DM+Vel@1
FLUX.2-klein-base-4B	16.5	29.0 × 1.75	19.7 × 1.19	21.4 × 1.29	77.5 × 4.69	30.5 × 1.85	78.7 × 4.76	28.0 × 1.70
FLUX.2-klein-base-9B	16.3	28.6 × 1.75	19.4 × 1.19	21.0 × 1.29	77.3 × 4.74	30.2 × 1.85	78.2 × 4.79	27.5 × 1.69

with $4.69\times$ and $4.74\times$ for DM@5 on the 4B and 9B models, respectively. DM+TE@1 costs $1.85\times$, while DM+Vel@1 costs $1.70\times$ and $1.69\times$, respectively. Transition anchoring avoids repeated fake-model updates and therefore adds less overhead than high-TTUR DM.

Figure 5 extends training from 1K to 4K student updates. CA-only collapses within the first 1K updates, and DM with TTUR $= 1$ deteriorates by 2K updates; increasing TTUR delays this deterioration. DM+IDA, Adaptive TE (DM+TE), transition anchoring, and DM+Vel exhibit more stable preference scores over the tested 4K-update horizon.

Perceptual refinement. As illustrated in Figure 6, the compared stabilizers stabilize optimization but can still inherit CA-related oversaturation and simplified local details. We alleviate these artifacts through final-step perceptual supervision: intermediate steps retain MSE anchoring, and perceptual loss is applied only at the final student step. Table 3 and figure 7 compare this refinement with MSE-based transition anchoring. Final-step perceptual supervision improves local-detail preservation and reduces oversaturation while maintaining comparable preference performance. Full perceptual-loss ablations are provided in Appendix B.2.

Table 3: Final-step perceptual loss on COCO10K.

Loss	FLUX-4B			FLUX-9B		
	HPS↑	Pick↑	FID-T↓	HPS↑	Pick↑	FID-T↓
Traj + MSE	0.3101	22.30	20.37	0.3200	22.89	21.12
Traj + Perc. (final)	0.3042	22.51	14.12	0.3204	22.97	16.22
Cons + MSE	0.3084	22.33	21.18	0.3182	22.84	21.12
Cons + Perc. (final)	0.3012	22.50	13.76	0.3203	23.07	14.58



Figure 7: Trajectory anchoring with MSE (top) and final perceptual loss (bottom) on FLUX.2-klein-base-4B. Each example pairs the full image (left) with the enlarged region marked in red (right).

7 Conclusion

In this work, we investigate the mechanism and stabilization of classifier score distillation in text-to-image generation. Mechanistically, controlled Gaussian-mixture experiments show that finer condition groupings yield more mode-discriminative classifier-score directions and improve CA-only distillation. Score-space interpolation in text-to-image models further shows that retaining more fine-condition score information prolongs the effective optimization phase, but does not prevent collapse. For stabilization, we first examine fake-side tracking strategies and student-side regularization within DM. Despite these improvements, such approaches still require an online estimate of the evolving student score. We therefore move beyond the reverse KL formulation and study transition anchoring, which constrains student transitions using frozen teacher dynamics. We further study final-step perceptual supervision to alleviate CA-associated oversaturation and local-detail degradation. Across the evaluated models, both stabilizer families improve stability over the tested training horizons, with different trade-offs between preference metrics, teacher-distribution agreement, and computational cost. Together, these findings support a classifer-score-centered view of step distillation that broadens the choice of stabilizing constraints beyond distribution matching.

AI USE STATEMENT

In this work, we used generative AI tools to refine author-developed hypotheses, provide feedback on research methodology and experimental design, examine author-formulated mathematical claims, assist with translation and writing refinement, format tables and visual materials, and help implement code that was subsequently verified by the authors. We also used Qwen3.8-27B to generate coarsened and progressively refined prompt variants derived from a JourneyDB subset. Generative AI was not used to originate the conceptual framework, hypotheses, mathematical claims, or proofs, nor was it used for dataset cleaning or reformatting, result interpretation, or qualitative or thematic data analysis. These tasks were either performed without generative AI assistance or were not applicable to this work.

ETHICS STATEMENT

This work focuses on the technical aspects of diffusion model distillation, and we identify no ethical concerns specific to the proposed methodology.

REPRODUCIBILITY STATEMENT

The models, evaluation metrics, evaluation datasets, and implementation details are described in the main paper and appendix to support reproducibility.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *JMLR*, 2025.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Clement Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proc. AAAI*, 2025.

Dong Chen, Fangyun Wei, Ziyu Wan, Dongdong Chen, Jiawei Zhang, Jinjing Zhao, Sirui Zhang, Yang Yue, Zhiyang Liang, Baining Guo, et al. Lens: Rethinking training efficiency for foundational text-to-image models. *arXiv preprint arXiv:2605.21573*, 2026a.

Zilong Chen, Chaorui Deng, Kunchang Li, Hongyi Yuan, and Haoqi Fan. Scaling properties of text conditioning in visual generation. *arXiv preprint arXiv:2607.29679*, 2026b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Proc. NeurIPS*, 2021.

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 *Proc. ICML*, 2024.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proc. CVPR*, 2026.

Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P. Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin, 2024. URL `https://diffusionflow.github.io/`.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. Sense-flow: Scaling distribution matching for flow-based text-to-image distillation. In *Proc. ICLR*, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Proc. NeurIPS*, 2025.

Alexi Gladstone, Heng Ji, and Yilun Du. Explorative modeling: Unlocking a third pretraining axis and end-to-end generation, 2026. URL https://arxiv.org/abs/2607.27372.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *TMLR*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proc. EMNLP*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Proc. NeurIPS Workshop*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Proc. NeurIPS*, 2023.

Yufei Huang, Liangyu Yuan, Changxi Chi, Yunfan Liu, Cheng Tan, Siyuan Li, Jingbo Zhou, Haitao Lin, Chang Yu, and Stan Z. Li. Steinsgate: Adding causality to diffusions for long video generation via path integral. In *Proc. ICLR*, 2026a.

Yushi Huang, Xiangxin Zhou, Ruoyu Wang, Chi Zhang, Jun Zhang, and Tianyu Pang. Reinforcing few-step generators via reward-tilted distribution matching. *arXiv preprint arXiv:2605.26108*, 2026b.

Zemin Huang, Zhengyang Geng, Weijian Luo, and Guo-jun Qi. Flow generator matching. *arXiv preprint arXiv:2410.19310*, 2024.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Proc. NeurIPS*, 2024.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *Proc. ICLR*, 2024.

Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. In *Proc. NeurIPS*, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Proc. NeurIPS*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

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

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Continuous adversarial flow models. *arXiv preprint arXiv:2604.11521*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proc. ICLR*, 2023.

Dongyang Liu, Ruoyi Du, David Liu, Dengyang Jiang, Liangchen Li, Qilong Wu, Zhen Li, Steven C. H. Hoi, Hongsheng Li, and Peng Gao. High-fidelity two-step image generation via teacher-aligned end-to-end distillation. *arXiv preprint arXiv:2606.12575*, 2026a.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *Proc. ICLR*, 2026b.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *Proc. NeurIPS*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. ICLR*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. CVPR*, 2022.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Proc. NeurIPS*, 2023.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proc. ICCV*, 2025.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. In *Proc. ACL*, 2023.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In *Proc. NeurIPS*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *Proc. ICLR*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *Proc. ICLR*, 2022.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proc. ECCV*, 2024b.

Christoph Schuhmann. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proc. ICML*, 2023.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. JourneyDB: A benchmark for generative image understanding. In *Proc. NeurIPS*, 2023.

Ruoyu Wang, Boye Niu, Xiangxin Zhou, Yushi Huang, Tongliang Liu, and Chi Zhang. Exploring the design space of reward backpropagation for flow matching. *arXiv preprint arXiv:2606.11075*, 2026.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Proc. NeurIPS*, 2023.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026a.

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

Zhuguanyu Wu, Ruihao Gong, Yang Yong, Yushi Huang, Xiangyu Fan, Lei Yang, Dahua Lin, and Xianglong Liu. Sgmd: Score gradient matching distillation for few-step video diffusion distillation. In *Proc. ICML*, 2026b.

Zhang Xiaofeng, Aaron Courville, Michal Drozdzal, and Adriana Romero-Soriano. The intricate dance of prompt complexity, quality, diversity, and consistency in t2i models. *arXiv preprint arXiv:2510.19557*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Proc. NeurIPS*, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *Proc. NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proc. CVPR*, 2024b.

Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-evoke: From linear-scaling supervision to endless world. *arXiv preprint arXiv:2608.13546*, 2026.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and XIAOJUAN QI. Text-to-3d with classifier score distillation. In *Proc. ICLR*, 2024.

Liangyu Yuan, Yufei Huang, Mingkun Lei, Tong Zhao, Ruoyu Wang, Changxi Chi, Yiwei Wang, and Chi Zhang. Improving diffusion generalization with weak-to-strong segmented guidance. In *Proc. CVPR*, 2026.

Chuyang Zhao, Yifei Song, Hongfa Wang, Jianlong Yuan, Yuan Zhang, Siming Fu, Zhineng Chen, Huilin Deng, Haoyang Huang, and Nan Duan. Perceptual flow matching for few-step generative modeling. *arXiv preprint arXiv:2607.03524*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *Proc. ICLR*, 2026a.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *Proc. ICLR*, 2026b.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *Proc. ICLR*, 2026c.

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

APPENDIX

Contents

A Analyzing Classifier Score Distillation 15
A.1 Toy settings 15
A.2 Score interpolation and natural-language specificity analysis 18
B Stabilize Classifier Score Distillation 20
B.1 Adaptive TE for low-TTUR distribution matching 20
B.2 Loss ablations 21
B.3 Limitations, discussion and future directions 22
C Experiments 23
C.1 Experimental setup 23
C.2 Full quantitative results and training curves 24
C.3 Transition anchoring ablations 26
C.4 Comparison with De-DMD baselines 27
C.5 Additional qualitative results 28

A ANALYZING CLASSIFIER SCORE DISTILLATION

A.1 TOY SETTINGS

Synthetic 2d toy design. To study how condition granularity affects CA in the synthetic toy setting, we use a 2d Gaussian mixture across all toy experiments. Following AutoGuidance (Karras et al., 2024), we adopt its Gaussian-mixture implementation and score-network design. The data distribution consists of 16 equally weighted isotropic Gaussian components uniformly arranged on a circle of radius 2.5, each with standard deviation 0.15. We define $G \in \{2, 4, 8, 16\}$ condition groups by partitioning adjacent components, such that each group contains $16/G$ modes. Increasing G therefore yields finer-grained conditions, while the component geometry, covariance, mixture weights, and marginal data distribution remain fixed.

For a group c, let $\mathcal{I}_c$ denote its associated component indices. The corresponding conditional distribution is

$$p_{\text{data}}(\mathbf{x} \mid c) = \frac{1}{|\mathcal{I}_c|} \sum_{k \in \mathcal{I}_c} \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \sigma_{\text{data}}^2 \mathbf{I}\right), \tag{A.1}$$

where $\sigma_{\text{data}} = 0.15$. The unconditional distribution is obtained by uniformly mixing all 16 components. This construction allows us to vary the information provided by the condition while keeping the component geometry, covariance, mixture weights, and marginal data distribution fixed.

For the distillation setting. the student is a one-step conditional generator

$$\mathbf{x}_0^\theta = G_\theta(\mathbf{z}, c), \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, 5^2\mathbf{I}), \tag{A.2}$$

with four magnitude-preserving hidden blocks of width 64 and an 8-dimensional condition embedding. We sample $t \sim \mathcal{U}(0, 1)$, set $\sigma_t = t/(1-t)$, and construct

$$\mathbf{x}_t = \text{sg}\left[\mathbf{x}_0^\theta + \sigma_t \boldsymbol{\epsilon}\right]. \tag{A.3}$$

The CA direction is

$$\boldsymbol{\Delta}_{\text{CA}} = \mathbf{s}_c(\mathbf{x}_t, \sigma_t, c) - \mathbf{s}_u(\mathbf{x}_t, \sigma_t), \tag{A.4}$$

The student is optimized with the following objective:

$$\mathcal{L}_{\text{CA}} = \left\| \mathbf{x}_0^\theta - \text{sg}\left[\mathbf{x}_0^\theta + \boldsymbol{\Delta}_{\text{CA}}\right] \right\|_2^2. \tag{A.5}$$

For CA+DM, we set TTUR (Yin et al., 2024a) = 5 for better fake score approximation.

Full comparison across teacher, CA and CA+DM Figure 8 extends the CA-only results in Figure 2 with the pretrained multi-step teacher and CA+DM. The teacher recovers the target distribution across condition groupings, whereas CA-only distillation is sensitive to granularity. Adding DM reduces this sensitivity and improves mode recovery under coarse conditions.

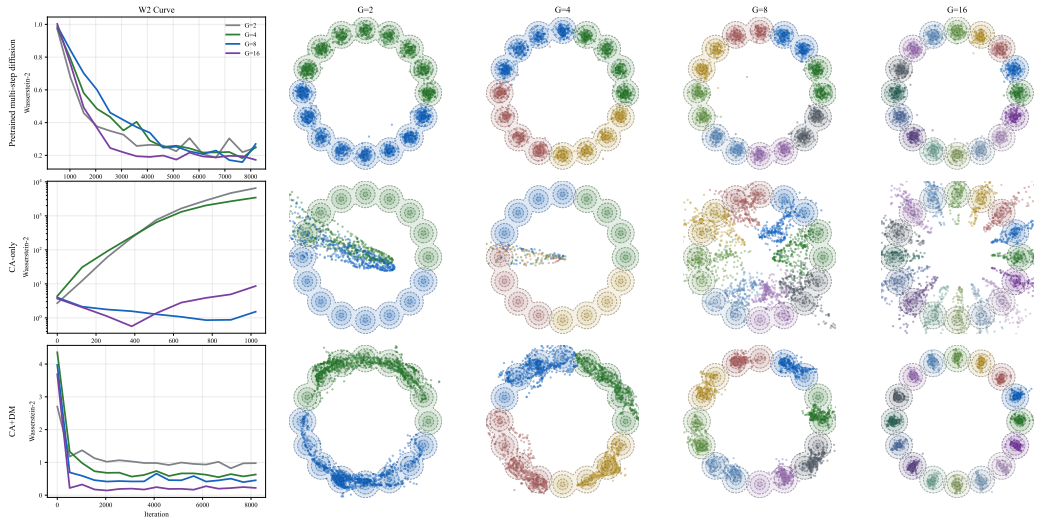


Figure 8: Controlled experiments on a 16-mode ring Gaussian mixture using learned teacher scores. Rows show the pretrained multi-step teacher, CA-only, and CA+DM. Finer conditions improve CA-only distillation, while DM stabilizes training under coarse conditions.

The gradient directions in Figure 9 exhibit the same trend. Coarse condition grouping produces ambiguous group-level directions, whereas finer conditions yield more mode-specific guidance. Incorporating DM reduces sensitivity to condition grouping, demonstrating the need for additional regularization even when the CSD direction is informative.

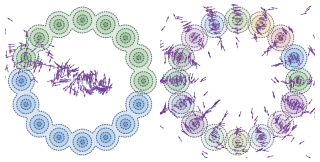


Figure 9: Visualization of classifier-score directions under coarse and fine condition granularity. Coarse grouping produces ambiguous group-level directions, while fine-grained conditions provide more mode-discriminative CA signals.

Condition granularity across learned and analytical scores. To further explore whether score-estimation error obscures the effect of condition granularity, we compare CA-only distillation using learned teacher scores and exact analytical scores. For the analytical score, under the Gaussian perturbation

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{A.6}$$

each component remains isotropic with variance $\tau_t^2 = \sigma_{\text{data}}^2 + \sigma_t^2$. The conditional and unconditional posterior are

$$r_k^{(c)}(\mathbf{x}_t, \sigma_t) = \frac{\exp\left(-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}_k\|_2^2}{2\tau_t^2}\right)}{\sum_{j \in \mathcal{I}_c} \exp\left(-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}_j\|_2^2}{2\tau_t^2}\right)}, \quad k \in \mathcal{I}_c, \quad r_k^{(u)}(\mathbf{x}_t, \sigma_t) = \frac{\exp\left(-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}_k\|_2^2}{2\tau_t^2}\right)}{\sum \exp\left(-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}_j\|_2^2}{2\tau_t^2}\right)}, \tag{A.7}$$

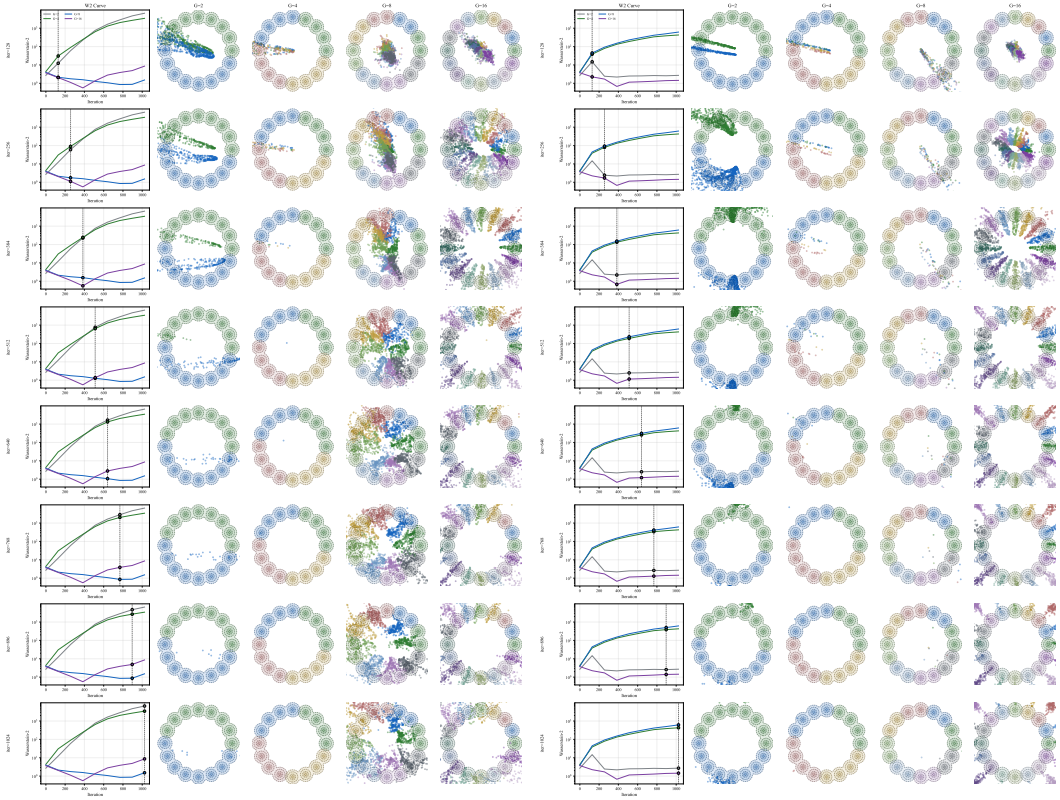
The corresponding analytical scores (Gu et al., 2025) are thus:

$$\mathbf{s}_c^*(\mathbf{x}_t, \sigma_t) = \sum_{k \in \mathcal{I}_c} r_k^{(c)}(\mathbf{x}_t, \sigma_t) \frac{\boldsymbol{\mu}_k - \mathbf{x}_t}{\tau_t^2}, \quad \mathbf{s}_u^*(\mathbf{x}_t, \sigma_t) = \sum_{k=1}^{16} r_k^{(u)}(\mathbf{x}_t, \sigma_t) \frac{\boldsymbol{\mu}_k - \mathbf{x}_t}{\tau_t^2}. \tag{A.8}$$

These closed-form scores remove neural score-estimation error from the CA signal.

As shown in Figure 10, the finest-grained setting, $G = 16$, consistently achieves the lowest Wasserstein-2 distance and the most accurate mode recovery under both (learned and analytical)

score sources. This agreement indicates that the benefit of fine-grained conditions is not merely an artifact of teacher approximation error. In practice, exact scores are generally unavailable, particularly for high-dimensional text-to-image models. To better resemble practical setting and given the agreement of trends on both cases, we thus keep learned score analysis as our default setting. Nevertheless, CA-only training remains unstable, with performance deteriorating after the optimal intermediate checkpoint. Finer conditions therefore strengthen CSD, but do not eliminate the need for additional regularization.



(a) Learned teacher scores (b) Analytical scores

Figure 10: CA-only distillation across condition granularities using learned teacher scores and analytical scores. The finest-grained setting $G = 16$ consistently achieves the best intermediate performance, while all settings eventually deteriorate without additional regularization.

Motivation behind the controlled setting. Condition granularity differs substantially across generative settings. Class labels in standard image datasets each cover thousands of diverse images, whereas an exact caption in a large-scale text-to-image dataset (Cai et al., 2025; Chen et al., 2026a; Esser et al., 2024; Labs, 2025) typically corresponds to only one or a few images. As summarized in Table 4, textual conditions therefore tend to specify a much narrower set of compatible modes than class labels. In an attempt to isolate the effect from dataset-specific factors, our synthetic setting fixes the marginal data distribution and varies the number of modes associated with each condition.

Table 4: Dataset-level illustration of condition granularity.

Setting	Condition	Approx. images/condition
CIFAR-10	Class label	5,000
ImageNet-1K	Class label	1,281
Large-scale T2I datasets	Exact caption	Typically close to 1

A.2 SCORE INTERPOLATION AND NATURAL-LANGUAGE SPECIFICITY ANALYSIS

Here we provide a detailed setting of the score interpolation experiment along with the results of a complementary analysis that probes the effect of increasing textual specificity on CSD.

Score interpolation setting. For each prompt P in a JourneyDB subset (Sun et al., 2023), we use Qwen3.8-27B (Qwen Team, 2026) to construct a coarsened version P_c whose semantic information is contained within P, *i.e.*, $I(P_c) \subseteq I(P)$. To exclude the effect of signal strength, we match the magnitude of the interpolated $\mathbf{g}_{\mathrm{CA}}^{(\beta)}$ to the original signal $\mathbf{g}_{\mathrm{CA}}^{(1)}$:

$$\bar{\mathbf{g}}_{\mathrm{CA}}^{(\beta)} = \mathbf{g}_{\mathrm{CA}}^{(\beta)} \cdot \mathrm{sg}\left[\frac{\left\|\mathbf{g}_{\mathrm{CA}}^{(1)}\right\|_2}{\left\|\mathbf{g}_{\mathrm{CA}}^{(\beta)}\right\|_2 + \epsilon}\right]. \tag{A.9}$$

We evaluate CA-only distillation with $\beta \in \{0, 0.25, 0.5, 0.75, 1, 1.25\}$ and CA+DM with $\beta \in \{0, 0.5, 1\}$ (TTUR=5), tracking HPSv2.1 (Wu et al., 2023) and CLIPScore (Hessel et al., 2021) throughout training.

The controlled score-space intervention in the main text shows that retaining more fine-condition information strengthens CSD while keeping the underlying generation task fixed. A natural follow-up question is whether this benefit can be realized simply by making natural-language prompts increasingly specific. Here we investigate this practical setting through progressively refined prompt variants.

Does increasing prompt specificity continue to benefit CSD? We use textual specificity as an auxiliary proxy for condition granularity, since adding semantic constraints generally narrows the set of outcomes compatible with a prompt (Mehrabi et al., 2023). Starting from each prompt in a JourneyDB subset (Sun et al., 2023), we use Qwen3.8-27B (Qwen Team, 2026) to construct six progressively refined variants G_0–G_5, with G_2 corresponding to the unchanged original prompt. We enforce semantic nesting,

$$I(G_0) \subseteq I(G_1) \subseteq \cdots \subseteq I(G_5), \tag{A.10}$$

such that each finer prompt preserves the visual information of its predecessor while introducing additional attributes, relations, scene context, actions, or compositional details (Hu et al., 2024; Huang et al., 2023). An example is shown below.

G0: Cyberpunk urban warfare scene.
G1: Cyberpunk urban warfare scene in a city ruin, daylight, dramatic lighting.
G2: Orange and black cyberpunk full body armor carrying a yellow ball in a gun fight action scene in an urban city ruin, daylight, warfare, dramatic lighting, photorealistic, reflections, smoke.
G3: An armored cyberpunk combatant in orange-and-black full-body armor carries a bright yellow ball while advancing through a gunfight in a ruined city street. Daylight filters through smoke, creating dramatic lighting, metallic reflections, and a photorealistic atmosphere of urban warfare.
G4: A lone cyberpunk combatant wearing scratched orange-and-black full-body armor runs through a shattered city boulevard, holding a bright yellow ball while exchanging gunfire with unseen opponents. Collapsed concrete buildings, abandoned vehicles, flying sparks, bullet impacts, and drifting smoke surround the figure. The full-body composition uses dramatic daylight, strong rim lighting, metallic reflections, and cinematic photorealism.
G5: A low-angle, full-body cinematic action photograph of a lone cyberpunk combatant in scratched, segmented orange-and-black armor sprinting from left to right through a devastated city boulevard. The figure secures a bright yellow ball beneath the left arm while firing a compact rifle with the right hand. Collapsed high-rises and burned vehicles frame the background, with concrete debris in the foreground and distant gunfire visible through dense gray smoke. Midday sunlight produces hard rim lighting, orange sparks, suspended dust, and crisp reflections across the armor. Shot with a 35 mm lens, shallow depth of field, realistic proportions, highly detailed materials, and photorealistic rendering.

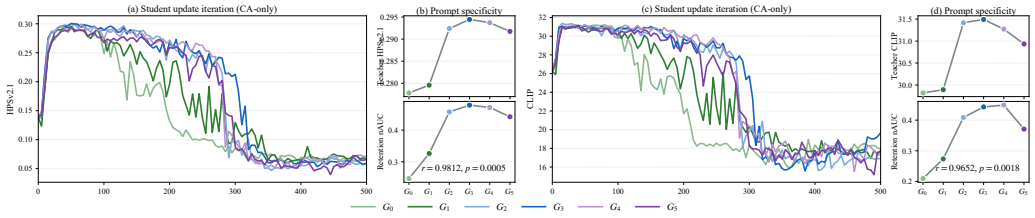


Figure 11: Prompt-specificity analysis on FLUX.2 Klein 4B. We compare CA-only training trajectories and retention nAUC across progressively refined prompts with the corresponding multistep-teacher performance.

As shown in Figure 11, all CA-only runs initially improve and eventually collapse. From G_0 to G_3, increasing prompt specificity generally prolongs the effective distillation phase. This trend does not persist indefinitely: at G_4 and G_5, additional specificity provides no consistent gain, with collapse occurring earlier than or comparably to G_3.

We hypothesize that the saturation arises from the teacher's limited ability to represent increasingly specific prompts. To examine this quantitatively, we measure the teacher performance by evaluating its multi-step ability to generate images for the same prompt groups. Teacher performance follows a similar non-monotonic trend, generally improving with specificity before saturating or reversing at higher levels. We further quantify the relationship between teacher performance and student distillation effectiveness using retention nAUC, defined as the normalized area under the student distillation curve. For a prompt group G_i,

$$\mathrm{nAUC}(G_i) = \frac{1}{T}\int_0^T \mathrm{clip}\left(\frac{q_i(t) - q_i(0)}{\max_{0\le\tau\le T} q_i(\tau) - q_i(0)}, 0, 1\right) \mathrm{d}t, \tag{A.11}$$

where $q_i(t)$ denotes the metric value (*e.g.*HPSv2.1) at update t, and T is the training horizon. A larger nAUC indicates that the student retains a greater fraction of its best observed improvement throughout training.

As shown in panels (b) and (d) of Figure 11, teacher performance closely covaries with student retention for both HPSv2.1 ($r_{\mathrm{P}} = 0.981$, $p < 0.001$) and CLIPScore ($r_{\mathrm{P}} = 0.965$, $p = 0.002$). These coupled trends are consistent with a teacher-supported regime: additional textual constraints benefit CSD only while they remain reliably represented by the pretrained teacher.

Why is vanilla textual specificity an imperfect measure of condition granularity? Despite enforcing semantic containment between successive prompt levels, the construction above remains controlled only at the language level. In particular, multiple layers of $I(G_i) \subseteq I(G_{i+1})$ does not imply a corresponding nested or monotonic relationship after tokenization and text encoding. Additional details may alter the text representation nonlinearly, interact with existing tokens, or introduce information that is redundant, stylistic, or weakly represented by the teacher. Consequently, multiple layers of progressively longer or more specific prompts do not provide a direct measure of condition granularity in the resulting teacher score field (Chen et al., 2026b). For this reason, we treat textual specificity as an auxiliary natural-language probe rather than the controlled definition of condition granularity.

B Stabilize Classifier Score Distillation

B.1 Adaptive TE for low-TTUR distribution matching

We detail the sample-wise calibration used in Adaptive TE, a low-TTUR DM variant based on SGMD's negative-residual (NR) potential (Wu et al., 2026b). Consider the linear perturbation path

$$\mathbf{x}_0^\theta = G_\theta(\mathbf{z}, \mathbf{c}), \qquad \mathbf{x}_t = (1-t)\mathbf{x}_0^\theta + t\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{B.1}$$

Let $q_{\theta,t}(\cdot \mid \mathbf{c})$ denote the noisy student distribution, with score $\mathbf{s}_q = \nabla_{\mathbf{x}_t} \log q_{\theta,t}(\mathbf{x}_t \mid \mathbf{c})$. We use $\mathbf{s}_p$ for the teacher score and $\mathbf{s}_\psi$ for the online fake score. Throughout, $t \in (0,1)$, gradients are row vectors, and expectations are over $(t, \mathbf{z}, \boldsymbol{\epsilon}, \mathbf{c})$ unless conditioned otherwise. Score arguments are omitted when clear. We retain the path factor $1-t$ explicitly below; it is suppressed in the main-text notation of Equation (5.1).

The denoising score identity (Zhou et al., 2024; Wu et al., 2026b) relates the sampled perturbation score to the marginal student score:

$$\widetilde{\mathbf{s}}_{t|0} = \frac{(1-t)\,\mathrm{sg}[\mathbf{x}_0^\theta] - \mathbf{x}_t}{t^2}, \qquad \mathbb{E}\big[\widetilde{\mathbf{s}}_{t|0} \mid \mathbf{x}_t, t, \mathbf{c}\big] = \mathbf{s}_q. \tag{B.2}$$

The stop-gradient operator preserves the forward value while blocking the direct dependence on $\mathbf{x}_0^\theta$ in the numerator; gradients through $\mathbf{x}_t$ remain active. The NR potential is

$$\ell_{\mathrm{NR}} = -\left\|\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi(\mathbf{x}_t, t, \mathbf{c})\right\|_2^2. \tag{B.3}$$

During student updates, ψ is fixed, but the fake-score network retains its input gradients. Let $\mathbf{J}_G = \partial G_\theta(\mathbf{z}, \mathbf{c})/\partial\theta$ and $\mathbf{J}_\psi = \partial \mathbf{s}_\psi(\mathbf{x}_t, t, \mathbf{c})/\partial \mathbf{x}_t$. Under this stop-gradient convention, $\partial \widetilde{\mathbf{s}}_{t|0}/\partial \mathbf{x}_0^\theta = -(1-t)\mathbf{I}/t^2$, giving

$$\begin{aligned} \nabla_{\mathbf{x}_0^\theta} \ell_{\mathrm{NR}} &= \frac{2(1-t)}{t^2}\left(\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi\right)\left(\mathbf{I} + t^2 \mathbf{J}_\psi\right), \\ \nabla_\theta \ell_{\mathrm{NR}} &= \nabla_{\mathbf{x}_0^\theta} \ell_{\mathrm{NR}}\, \mathbf{J}_G. \end{aligned} \tag{B.4}$$

Evaluating this gradient requires a vector–Jacobian product through the input of the frozen fake-score network.

For the DM component, we use the detached surrogate

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}\big[(1-t)\left\langle \mathrm{sg}[\mathbf{s}_\psi - \mathbf{s}_p], \mathbf{x}_0^\theta \right\rangle\big], \tag{B.5}$$

which does not differentiate through either score network. For a fixed sample-dependent coefficient λ, the NR-regularized objective is

$$\mathcal{L}_\theta = \mathcal{L}_{\mathrm{DM}} + \mathbb{E}[\lambda \ell_{\mathrm{NR}}]. \tag{B.6}$$

Combining the DM and NR gradients and adding and subtracting $\mathbf{s}_q$ yields

$$\begin{aligned} \nabla_\theta \mathcal{L}_\theta = &-\mathbb{E}[(1-t)(\mathbf{s}_p - \mathbf{s}_q)\mathbf{J}_G] \\ &+ \mathbb{E}\left[(1-t)(\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi)\left[\frac{2\lambda}{t^2}\left(\mathbf{I} + t^2\mathbf{J}_\psi\right) - \mathbf{I}\right]\mathbf{J}_G\right] \\ &+ \mathbb{E}\big[(1-t)(\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_q)\mathbf{J}_G\big]. \end{aligned} \tag{B.7}$$

The first term is the ideal reverse-KL gradient. The remaining terms separate a sampled calibration residual from a sample-to-marginal discrepancy. Empirically, vanilla NR failed to stabilize the distillation with TTUR=1 (as illustrated in Figure 12).

The NR-term gradient depends on both the timestep and the local fake-score Jacobian. A scalar coefficient cannot generally invert this transformation, but its action can be fitted to the sampled residual. We therefore calibrate a sample-level coefficient rather than using a single fixed weight. Define

$$\widehat{\boldsymbol{\delta}} = \mathrm{sg}\big[\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi\big], \qquad \widehat{\mathbf{u}} = \widehat{\boldsymbol{\delta}}\left(\mathbf{I} + t^2\mathbf{J}_\psi\right). \tag{B.8}$$

By Equation (B.2), the sampled residual satisfies

$$\mathbb{E}\Big[\widehat{\boldsymbol{\delta}} \mid \mathbf{x}_t, t, \mathbf{c}\Big] = \mathbf{s}_q - \mathbf{s}_\psi.$$

$\widehat{\boldsymbol{\delta}}$ provides a sampled target for compensating the fake-score mismatch, we therefore refer to it as the TE term for sample-level tracking-error. Before multiplication by $\mathbf{J}_G$, the sampled residual in the second term of Equation (B.7) is proportional to $(2\lambda/t^2)\widehat{\mathbf{u}} - \widehat{\boldsymbol{\delta}}$. For each sample independently, we solve the regularized local fit

$$\min_{\lambda \geq 0} \left\{ \left\| \frac{2\lambda}{t^2}\widehat{\mathbf{u}} - \widehat{\boldsymbol{\delta}} \right\|_2^2 + \varepsilon \left(\frac{2\lambda}{t^2} \right)^2 \right\}, \qquad \varepsilon > 0. \tag{B.9}$$

All norms and inner products in this calibration are evaluated over the coordinates of a single sample. Its detached solution and the resulting sample-wise TE regularizer are

$$\widehat{\lambda}_{\mathrm{TE}} = \mathrm{sg}\left[\frac{t^2}{2} \frac{\max\left(\left\langle \widehat{\mathbf{u}}, \widehat{\boldsymbol{\delta}} \right\rangle, 0 \right)}{\|\widehat{\mathbf{u}}\|_2^2 + \varepsilon} \right], \widehat{\lambda}_{\mathrm{TE}} \ell_{\mathrm{TE}} = -\widehat{\lambda}_{\mathrm{TE}} \left\| \widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_\psi(\mathbf{x}_t, t, \mathbf{c}) \right\|_2^2. \tag{B.10}$$

After applying the adaptive coefficient, it manages to sustain the distillation (Figure 12). In practice, the coefficient is detached during student updates, the changes of coefficient througout the distillation is also illustrated in Figure 12.

Limitation. Despite its empirical effectiveness, the calibration minimizes the regularized sample-level objective, not the full parameter-space discrepancy in Equation (B.7). Moreover, the sample-to-marginal term need not vanish: although $\widetilde{\mathbf{s}}_{t|0} - \mathbf{s}_q$ has zero conditional mean given $(\mathbf{x}_t, t, \mathbf{c})$, its product with $\mathbf{J}_G$ need not, because the generator Jacobian depends on the underlying latent sample. Adaptive TE is therefore an empirical low-TTUR regularizer rather than an exact correction of the marginal DM gradient, and it retains the online fake-score model.

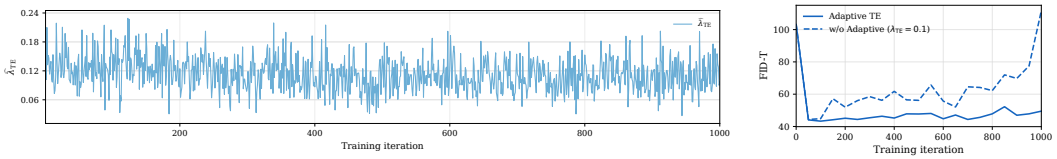


Figure 12: Adaptive TE on FLUX.2-klein-base-4B. Left: evolution of the adaptive coefficient $\widehat{\lambda}_{\mathrm{TE}}$. Right: FID-T comparison between adaptive weighting and a fixed coefficient $\lambda = 0.1$.

B.2 Loss ablations

CA is derived from the same conditional-unconditional score gap as CFG and can exhibit similar over-guidance behavior, including brightness and saturation drift and loss of local detail (Ho & Salimans, 2021).

We investigate whether perceptual supervision can alleviate these issues within transition anchoring. Following (Zhao et al., 2026), we use ConvNeXt (Liu et al., 2022) as the perceptual feature extractor and compare two variants against vanilla MSE on FLUX.2-klein-base-4B and FLUX.2-klein-base-9B: (1) *Perceptual (full)*, which applies perceptual matching at every student step after denoising the student and teacher states to image space; and (2) *Perceptual (final)*, which applies perceptual matching only at the final student step while retaining latent-space MSE for intermediate steps.

As illustrated in Table 5 and Figure 13, full perceptual matching is not consistently beneficial and can introduce substantial degradation. We attribute this to the role of intermediate teacher states: directly denoising to zero provides a reliable stabilizing anchor, but the resulting images are not necessarily high-quality perceptual references. Applying perceptual supervision to these intermediate predictions can therefore introduce misleading image-space constraints. To preserve the anchoring effect of intermediate states while still benefiting from perceptual supervision, we apply perceptual matching only at the final student step, ensuring that the teacher rollout is reliable enough for perceptual supervision.

We also provide qualitative comparisons in Figures 25 and 26. Compared with the teacher, vanilla DM and MSE-based transition anchoring often produce sharper outputs, but can also exhibit brightness drift, oversaturation, or simplified local structure. Applying perceptual supervision only at the

final step alleviates oversaturation and preserves local details while maintaining comparable quantitative performance.

Table 5: Loss ablations on FLUX.2-klein-base-4B and FLUX.2-klein-base-9B, evaluated on COCO10K.

Method	Loss	FLUX.2-klein-base-4B						FLUX.2-klein-base-9B					
		Aes.↑	HPS↑	CLIP↑	Pick↑	IR↑	FID-T↓	Aes.↑	HPS↑	CLIP↑	Pick↑	IR↑	FID-T↓
Traj	MSE	5.752	0.3101	30.90	22.30	1.093	20.37	5.706	0.3200	31.37	22.89	1.287	21.12
	Perceptual (full)	5.526	0.2760	30.50	21.64	0.646	16.53	5.560	0.2960	30.51	21.99	0.926	19.12
	Perceptual (final)	5.783	0.3042	31.07	22.51	1.064	14.12	5.674	0.3204	31.39	22.97	1.274	16.22
Cons	MSE	5.791	0.3084	30.79	22.33	1.080	21.18	5.799	0.3182	31.20	22.84	1.231	21.12
	Perceptual (full)	5.556	0.2682	28.99	20.79	0.477	39.37	5.573	0.2792	29.65	21.20	0.634	37.93
	Perceptual (final)	5.798	0.3012	31.02	22.50	1.057	13.76	5.709	0.3203	31.33	23.07	1.264	14.58

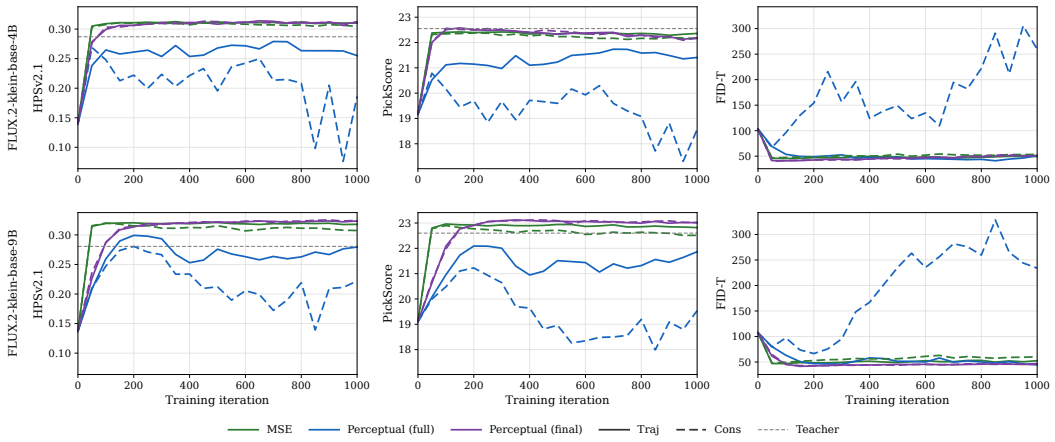


Figure 13: Loss-ablation training curves on FLUX.2-klein-base-4B (top) and FLUX.2-klein-base-9B (bottom). Gray dashed lines indicate teacher performance.

B.3 Limitations, discussion and future directions

Limitation Our empirical study is limited to text-to-image distillation on the Stable Diffusion (Esser et al., 2024) and FLUX (Labs, 2025) model families, and its conclusions may not directly generalize to other architectures, modalities, or scales. Although final-step perceptual supervision alleviates oversaturation and simplified local structure in our tested transition anchoring setting, its effectiveness has not been systematically validated across all model scales. Moreover, well-tuned distribution-matching methods, particularly when combined with complementary objectives such as adversarial training (Yin et al., 2024a; Ge et al., 2026) or reinforcement learning (Zheng et al., 2026b; Wang et al., 2026), may retain a higher performance ceiling. Our controlled comparison does not exhaustively investigate such compound training recipes.

Discussion and future directions. For potential future works, exploring alternative rollout constructions and combinations of transition anchoring with complementary few-step distillation techniques remains an important direction. This analysis could further be extended to video and world-model distillation, particularly for long-horizon generation (Yin et al., 2026; Huang et al., 2026a), where the computational and optimization burden of online distribution matching may become increasingly significant. More broadly, we hope this work contribute a step further towards raising the awareness of understanding CFG across different settings. In ImageNet, CFG is effective on early diffusion based attempts, while AutoGuidance (Karras et al., 2024) surpasses it on EDM2, RAE (Zheng et al., 2026a) further reveals that it's even inferior to unguided generation. In contrast, modern text-conditioned models still rely heavily on CFG if without additional post-training, and the same classifier-score signal can be strong enough to drive few-step distillation.

C EXPERIMENTS

C.1 EXPERIMENTAL SETUP

Implementation details. Table 6 summarizes the hyperparameters used across four model configurations from two model families. Because the objectives induce different gradient scales and optimization dynamics, a single shared hyperparameter configuration can systematically favor some methods over others. We therefore tune method-specific hyperparameters under a comparable search budget using the same validation protocol. The prompt set used for distillation is a subset constructed from JourneyDB prompt set. To isolate the contribution of each stabilizer, we do not combine the evaluated methods with auxiliary stabilization techniques beyond those explicitly included in our comparison (Ge et al., 2026; Yin et al., 2024a).

Table 6: Hyperparameters settings. BS reports the per-GPU batch size multiplied by four GPUs; TTUR is the number of fake-score-network updates per student update.

Model	Method	BS	$w_{\mathrm{CA}}(\alpha)$	w_{DM}	w_{path}	K	TTUR	η_{stu}	η_{fake}
SD3.5-M	CA-only	32×4	4	–	–	–	–	5×10^{-6}	–
	+ Traj	32×4	4	–	0.5	4	–	5×10^{-6}	–
	+ Vel	32×4	4	–	0.5	1	–	5×10^{-6}	–
	+ Cons	32×4	4	–	0.1	1	–	5×10^{-6}	–
	+ DM	32×4	4	1	–	–	1	5×10^{-6}	4×10^{-5}
	+ DM+IDA	32×4	4	1	–	–	1	5×10^{-6}	4×10^{-5}
	+ DM+TE	32×4	4	1	–	–	1	1×10^{-6}	1×10^{-6}
	+ DM+Vel	32×4	4	1	0.5	1	1	5×10^{-6}	4×10^{-5}
SD3.5-L	CA-only	16×4	4	–	–	–	–	2×10^{-6}	–
	+ Traj	16×4	4	–	0.5	4	–	2×10^{-6}	–
	+ Vel	16×4	4	–	0.5	1	–	2×10^{-6}	–
	+ Cons	16×4	4	–	0.1	1	–	2×10^{-6}	–
	+ DM	16×4	4	1	–	–	5	1×10^{-7}	5×10^{-7}
	+ DM+IDA	16×4	4	1	–	–	5	2×10^{-7}	5×10^{-7}
	+ DM+TE	16×4	4	1	–	–	1	1×10^{-7}	5×10^{-7}
	+ DM+Vel	16×4	4	1	0.5	1	1	2×10^{-6}	2×10^{-5}
FLUX-4B	CA-only	16×4	4	–	–	–	–	1×10^{-6}	–
	+ Traj	16×4	4	–	0.5	4	–	1×10^{-6}	–
	+ Vel	16×4	4	–	0.5	1	–	1×10^{-6}	–
	+ Cons	16×4	4	–	0.5	1	–	1×10^{-6}	–
	+ DM	16×4	4	1	–	–	5	5×10^{-7}	5×10^{-7}
	+ DM+IDA	16×4	4	1	–	–	5	1×10^{-6}	1×10^{-6}
	+ DM+TE	16×4	4	1	–	–	1	1×10^{-6}	1×10^{-6}
	+ DM+Vel	16×4	4	1	0.5	1	1	1×10^{-6}	1×10^{-5}
FLUX-9B	CA-only	8×4	4	–	–	–	–	1×10^{-6}	–
	+ Traj	8×4	4	–	0.5	4	–	1×10^{-6}	–
	+ Vel	8×4	4	–	0.5	1	–	1×10^{-6}	–
	+ Cons	8×4	4	–	0.5	1	–	1×10^{-6}	–
	+ DM	8×4	4	1	–	–	5	5×10^{-7}	5×10^{-7}
	+ DM+IDA	8×4	4	1	–	–	5	1×10^{-6}	1×10^{-6}
	+ DM+TE	8×4	4	1	–	–	1	5×10^{-7}	5×10^{-7}
	+ DM+Vel	8×4	4	1	0.5	1	1	1×10^{-6}	1×10^{-5}

We optimize the student and, when used, the online fake-score model with AdamW, using $(\beta_1, \beta_2) = (0, 0.999)$, weight decay 10^{-4}. Learning rates are constant without warm-up. Training uses four GPUs with fully sharded data parallelism, without gradient accumulation. Computational-cost measurements are performed on four NVIDIA H200 GPUs.

C.2 FULL QUANTITATIVE RESULTS AND TRAINING CURVES

Tables 7 and 8 provide the full six-metric results corresponding to Table 1.

Table 7: Full six-metric COCO10K results for the SD series. @x denotes a TTUR ratio of x; @1/5 denotes TTUR$=1$ for SD3.5-M and TTUR$=5$ for SD3.5-L. Transition anchoring, distribution matching, and hybrid denote the method families.

Method	SD3.5-M						SD3.5-L					
	Aes.↑	HPS↑	CLIP↑	Pick↑	IR↑	FID-T↓	Aes.↑	HPS↑	CLIP↑	Pick↑	IR↑	FID-T↓
Teacher	5.319	0.2809	31.80	22.37	0.955	–	5.500	0.2950	31.84	22.78	1.062	–
Stabilizers used during Classifier Score Distillation												
+ DM@1/5	5.466	0.3025	31.69	22.70	1.138	8.79	5.568	0.2885	31.70	22.66	1.016	13.95
+ DM+IDA@1/5	5.514	0.2984	**31.87**	22.68	1.118	11.98	5.562	0.2999	31.91	22.89	1.091	8.44
+ DM+TE@1	5.406	0.2889	31.40	22.37	0.953	7.87	5.514	0.2869	31.71	22.71	1.005	11.09
+ Traj	5.592	0.3102	31.77	**22.75**	**1.221**	11.18	5.642	0.3141	31.74	22.87	**1.205**	9.97
+ Vel	5.633	0.3099	31.68	22.67	1.200	12.96	**5.685**	0.3104	31.65	22.64	1.162	15.04
+ Cons	**5.723**	**0.3174**	31.79	22.57	1.216	19.19	5.677	**0.3179**	**32.11**	22.57	1.195	15.29
+ DM+Vel@1	5.473	0.2999	31.57	22.68	1.094	**7.09**	5.569	0.3068	31.76	**23.03**	1.171	**7.69**

Table 8: Full six-metric COCO10K results for the FLUX.2-klein-base series. @x denotes a TTUR ratio of x. Transition anchoring, distribution matching, and hybrid denote the method families.

Method	FLUX.2-klein-base-4B						FLUX.2-klein-base-9B					
	Aes.↑	HPS↑	CLIP↑	Pick↑	IR↑	FID-T↓	Aes.↑	HPS↑	CLIP↑	Pick↑	IR↑	FID-T↓
Teacher	5.428	0.2870	**31.38**	**22.55**	0.949	–	5.258	0.2806	**31.72**	22.60	0.999	–
Stabilizers used during Classifier Score Distillation												
+ DM@5	5.591	0.2932	31.06	21.85	0.935	13.97	5.552	0.3006	30.60	22.11	0.983	23.37
+ DM+IDA@5	5.538	0.2963	31.09	22.22	0.948	10.76	5.545	0.3009	30.92	22.38	1.030	17.34
+ DM+TE@1	5.710	0.2954	31.02	21.95	0.947	16.74	5.588	0.2977	31.08	22.17	1.023	21.91
+ Traj	5.752	**0.3101**	30.90	22.30	**1.093**	20.37	5.706	**0.3200**	31.37	**22.89**	1.287	21.12
+ Vel	**5.836**	0.3085	30.67	22.29	1.089	22.65	5.718	0.3185	31.23	22.79	**1.294**	24.49
+ Cons	5.791	0.3084	30.79	22.33	1.080	21.18	**5.799**	0.3182	31.20	22.84	1.231	21.12
+ DM+Vel@1	5.592	0.2964	31.28	22.30	1.019	**9.76**	5.442	0.2944	31.19	22.39	1.025	**13.15**

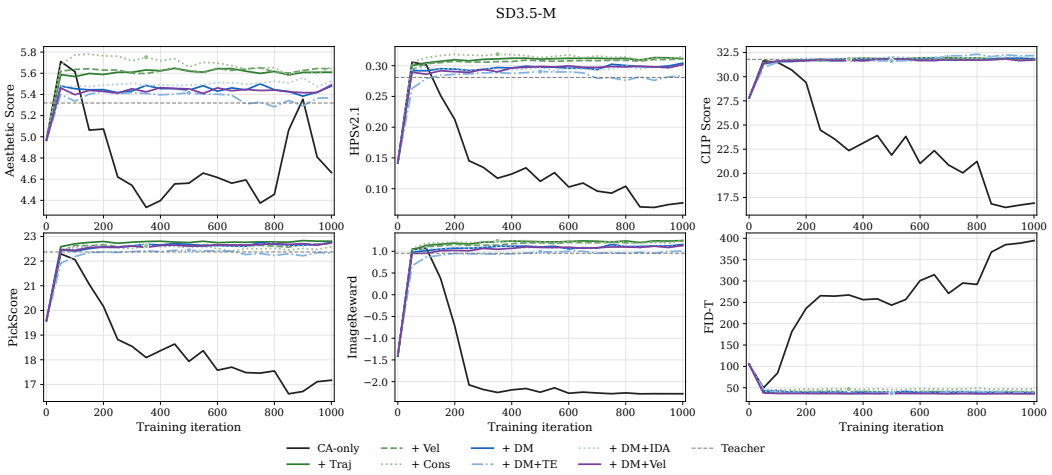


Figure 14: Full training curves for SD3.5-M.

We report the complete six-metric training curves for all four model scales in Figures 14 to 17. All curves use the 1K evaluation protocol and raw measurements at 50-iteration intervals. CA-only diagnostic curves are retained to expose collapse. Gray dashed lines show Teacher performance.

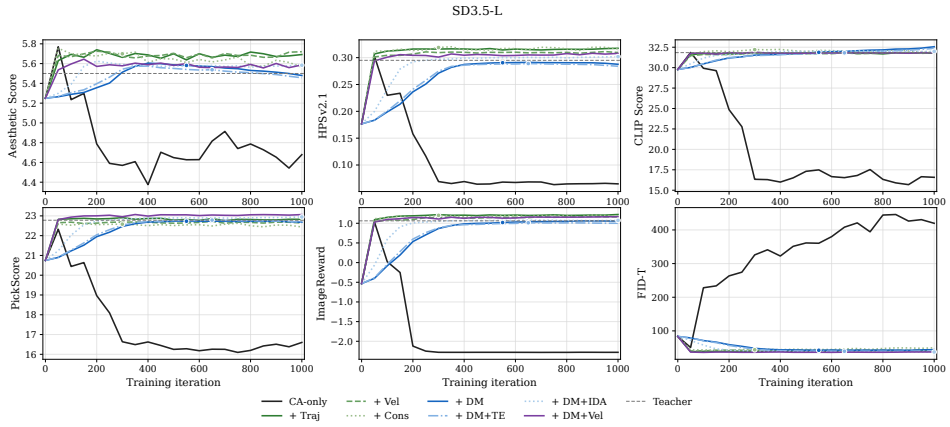


Figure 15: Full training curves for SD3.5-L.

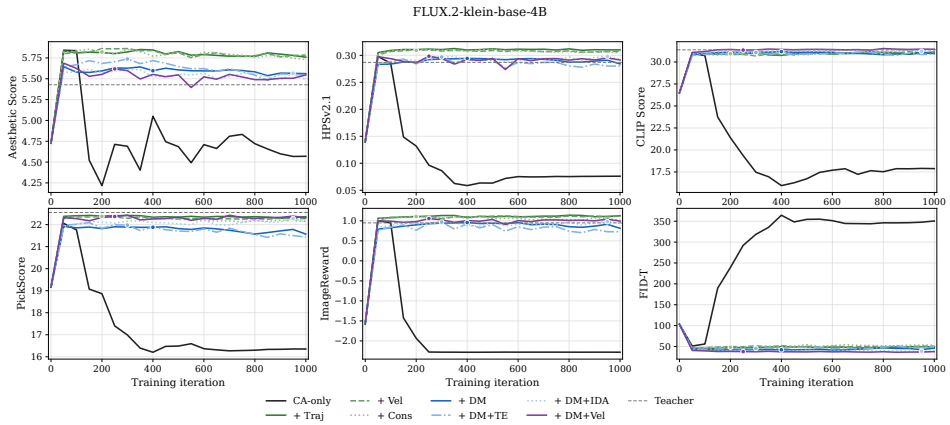


Figure 16: Full training curves for FLUX.2-klein-base-4B.

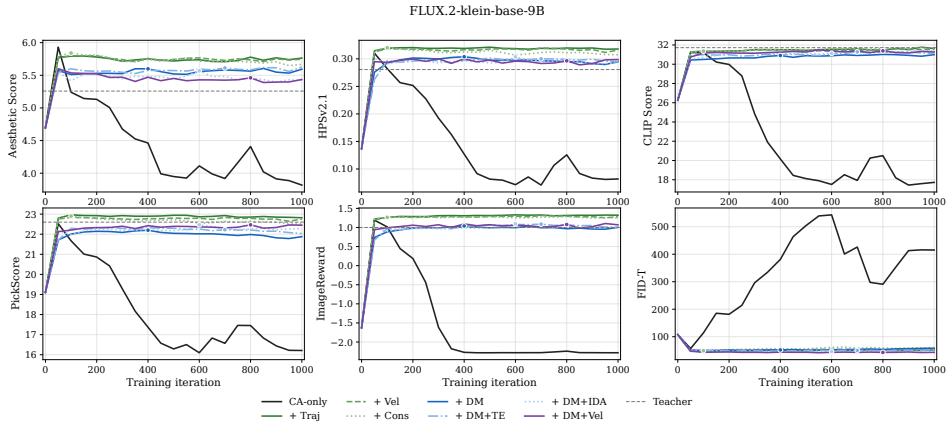


Figure 17: Full training curves for FLUX.2-klein-base-9B.

C.3 TRANSITION ANCHORING ABLATIONS

Trajectory ablations. We ablate the number of trajectory steps ($K \in \{1, \ldots, 6\}$), trajectory-loss ratio ($r \in \{0.05, 0.10, 0.50, 1.00\}$), and velocity CFG weight ($w \in \{1, \ldots, 5\}$) for FLUX.2-klein-base-4B. Figure 18 reports the corresponding training curves.

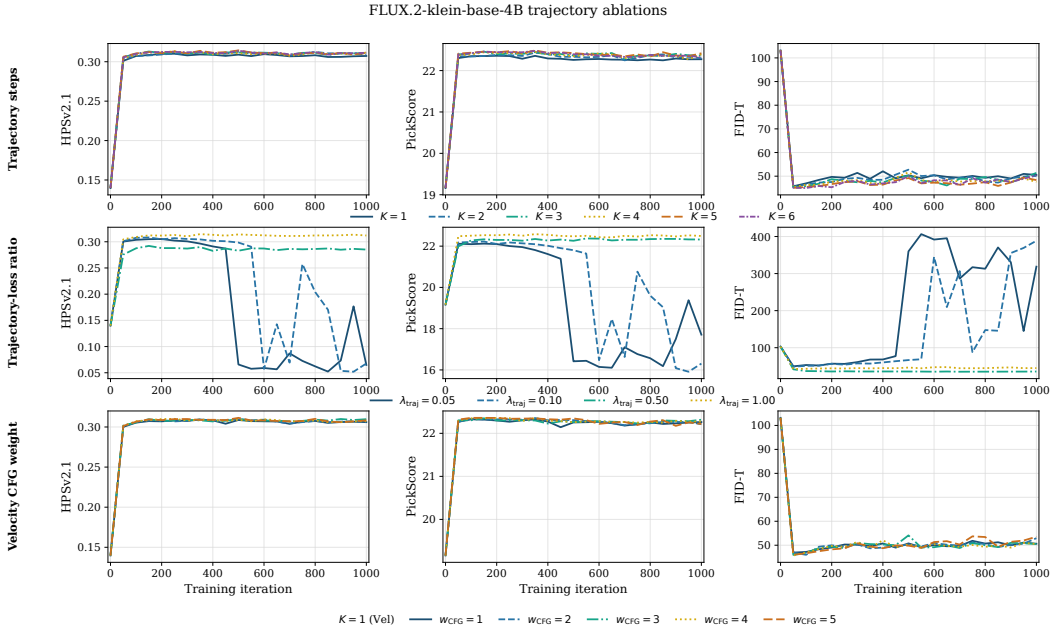


Figure 18: Transition anchoring ablations on FLUX.2-klein-base-4B. Rows vary the number of trajectory steps K, trajectory-loss ratio λ_{traj}, and velocity CFG weight w_{CFG} at $K = 1$ (Vel), respectively.

Trajectory-based objectives were originally developed for diffusion distillation (Salimans & Ho, 2022; Kim et al., 2024), motivating the question of how effectively they distill text-to-image models without CA. We compare MSE-based Traj and Cons with and without CA in Figure 19. Although omitting CA yields lower FID-T, indicating closer agreement with the teacher distribution, it produces substantially lower HPS and PickScore. These results highlight the contribution of CA to preference-oriented performance beyond standalone transition anchoring.

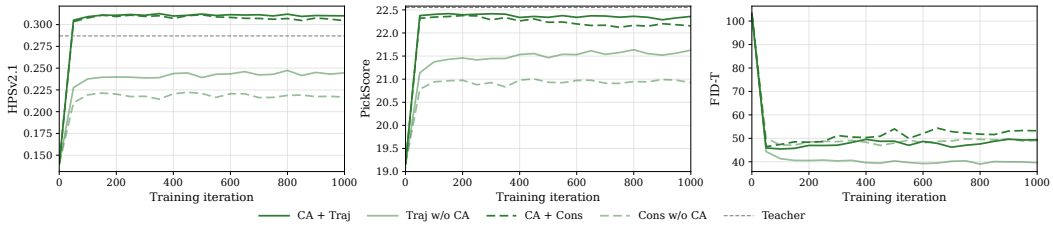


Figure 19: Traj and Cons with and without CA on FLUX.2-klein-base-4B, evaluated on COCO1K during training. Gray dashed lines indicate teacher performance.

C.4 Comparison with De-DMD baselines

Here we compare transition anchoring variants (Traj, Vel) with additional regularizers from decoupled DMD (Liu et al., 2026b) on FLUX.2-klein-base-4B. The unstable CA-only, DataKL (Liu et al., 2026b), and GAN curves are shown over the full training horizon. Figure 20 tracks their clean-latent mean and standard deviation alongside HPSv2.1. CA-only and GAN progressively deteriorate under unstable training. DataKL preserves the latent statistics, but its declining HPSv2.1 reveals a corresponding loss in image quality.

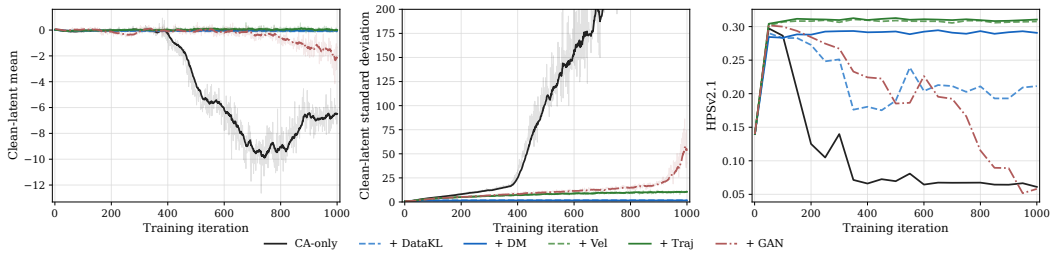


Figure 20: Comparison with GAN and DataKL, proposed in Decoupled DMD (Liu et al., 2026b), on FLUX.2-klein-base-4B, reporting the clean-latent mean and standard deviation alongside HPSv2.1. The latent-statistic curves show an exponential moving average over 25 updates, with the unsmoothed measurements shown faintly underneath.

C.5 ADDITIONAL QUALITATIVE RESULTS

We provide qualitative comparisons for all four model scales, followed by perceptual-loss comparisons on FLUX.2-klein-base-4B.



Figure 21: Qualitative comparison on SD3.5-M.

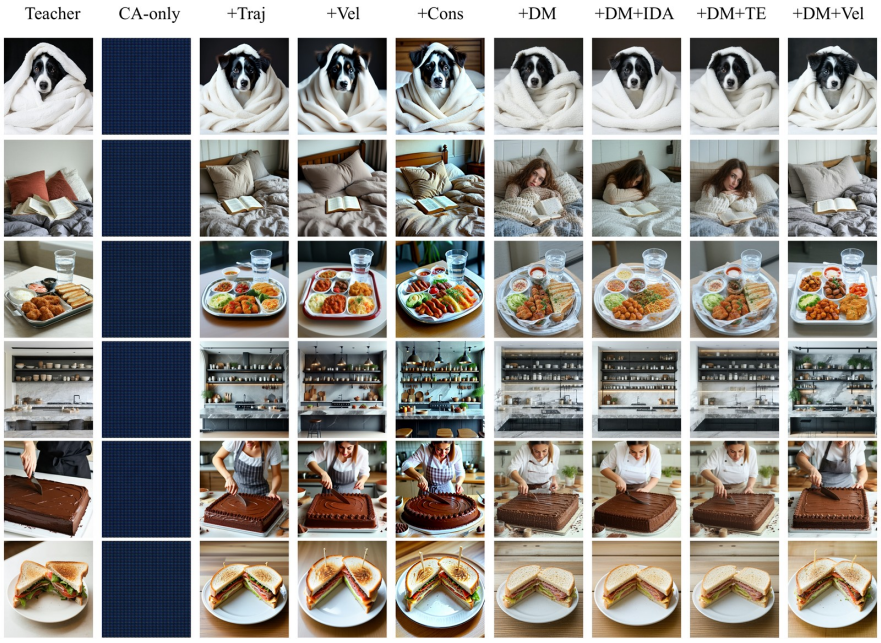


Figure 22: Qualitative comparison on SD3.5-L.



Figure 23: Qualitative comparison on FLUX.2-klein-base-4B.

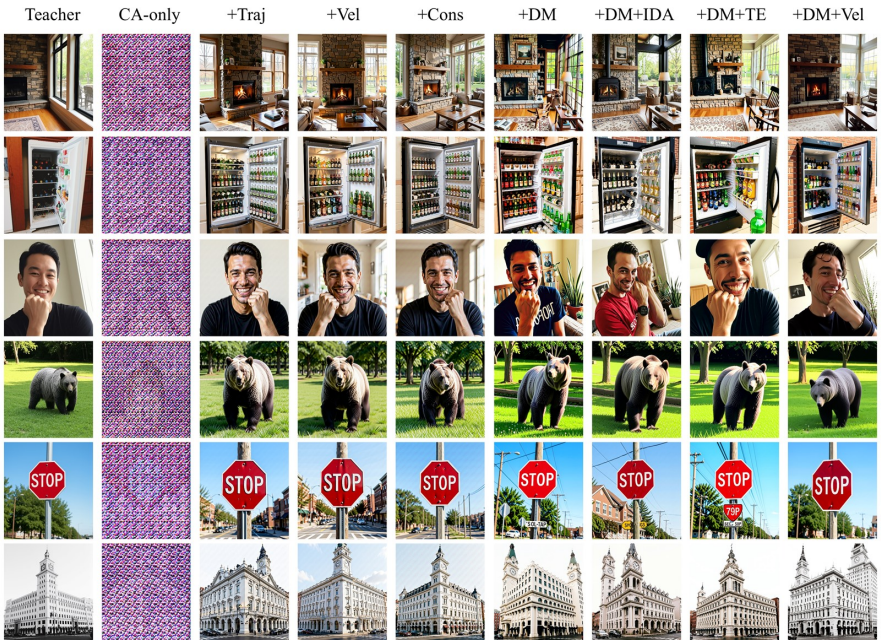


Figure 24: Qualitative comparison on FLUX.2-klein-base-9B.



Figure 25: Qualitative comparison of MSE and perceptual losses for transition anchoring on FLUX.2-klein-base-4B, with teacher and DM references. Red boxes indicate the enlarged regions shown below each image.

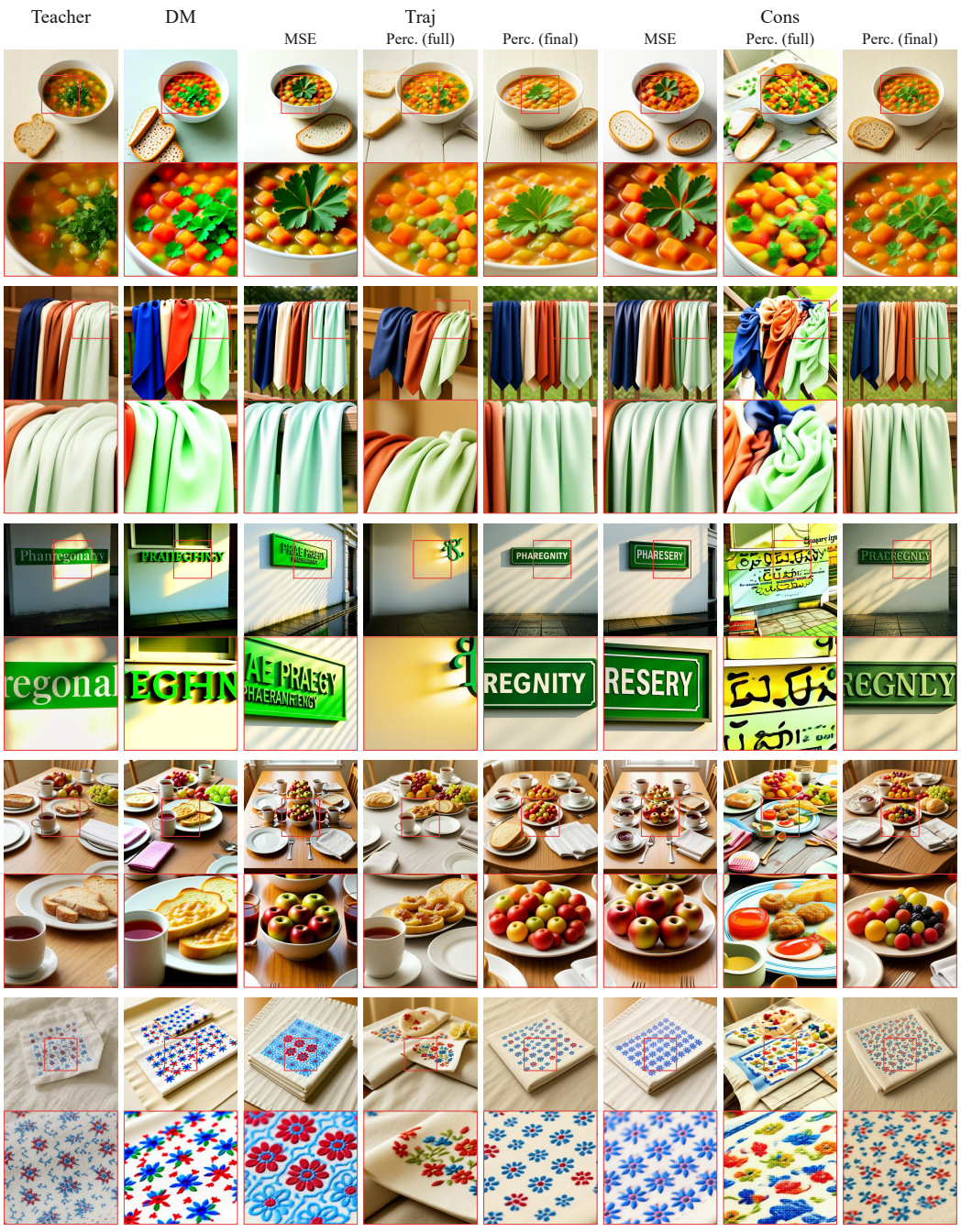


Figure 26: Additional qualitative comparisons of MSE and perceptual losses on FLUX.2-klein-base-4B, using the same layout as Figure 25.

Prompts for qualitative sweep over 4 models.

SD3.5-M.

1. A bowl of apples and bananas sits on a woven cloth.
2. They are all working in the kitchen together.
3. There is a sign on a building that says best buy.
4. A group of office workers posing for a picture together.
5. The frontal facade of an old Gothic styled building.
6. A plate full of food including broccoli, quinoa, black beans, and vegetables.

SD3.5-L.

1. A black and white dog wrapped in a white blanket.
2. A messy bed with a pillow and a book.
3. There is a tray full of food and a cup of water on the tray.
4. A kitchen with shelves and gray marble countertops.
5. A woman cutting up a large chocolate sheet cake.
6. A sandwich is cut in half on a white plate.

FLUX.2-klein-base-4B.

1. Two white couches and a brown table sit in a room.
2. A plate of food with bread, french fries, and meat on it.
3. A young girl is eating her meal with other plates of food present.
4. A tennis player bounds toward the end of the court after hitting the ball.
5. A building with spires stands in the background near a traffic light showing a digital man with his hands on his hips.
6. An open box with pizza on the ground.

FLUX.2-klein-base-9B.

1. Two white couches and a brown table sit in a room.
2. A refrigerator door left open that has a lot of bottles on the door.
3. A man is smiling with his fist aimed at his cheek.
4. A gray bear in the park with grass and trees at the back.
5. A stop sign is shown attached to a pole.
6. A large white building with a big clock tower at one corner.

Prompts for the perceptual-loss comparisons.

Figure 25.

1. Ripe red tomatoes arranged on a stainless-steel tray beside a folded beige cloth.
2. A line of footprints crossing an open snowfield toward a distant red shed.
3. A person walking through a snowy park under an overcast winter sky, with bare trees and footprints in the snow.
4. A minimalist coffee shop storefront with a large sign reading exactly “LATENT SPACE”, smaller text underneath reading “Coffee & Models”, realistic street photography.
5. A person standing inside a café at dusk, warm indoor lighting mixed with cool blue light from the street outside.

Figure 26.

1. A bowl of vegetable soup topped with parsley beside a slice of bread on a pale table.
2. Silk scarves in navy, cream, rust, and pale green draped over a wooden rail.
3. A green pharmacy sign casting light across a white wall and damp sidewalk.
4. A breakfast table set with plates, cups, fruit, bread, jam, and folded napkins.
5. An embroidered cloth with small blue and red motifs folded beside plain linen.