TEACHER AS REFERENCE: LEARNING REAL-DATA CORRECTIONS FOR FEW-STEP DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step distillation makes text-to-image generation efficient, but teacher matching alone does not specify how real observations should correct student predictions. We introduce *Reference-Informed Non-adversarial Few-step Distillation* (RINFD), which injects real-data information through two distinct training signals. Teacher-to-Data correction (T2D) learns teacher-referenced displacements from reconstruction–observation pairs on noisy bridge queries, then supplies prediction-level corrections at re-noised student outputs. Frequency-sensitive sliced Wasserstein distillation (FD-SWD) provides distribution-level supervision by aligning local, multi-scale statistics of predicted clean endpoints with real latents. Both signals act on the same clean prediction alongside teacher-guided distillation, without a discriminator. Controlled SDXL experiments support their complementarity: the default combination exceeds all tested FD-SWD-only weights on PickScore and ImageReward, and bridge sampling improves on reconstruction-only correction training. Weight sweeps reveal a balance between preference, color fidelity, and composition, with no single setting leading every metric. Across SDXL, SD3.5 Large, and FLUX.1-dev, the complete framework achieves the best average rank on two backbones and the second-best on the third among the evaluated four-step students, with the highest GenEval on both flow backbones.

1 INTRODUCTION

Diffusion models generate high-quality images but conventionally require many sequential network evaluations (Rombach et al., 2022; Black Forest Labs, 2024a). Few-step distillation reduces this cost by training a student to produce an image in only a few evaluations, enabling practical deployment in interactive and resource-constrained settings (Yin et al., 2024a; Ge et al., 2026). DMD distills a student model by estimating an online generator field, while ISG uses teacher dynamics to stabilize training (Yin et al., 2024a; Ge et al., 2026). Real data can provide additional supervision, while central design choice is how real data should influence the student beyond teacher matching. Adversarial distillation introduces this information through a learned discriminator (Yin et al., 2024b), whereas statistical objectives compare generated and real feature distributions (Starodubcev et al., 2026). These approaches motivate a complementary question: how can real data supply both corrections to individual predictions and constraints on their aggregate statistics within a non-adversarial framework?

We propose *Reference-Informed Non-adversarial Few-step Distillation* (RINFD), with two real-data supervision mechanisms. Teacher-to-Data correction (T2D) uses teacher reconstructions to construct paired displacements, learns a correction predictor on noisy bridge queries, and applies it to unpaired student predictions. Frequency-sensitive sliced Wasserstein distillation (FD-SWD) directly aligns local, multi-scale statistics of predicted clean endpoints with real latents. The first mechanism learns from reconstruction–observation correspondence; the second constrains aggregate output statistics. They act on the same clean prediction alongside unchanged DMD and ISG objectives, and all auxiliary modules are used only during training.

The two signals expose different information from real data, making their combination useful but their balance consequential. At four NFE, RINFD achieves the best average rank on SDXL and FLUX.1-dev and the second-best on SD3.5 Large. Controlled SDXL ablations (Sec. 5.3) show pref-

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.

erence gains beyond the tested FD-SWD-only weights and support bridge sampling for the learned correction. Weight sweeps identify different preference and color-fidelity operating points, while the combined objective incurs an SDXL compositional trade-off.

Our contributions are:

- **A framework for non-adversarial real-data supervision.** We combine learned prediction-level correction and direct distribution-level alignment within few-step distillation, with no inference-time auxiliary modules.
- **Two distinct supervision mechanisms.** T2D learns from teacher-constructed pairs through bridge regression; FD-SWD matches frequency-sensitive endpoint statistics. We characterize the correction's regression target and make explicit how the two signals enter the student update.
- **Evidence of complementarity and balance.** Three-backbone comparisons establish competitive four-step generation. Controlled SDXL component, query-design, and weight ablations demonstrate joint preference gains within the tested range and characterize their trade-off with compositional accuracy.

2 Related Work

Diffusion models. Diffusion models learn to reverse a corruption process through denoising or score estimation (Song & Ermon, 2019; Song et al., 2021b). DDIM introduces a deterministic, non-Markovian sampling process that enables faster generation with fewer denoising evaluations (Song et al., 2021a). Flow matching generalizes this perspective by learning vector fields along prescribed probability paths (Lipman et al., 2023), while rectified flow constructs straighter transport trajectories that are easier to integrate with coarse numerical solvers (Liu et al., 2022). These formulations have supported increasingly powerful text-to-image models with strong visual quality and compositional capabilities (Rombach et al., 2022; Podell et al., 2024; Labs et al., 2025).

Few-step distillation. Few-step distillation compresses the iterative sampling process of a pre-trained diffusion or flow model into a small number of function evaluations. Progressive distillation repeatedly halves the number of sampling steps by matching composed teacher updates (Salimans & Ho, 2022), whereas consistency-based methods enforce agreement among predictions along the same sampling trajectory (Song et al., 2023; Kim et al., 2024). Other approaches directly regress teacher outputs or distill multi-step transitions and trajectories (Luhman & Luhman, 2021; Luo et al., 2023; Zheng et al., 2024). DMD instead aligns teacher and generator distributions through their estimated score fields (Yin et al., 2024a). DMD2 removes the dependence on large precomputed regression datasets and incorporates real-data supervision through an adversarial objective (Yin et al., 2024b), while ADD similarly combines teacher-based distillation with adversarial learning (Sauer et al., 2024). Building on DMD2, Scale-wise Distillation (SwD) introduces patch-level MMD across feature scales to provide additional multi-scale distribution supervision (Starodubcev et al., 2026). SenseFlow extends distribution matching to large flow models using intra-segment guidance (Ge et al., 2026), while Decoupled DMD separates the effects of classifier-free guidance augmentation and distribution matching (Liu et al., 2025). Recent work has also explored teacher score discrepancies involving real and generated samples for few-step video generation (Chen et al., 2026).

Statistical distribution objectives. Statistical objectives provide direct statistical measures of distribution discrepancy without learning a real–fake discriminator. Maximum Mean Discrepancy (MMD) matches kernel mean embeddings, while sliced Wasserstein distance compares distributions through one-dimensional projections (Deshpande et al., 2018; Du et al., 2023). Frequency Distribution Loss further applies sliced Wasserstein matching to Fourier amplitude and phase statistics (Ni et al., 2024). Such objectives enable stable distribution matching over local, global, and frequency-domain representations and have also been incorporated into diffusion distillation, as exemplified by the patch-level MMD objective in SwD (Starodubcev et al., 2026).

3 PRELIMINARIES AND PROBLEM FORMULATION

Noising and clean predictions. For a clean latent $\boldsymbol{z}_0$, denote

$$\boldsymbol{z}_t = \alpha_t \boldsymbol{z}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{1}$$

For SDXL, $D_F^w(\boldsymbol{z}_t, t, c) = (\boldsymbol{z}_t - \sigma_t \epsilon_F^w(\boldsymbol{z}_t, t, c))/\alpha_t$ denotes a clean-latent prediction at guidance w. The frozen teacher is T. A few-step student G_θ produces $\boldsymbol{z}_0^G$, the clean endpoint predicted from the student state at a sampled coarse step and used by the training losses. For other prediction parameterizations, D_F^w denotes the corresponding clean-latent conversion.

Teacher-guided distillation. DMD2 uses an online generator-field estimate and a frozen teacher target (Yin et al., 2024a;b). For ideal conditional scores $s_p = \nabla_{\boldsymbol{z}_t} \log p_t(\boldsymbol{z}_t \mid c)$ and the CFG convention $s_T^w = s_T^c + (w-1)(s_T^c - s_T^u)$, its score-difference direction can be written as

$$s_G^c - s_T^w = (s_G^c - s_T^c) - (w-1)(s_T^c - s_T^u). \tag{2}$$

This separates conditional teacher matching from guidance augmentation, whose empirical roles are studied by Decoupled DMD (Liu et al., 2025). Both depend on the teacher; neither specifies a correction derived from a particular real observation. We retain the original DMD target, guidance, normalization, and timestep rule, together with ISG (Ge et al., 2026), and construct an additional source of real-data supervision.

Two forms of real-data supervision. We study how real observations can guide the student through learned corrections and direct statistical constraints. T2D constructs displacement targets from teacher reconstructions and learns to apply corrections at student queries. FD-SWD matches generated and real patch distributions. Both supply gradients to the same clean prediction, but use different information: paired correspondence and aggregate statistics. The next section defines their construction and joint optimization.

4 METHOD

RINFD augments teacher-guided distillation with two real-data signals at the student clean prediction: a learned correction and a direct statistical constraint. Fig. 1 shows their construction: panel (a) learns from reconstruction–observation pairs, panel (b) applies the correction at student queries, and panel (c) matches endpoint patch statistics. We first describe generator tracking, then the two supervision mechanisms and their joint optimization.

4.1 DYNAMIC GENERATOR TRACKING

DMD requires an estimate of the evolving generator field. We obtain this estimate by re-noising detached student predictions and training an adapter on a frozen backbone:

$$\widehat{v}_g = v_{\text{ref}} + \Delta_g^v, \qquad \mathcal{L}_{R_g} = \mathbb{E}\left[\left\|\widehat{v}_g(\boldsymbol{z}_t^G, t, c) - u(\boldsymbol{z}_0^G, \boldsymbol{\epsilon}, t)\right\|_2^2\right]. \tag{3}$$

Here $u = \boldsymbol{\epsilon}$ for epsilon prediction and $u = \boldsymbol{\epsilon} - \boldsymbol{z}_0^G$ for the linear flow-matching parameterization. For a fixed generator, the unrestricted population minimizer is the conditional prediction induced by p_t^G. Consequently, $\widehat{v}_g$ estimates the full generator field, whereas Δ_g^v represents its prediction-space residual relative to the frozen reference. The finite adapter approximates this field and is updated online to track the changing generator distribution.

4.2 TEACHER-CONSTRUCTED CORRECTION

The generator-field estimator supports teacher matching, but supplies no real-data signal. We obtain such targets by asking the teacher to reconstruct real latents. Given a real latent–caption pair $(\boldsymbol{z}_r, c)$, sample a source timestep s, Gaussian noise $\boldsymbol{\epsilon}_s$, and teacher guidance w, and construct

$$\boldsymbol{y}_s^r = \alpha_s \boldsymbol{z}_r + \sigma_s \boldsymbol{\epsilon}_s, \qquad \boldsymbol{z}_T = D_T^w(\boldsymbol{y}_s^r, s, c), \qquad \boldsymbol{\Delta} = \boldsymbol{z}_r - \boldsymbol{z}_T. \tag{4}$$

The reconstruction and its source observation define the training correspondence. Their displacement contains reconstruction uncertainty as well as systematic teacher error. Generated predictions

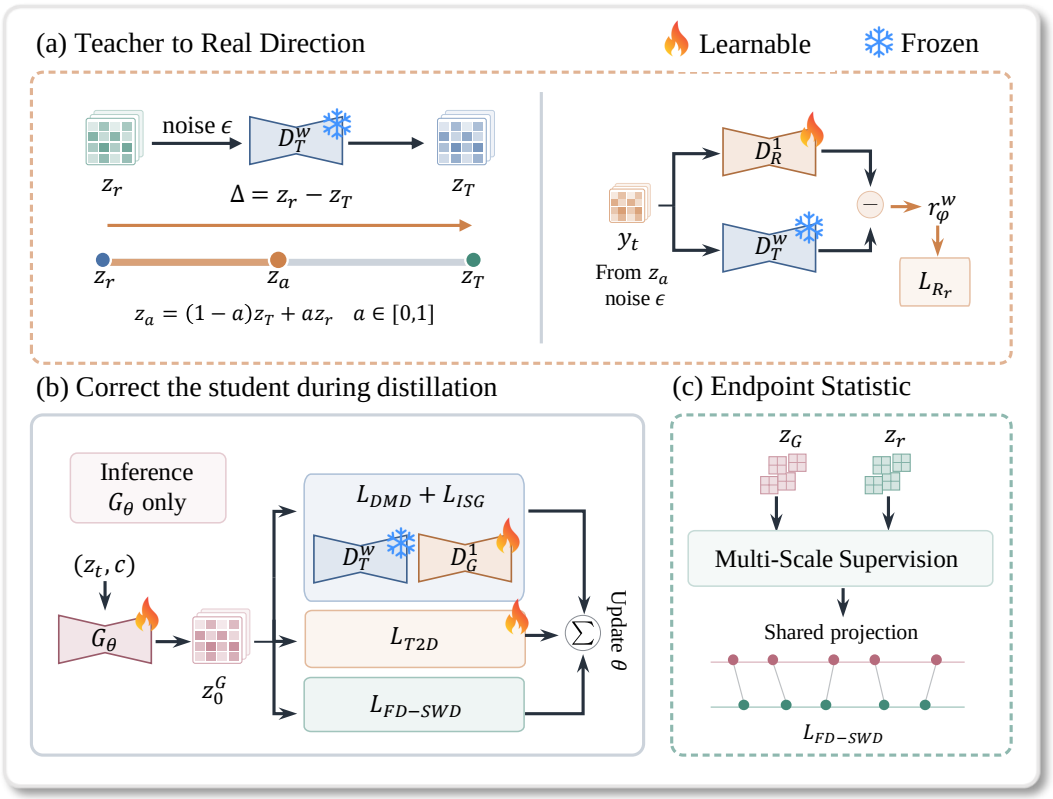


Figure 1: Overview of RINFD, a non-adversarial framework for few-step distillation. Built on teacher-guided DMD and ISG, the framework introduces two real-data supervision signals with different scopes: T2D learns a teacher-to-real displacement on bridge queries from reconstruction–observation pairs and transfers it to student predictions, while FD-SWD aligns predicted clean endpoints with real latents through local, multi-scale, and frequency-sensitive statistics. The generator-field estimator is updated online to track the evolving student distribution. All auxiliary modules are used only during training, introducing no inference-time overhead.

have no paired real target, so the displacement must be learned as a function of the query state. Training only at reconstructions would restrict the clean inputs to one endpoint of each pair. We therefore vary the regression queries by sampling a bridge coordinate a and forming

$$\boldsymbol{z}_a = \boldsymbol{z}_T + a\boldsymbol{\Delta}, \qquad \boldsymbol{y}_t = \alpha_t \boldsymbol{z}_a + \sigma_t \boldsymbol{\epsilon}_t, \tag{5}$$

where t and $\boldsymbol{\epsilon}_t$ are sampled independently of the source perturbation. The distribution of a mixes uniform sampling on $[0, 1]$ with optional endpoint samples. The same w is reused for source reconstruction and the bridge reference.

Let D_R^1 be the clean prediction of the frozen backbone with a trainable correction adapter. Let sg denote stop-gradient. Define

$$r_\psi^w(\boldsymbol{y}_t, t, c) = D_R^1(\boldsymbol{y}_t, t, c) - \mathrm{sg}(D_T^w(\boldsymbol{y}_t, t, c)). \tag{6}$$

Only the correction adapter is optimized with

$$\mathcal{L}_{R_r} = \mathbb{E}\left[\left\| r_\psi^w(\boldsymbol{y}_t, t, c) - \mathrm{sg}(\boldsymbol{\Delta}) \right\|_2^2\right], \tag{7}$$

The target remains $\boldsymbol{\Delta}$ at every a, including the real endpoint. Unlike R_g, this regression is defined by a fixed real-data distribution and frozen teacher, rather than the changing generator.

What the adapter learns. The population regression target depends on whether teacher guidance is fixed or randomly sampled.

Proposition 1 (Population target of bridge regression). *Let $X = (\boldsymbol{y}_t, t, c)$, $B_w(X) = D_T^w(X)$, and $A(X) = D_R^1(X)$. Assuming finite second moments, the squared-risk minimizer over unrestricted predictors satisfies*

$$A^*(X) = \mathbb{E}[B_w(X) + \boldsymbol{\Delta} \mid X]. \tag{8}$$

For fixed w, the optimal residual is $r^*(X) = \mathbb{E}[\boldsymbol{\Delta} \mid X]$. For random w not supplied to A, only its conditional average satisfies $\mathbb{E}[A^*(X) - B_w(X) \mid X] = \mathbb{E}[\boldsymbol{\Delta} \mid X]$.

Proposition 1 connects target construction to query design: even with identical paired displacements, changing the query law changes the conditional regression problem. Bridge sampling therefore determines where the predictor learns to interpret the correction targets. Fixed guidance yields the conditional-mean displacement; random guidance contributes the reference-averaging term derived in Appendix A.

4.3 Transferring the Correction to the Generator

T2D transfers the correction predictor to re-noised student predictions using the same teacher reference as in bridge regression. At a common query, the correction is the difference between the adapted prediction and the frozen teacher prediction:

$$r_\psi^w = D_R^1 - D_T^w. \tag{9}$$

The paired objective trains this prediction difference toward $\boldsymbol{\Delta}$; generator training reuses the learned predictor at student queries. For a student prediction $\boldsymbol{z}_0^G$, sample an independent probe timestep and noise, and evaluate the correction under stop-gradient

$$\overline{\boldsymbol{y}}_t^G = \alpha_t \,\mathrm{sg}(\boldsymbol{z}_0^G) + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{h} = \gamma r_\psi^w(\overline{\boldsymbol{y}}_t^G, t, c). \tag{10}$$

The teacher-guidance scalar is shared with DMD within the generator update; the probe timestep and noise are sampled independently. The adapter remains at guidance one, preserving the adapted-minus-reference convention used in regression. Here γ scales the displacement. We use the predicted displacement to form a detached regression target:

$$\mathcal{L}_{\mathrm{T2D}} = \left\| \boldsymbol{z}_0^G - \mathrm{sg}(\boldsymbol{z}_0^G + \boldsymbol{h}) \right\|_2^2. \tag{11}$$

The detached target gives $\nabla_{\boldsymbol{z}_0^G} \mathcal{L}_{\mathrm{T2D}} = -2\boldsymbol{h}$: the predicted correction directly specifies a clean-prediction gradient, with no differentiation through the correction adapter.

4.4 Frequency-Sensitive Endpoint Statistics

FD-SWD complements T2D by directly matching generated and real patch statistics. Both objectives act on $\boldsymbol{z}_0^G$, the clean endpoint predicted from an intermediate student state at the sampled coarse step, rather than exclusively on the final four-step output. Their distinction is the supervision supplied: a teacher-referenced correction versus a local statistical constraint. For a patch p at scale s, let $\mu(p)$ be its channel-wise spatial mean and $\widetilde{p} = p - \mu(p)$. We use three patch descriptors:

$$\phi_s^L(p) = \mu(p), \qquad \phi_s^M(p) = \mathrm{vec}(\mathrm{AvgPool}_2(\widetilde{p})), \qquad \phi_s^H(p) = \mathrm{vec}(\widetilde{p}). \tag{12}$$

The descriptors separate patch means from pooled and full-resolution spatial variation. Their frequency sensitivities overlap; in particular, ϕ^H retains all centered variation rather than an isolated high-frequency band. For each summary, generated and real feature sets are projected onto shared random unit directions. We sort each projected sample set and average the absolute differences between equal-rank values (Deshpande et al., 2018). The overall loss is

$$\mathcal{L}_{\text{FD-SWD}} = \sum_{s \in \mathcal{S}} a_s \sum_{b \in \{L,M,H\}} \lambda_b \,\mathrm{SWD}\left(\phi_s^b(\boldsymbol{z}_0^G), \phi_s^b(\boldsymbol{z}_0^r)\right). \tag{13}$$

FD-SWD constrains aggregate patch statistics, while T2D uses reconstruction–observation correspondence to learn a query-conditioned correction. Matching patch distributions alone does not determine these paired displacements. The two objectives thus supply distinct supervision to the same clean prediction; Section 5.3 tests their complementarity.

4.5 Optimization and Inference

The complete generator objective is

$$\mathcal{L}_G = \lambda_D \mathcal{L}_{\text{DMD}} + \lambda_I \mathcal{L}_{\text{ISG}} + \lambda_T \mathcal{L}_{\text{T2D}} + \lambda_S \mathcal{L}_{\text{FD-SWD}}. \tag{14}$$

At a clean prediction, the real-data contribution to the gradient is $-2\lambda_T \boldsymbol{h} + \lambda_S \nabla_{\boldsymbol{z}_0^G} \mathcal{L}_{\text{FD-SWD}}$. Thus, the learned correction and statistical matching enter the same student update through distinct signals, with their contributions regulated by λ_T and λ_S. Section 5.3 evaluates their combined benefits and trade-offs.

Training alternates generator updates, online generator-field regression, and real-pair correction regression, adding forward evaluations and adapter updates. Inference uses only G_θ on the original few-step grid, with no auxiliary estimator or added function evaluation.

5 Experiments

5.1 Experimental Setup

Datasets, Benchmark and Metrics. Following DMD2 (Yin et al., 2024b), we train on an aesthetically and resolution-filtered subset (aesthetic score 5.0, 1024 pixels at least) of LAION-5B (Schuhmann et al., 2022). We evaluate four-step students on COCO-5K (Lin et al., 2014) and GenEval (Ghosh et al., 2023). CLIP measures text–image alignment; HPSv2, PickScore, and ImageReward provide complementary learned preference signals; CFM (Fang et al., 2026) measures perceptual color fidelity; and GenEval evaluates compositional correctness. Since these metrics capture different aspects of generation quality, we do not treat any single metric as a universal measure.

Aggregate ranking. We summarize each backbone with the equal-weight mean rank over CLIP, HPSv2, PickScore, ImageReward, CFM, and GenEval, excluding the teacher. Ties receive the same competition rank. These correlated metrics are also reported individually; ranks summarize the measured trade-offs and do not establish statistical significance.

We evaluate three large-scale text-to-image backbones: FLUX.1 dev (Black Forest Labs, 2024a), Stable Diffusion 3.5 Large (Stability AI, 2024a), and SDXL (Podell et al., 2024), which span different model sizes and generative paradigms. Our students are trained on four NVIDIA H20 GPUs with a per-device batch size of one and two gradient-accumulation steps, giving a total batch size of eight, matching the setup used for the SenseFlow comparison. Appendix C gives the full protocol and reproducibility details.

5.2 Main Results

Comparison Baselines. For SDXL, we compare against LCM (Luo et al., 2023), Flash-SDXL (Chadebec et al., 2025), SDXL-Lightning (Lin et al., 2024), Hyper-SD (Ren et al., 2024), DMD2 (Yin et al., 2024b), SwD (Starodubcev et al., 2026) and SenseFlow (Ge et al., 2026). For SD 3.5 Large, we compare with SD 3.5 Large Turbo (Stability AI, 2024b), SwD and SenseFlow. For FLUX.1-dev, we compare with FLUX.1-schnell (Black Forest Labs, 2024b), Hyper-FLUX (Ren et al., 2024), FLUX-Turbo-Alpha (Alimama-Creative Team, 2024), SwD and SenseFlow. All baselines are evaluated under their official 4-step configurations.

Quantitative Results. As illustrated in Tab. 1, we compare RINFD with representative 4-step distillation methods on one diffusion backbone and two flow backbones. RINFD attains the best average rank on SDXL (2.17) and FLUX.1-dev (2.33), and the second-best on SD3.5 Large (2.00 versus 1.67 for SwD). It leads ImageReward on SDXL and FLUX.1 dev, and leads GenEval on both flow backbones, where its 0.7024 on FLUX.1-dev exceeds every four-step comparator as well as the 50-step teacher (0.6658). Notably, this is achieved without a discriminator, whereas the strongest single competitor per backbone varies.

Different Compositional Behavior. Aggregate GenEval separates the backbones. On SDXL, RINFD reaches 0.5551, above its teacher (0.5362) and SenseFlow (0.5463) but below SwD (0.5650) and DMD2 (0.5839); on SD3.5 Large and FLUX.1-dev it reaches 0.7336 and 0.7024, the best in both blocks. Appendix C.3 localizes the SDXL gap to a counting category. RINFD

Method	NFE	CLIP ↑	HPSv2 ↑	PickScore ↑	ImageReward ↑	CFM ↑	GenEval ↑	Avg. Rank ↓
Stable Diffusion XL								
SDXL	80	0.3398	0.2822	21.10	0.2858	13.03	0.5362	–
LCM-SDXL	4	0.3308	0.2730	20.84	0.1052	**13.23**	0.5002	6.17
Flash-SDXL	4	0.3236	0.2724	20.86	0.0856	13.08	0.4702	7.17
SDXL-Lightning	4	0.3314	0.2818	21.31	0.2284	12.91	0.5328	5.50
Hyper-SDXL	4	0.3353	<u>0.2888</u>	21.32	0.4367	12.72	0.5354	4.17
DMD2-SDXL	4	**0.3412**	0.2849	21.24	0.4533	12.34	**0.5839**	3.50
SwD-SDXL	4	<u>0.3398</u>	0.2824	<u>21.41</u>	<u>0.6116</u>	13.12	<u>0.5650</u>	<u>2.67</u>
SenseFlow	4	0.3394	0.2848	21.23	0.1741	12.94	0.5463	4.67
RINFD	4	0.3336	**0.2916**	**21.77**	**0.6517**	<u>13.15</u>	0.5551	**2.17**
Stable Diffusion 3.5 Large								
SD 3.5	80	0.3334	0.2901	21.71	0.9408	13.13	0.7069	–
SD 3.5 Turbo	4	0.3284	0.2826	21.57	0.7748	<u>13.14</u>	0.6845	3.67
SwD-SD 3.5	4	**0.3348**	<u>0.2911</u>	**21.78**	**1.0235**	13.06	<u>0.7215</u>	**1.67**
SenseFlow	4	<u>0.3329</u>	**0.2927**	21.69	0.9671	12.86	0.6988	2.67
RINFD	4	0.3318	0.2900	<u>21.73</u>	<u>0.9951</u>	**13.16**	**0.7336**	<u>2.00</u>
FLUX.1-dev								
FLUX.1-dev	50	0.3264	0.2937	22.02	0.9245	12.90	0.6658	–
FLUX.1-schnell	4	**0.3328**	0.2886	21.39	0.6952	12.83	0.6436	4.33
Hyper-FLUX	4	0.3262	0.2886	21.95	<u>0.9240</u>	12.92	0.6699	3.50
FLUX-Turbo-Alpha	4	0.3264	0.2843	21.85	0.8553	**13.08**	0.6162	4.67
SwD-FLUX	4	0.3275	**0.2940**	<u>21.96</u>	0.9195	9.75	<u>0.6737</u>	<u>2.83</u>
SenseFlow	4	0.3267	0.2935	**21.98**	0.8759	<u>12.97</u>	0.6420	3.17
RINFD	4	<u>0.3324</u>	<u>0.2939</u>	21.93	**0.9935**	12.91	**0.7024**	**2.33**

Table 1: Four-step comparisons on COCO-5K and GenEval. Teachers serve as references and are excluded from ranking. Avg. Rank averages ranks over all six metrics within each backbone block. Ties receive the same competition rank; for example, two students tied for second both receive rank 2, and the next receives rank 4. **Bold**/<u>Underline</u>: best/second best in distilling the same teacher.

leads all SDXL four-step students on two-object composition (78.54%) but reaches only 37.50% on counting, against 52.50% for SwD and 50.31% for DMD2; the remaining categories are within a few points of the strongest comparators. Counting is the largest SDXL deficit, rather than a uniform failure across categories. On FLUX.1-dev, RINFD attains 78.44% counting, second only to SenseFlow (80.00%) and above its own teacher (71.56%).

Qualitative Results Fig. 2 presents qualitative comparisons across a set of prompts. The displayed examples illustrate differences in lighting, structure, and prompt adherence, complementing the quantitative comparisons. Further examples appear in Appendix C.4.

5.3 Ablation Study

Tab. 2 examines how the two real-data signals contribute individually, combine, and respond to their relative weighting under a shared SDXL protocol. Bridge-versus-$a = 0$ comparisons further test query design within T2D under the same guidance convention. Results are point estimates for the reported runs.

Frequency-sensitive descriptors improve on plain patch matching. Replacing Plain-SWD with FD-SWD at matched patch scales improves PickScore ($21.43 \rightarrow 21.68$), ImageReward ($0.5386 \rightarrow 0.6042$), CFM ($12.96 \rightarrow 13.08$), and GenEval ($0.5642 \rightarrow 0.5732$), at an HPSv2 difference of 0.0003. Thus, at the default weight, separating patch means from pooled and full-resolution variation improves preference scores while preserving composition better than plain patch matching.

Both objectives contribute real-data supervision. At the default weights, FD-SWD improves all reported metrics over Base. T2D alone improves CLIP and the three preference proxies, with GenEval approximately unchanged ($0.5717 \rightarrow 0.5726$). Each signal therefore contributes useful real-data supervision, with the statistical objective improving a broader set of metrics in this comparison.

Combining the signals yields complementary preference gains. At $\lambda_S = 0.25$, adding T2D raises PickScore ($21.68 \rightarrow 21.77$), ImageReward ($0.6042 \rightarrow 0.6517$), and CFM ($13.08 \rightarrow 13.15$). Across



Figure 2: Qualitative comparisons of four-NFE students on SD3.5 Large and FLUX.1-dev. The examples compare prompt adherence, global composition, color rendition, and local texture across methods, complementing the quantitative evaluation.

Method	NFE	CLIP ↑	HPSv2 ↑	PickScore ↑	ImageReward ↑	CFM ↑	GenEval ↑
Component ablation							
Base (DMD + ISG)	4	0.3297	0.2800	21.27	0.4273	13.01	0.5717
Base + Plain-SWD	4	0.3326	**0.2934**	21.43	0.5386	12.96	0.5642
Base + FD-SWD ($\lambda_S = 0.25$)	4	0.3333	<u>0.2931</u>	<u>21.68</u>	0.6042	13.08	**0.5732**
Base + T2D (bridge, $\lambda_T = 0.25$)	4	0.3311	0.2877	21.38	0.4487	12.99	0.5726
Full ($\lambda_S = 0.25,\ \lambda_T = 0.25$)	4	0.3336	0.2916	**21.77**	<u>0.6517</u>	13.15	0.5551
Loss-weight sensitivity							
Base + FD-SWD ($\lambda_S = 0.125$)	4	0.3347	0.2897	21.33	0.5272	<u>13.16</u>	0.5586
Base + FD-SWD ($\lambda_S = 0.50$)	4	**0.3410**	0.2883	21.49	0.5985	13.01	<u>0.5729</u>
Full ($\lambda_S = 0.25,\ \lambda_T = 0.125$)	4	<u>0.3382</u>	0.2915	21.59	**0.6895**	13.14	0.5549
Full ($\lambda_S = 0.25,\ \lambda_T = 0.50$)	4	0.3359	0.2897	21.65	0.5664	**13.17**	0.5577
Bridge-sampling ablation							
Base + T2D ($a = 0$)	4	0.3333	0.2823	21.29	0.4726	13.07	0.5401
Base + T2D ($a = 0$) + FD-SWD	4	0.3351	0.2916	21.31	0.3163	13.07	0.5451

Table 2: SDXL component, loss-weight, and bridge-sampling ablations at four NFE. Weight sweeps vary λ_S without T2D, or λ_T within Full at fixed $\lambda_S = 0.25$. The last block uses reconstruction-only queries ($a = 0$), with $\lambda_T = 0.25$ and $\lambda_S = 0.25$ when present; other T2D rows use bridge sampling. All rows share the Base objective and training/evaluation protocol. **Bold**/<u>underline</u>: best/second-best distinct values per metric across all rows, with ties marked equally.

FD-SWD-only weights $\lambda_S \in \{0.125, 0.25, 0.50\}$, the highest PickScore (21.68) and ImageReward (0.6042) both remain below default Full. Full also exceeds this ImageReward maximum at $\lambda_T = 0.125$ (0.6895), so the gain appears at two tested correction strengths. Within this sweep, the results support complementary supervision beyond rescaling endpoint statistics.



Figure 3: Qualitative component ablation of RINFD on SDXL. All variants use matched prompts and random seeds. Combining all modules yields the strongest overall visual quality and aesthetic appeal, producing more polished images than the partial variants.

Balancing the signals selects different operating points. With $\lambda_S = 0.25$ fixed, reducing λ_T from 0.25 to 0.125 improves ImageReward to 0.6895, the best value in the table, while reducing PickScore from 21.77 to 21.59. Increasing λ_T to 0.50 achieves the highest CFM (13.17), with CLIP 0.3359, but lowers PickScore to 21.65 and ImageReward to 0.5664. The sweep reveals distinct operating points: $\lambda_T = 0.125$ favors ImageReward, 0.25 favors PickScore, and 0.50 favors CFM. The default $(\lambda_S, \lambda_T) = (0.25, 0.25)$ combines the highest PickScore with the second-highest ImageReward across the reported ablations.

The SDXL preference gains come with a compositional trade-off: default Full changes GenEval from FD-SWD-only's 0.5732 to 0.5551 and HPSv2 from 0.2931 to 0.2916. Full's GenEval spans 0.5549–0.5577 across the tested T2D weights, motivating structure-aware balancing. Appendix C.3 separately reports cross-method category results. These controlled SDXL comparisons assess component effects; the flow-backbone results establish the performance of the complete framework.

Bridge sampling improves correction transfer. Without FD-SWD, replacing reconstruction-only queries ($a = 0$) with bridge sampling improves HPSv2 ($0.2823 \rightarrow 0.2877$), PickScore ($21.29 \rightarrow 21.38$), and GenEval ($0.5401 \rightarrow 0.5726$). With FD-SWD, the contrast is larger: bridge training reaches ImageReward 0.6517 versus 0.3163, PickScore 21.77 versus 21.31, and GenEval 0.5551 versus 0.5451. With paired targets and guidance retained, this contrast supports the query-design rationale in Proposition 1. Appendix A.6 provides complementary held-out regression and student-query diagnostics.

6 Conclusion

We presented RINFD, which injects real-data supervision through teacher-referenced prediction correction and frequency-sensitive statistical alignment. Both signals guide the same student clean prediction without adversarial training or added inference cost. Across SDXL, SD3.5 Large, and FLUX.1-dev, the framework delivers competitive four-step generation. It achieves the best average rank on two backbones and the highest GenEval on both flow backbones among the evaluated students. Controlled SDXL ablations support complementary preference gains and reveal how the balance between the signals affects preference, color fidelity, and composition. These findings support combining the two forms of supervision and motivate structure-aware balancing.

AI USE DISCLOSURE

Generative AI systems, including OpenAI ChatGPT and Codex, were used as assistive tools for manuscript organization and language editing, LaTeX preparation, literature discovery, code inspection and debugging, and drafting and checking mathematical derivations. The human authors are responsible for the research decisions, training runs, data processing, evaluation, and scientific interpretation, and for verifying AI-assisted text, code, derivations, and citations before submission. No experimental measurement was generated or inferred by an AI system.

ETHICS STATEMENT

This work studies efficiency-oriented distillation of existing text-to-image models using existing training data and pretrained checkpoints. It does not introduce a new dataset, collect personal information, or involve a human-subject study; the reported human-preference metrics are automated predictors. The method nevertheless inherits the data provenance, representational biases, memorization risks, and safety limitations of its teachers and training corpus. More efficient generation may broaden beneficial access to generative models, but may also lower the cost of producing misleading, harmful, or infringing content. The proposed non-adversarial real-data correction does not constitute a safety mechanism and should not be interpreted as removing these risks.

REPRODUCIBILITY STATEMENT

Sections 3–4 define the distributions, objectives, gradient routes, and inference procedure. The appendix provides the bridge-regression proof, conditional query-transfer analysis, training procedure, reported hyperparameters and training-step budgets, dataset processing, and metric definitions. The appendix include configurations for the reported runs, together with the timestep distributions and weights of DMD and the paired T2D correction, random seeds, metric versions, and per-prompt evaluation outputs needed to regenerate the tables. All auxiliary branches are used only during training; the released inference configuration will contain only the few-step generator.

REFERENCES

Alimama-Creative Team. FLUX.1-Turbo Alpha. Hugging Face model card, 2024. URL https://huggingface.co/alimama-creative/FLUX.1-Turbo-Alpha. Official model card.

Black Forest Labs. FLUX.1-dev model card. Hugging Face model card, 2024a. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Official model card.

Black Forest Labs. FLUX.1-schnell model card. Hugging Face model card, 2024b. URL https://huggingface.co/black-forest-labs/FLUX.1-schnell. Official model card.

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

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. In *ICML Workshop*, 2026.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced wasserstein distance. In *CVPR*, 2018.

Chao Du, Tianbo Li, Tianyu Pang, Shuicheng Yan, and Min Lin. Nonparametric generative modeling with conditional sliced-Wasserstein flows. In *ICML*, 2023.

Zhengyao Fang, Zexi Jia, Yijia Zhong, Pengcheng Luo, Jinchao Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Too vivid to be real? benchmarking and calibrating generative color fidelity. In *CVPR*, 2026.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

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

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv: 2506.15742*, 2025.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv: 2402.13929*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *ECCV*, 2014.

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

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *NeurIPS Workshop*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, and Lin Ma. Misalignment-robust frequency distribution loss for image transformation. *arXiv preprint arXiv:2402.18192*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

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 *NeurIPS*, 2024.

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

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

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, volume 32, 2019.

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

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

Stability AI. Stable Diffusion 3.5 Large model card. Hugging Face model card, 2024a. URL `https://huggingface.co/stabilityai/stable-diffusion-3.5-large`. Official model card.

Stability AI. Stable Diffusion 3.5 Large Turbo model card. Hugging Face model card, 2024b. URL `https://huggingface.co/stabilityai/stable-diffusion-3.5-large-turbo`. Official model card.

Nikita Starodubcev, Ilya Drobyshevskiy, Denis Kuznedelev, Artem Babenko, and Dmitry Baranchuk. Scale-wise distillation of diffusion models. In *ICLR*, 2026.

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

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 *NeurIPS*, 2024b.

Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping. *arXiv preprint arXiv:2402.19159*, 2024.

A Regression Target and Transfer Analysis

The main construction raises two separate questions: what correction is identified by the paired objective, and under what conditions can it be queried at generated states? We first derive the regression target, then state a conditional transfer bound. The score identities at the end of this section clarify the distinction from direct field correction.

A.1 Proof of Proposition 1

Let $Y = B_w(X) + \mathbf{\Delta}$ and assume finite second moments. The identity

$$\mathbb{E}\|A(X) - Y\|^2 = \mathbb{E}\|A(X) - \mathbb{E}[Y \mid X]\|^2 + \mathbb{E}\|Y - \mathbb{E}[Y \mid X]\|^2 \tag{15}$$

follows by expanding the square and conditioning the cross term on X. The second term does not depend on A, giving Eq. 8. For fixed w, $B_w(X)$ is deterministic given X, so $A^* - B_w = \mathbb{E}[\mathbf{\Delta} \mid X]$. With random guidance, write $\overline{B}_\mu(X) = \mathbb{E}_\mu[B_w(X) \mid X]$ and $m_\mu(X) = \mathbb{E}_\mu[\mathbf{\Delta} \mid X]$ for the bridge training law μ. Then

$$A^*(X) - B_w(X) = m_\mu(X) + \overline{B}_\mu(X) - B_w(X). \tag{16}$$

The last term vanishes only after averaging under the training conditional law of $w \mid X$. Sharing the same marginal guidance range between training and injection need not preserve that conditional law when the query distribution changes. A finite LoRA regressor adds approximation and optimization error to this characterization.

A.2 A paired displacement prediction identity

For fixed guidance and the training query law, let $m(X) = \mathbb{E}[\mathbf{\Delta} \mid X]$. Then

$$\mathbb{E}\|\mathbf{\Delta} - \eta m(X)\|^2 = \mathbb{E}\|\mathbf{\Delta}\|^2 - \eta(2-\eta)\mathbb{E}\|m(X)\|^2. \tag{17}$$

This follows from $\mathbb{E}\langle \mathbf{\Delta}, m(X)\rangle = \mathbb{E}\|m(X)\|^2$. For $0 < \eta < 2$, the paired displacement prediction error decreases whenever m is nonzero. This is a regression result under the training query law; it does not establish improvement of generator quality or convergence to the real-data distribution.

Finite-predictor diagnostic. The same comparison can be made for a fitted predictor without assuming population optimality. Under a fixed paired-query law, let $M = \mathbb{E}\|\mathbf{\Delta}\|^2 > 0$, $E_r = \mathbb{E}\|\mathbf{\Delta} - r\|^2$, and $Q = \mathbb{E}\|r\|^2$, all finite. Expanding the square gives, for $0 < \eta \le 1$,

$$\mathbb{E}\|\mathbf{\Delta} - \eta r\|^2 = (1-\eta)M + \eta E_r - \eta(1-\eta)Q. \tag{18}$$

Hence $E_r/M < 1$ is sufficient for lower paired-displacement prediction error than zero correction for every such η. The identity also holds for finite-sample averages, directly linking this criterion to the NMSE diagnostic in Appendix A.6. It concerns displacement prediction on the same paired queries, not distance from a bridge state to the real endpoint or improvement of generator quality. NMSE above one fails this sufficient condition but does not establish that every scaled prediction is harmful; η here is an analysis variable, not a newly introduced training schedule.

A.3 Coverage requirement for generator queries

Let μ_X be the bridge query distribution and ν_X the generator probe distribution at a particular training update. For fixed guidance, suppose ν_X is absolutely continuous with respect to μ_X and $d\nu_X/d\mu_X \le \kappa$. With $r = A - B$ and the same reference at training and injection,

$$\mathbb{E}_{\nu_X}\|r - m_\mu\|^2 \le \kappa\, \mathbb{E}_{\mu_X}\|r - m_\mu\|^2 = \kappa\big(\mathcal{R}_\mu(A) - \mathcal{R}_\mu(A^*)\big), \tag{19}$$

where $\mathcal{R}_\mu$ is the unnormalized squared regression risk. The inequality follows by a change of measure and the equality from Eq. 15. The bound separates regression error from query coverage. The experiments do not establish a finite global κ: Gaussian noising alone is insufficient, and the generator query law changes during training. Moreover, the quantity being transferred is the training conditional-mean correction, whose usefulness for generator updates must be evaluated separately.

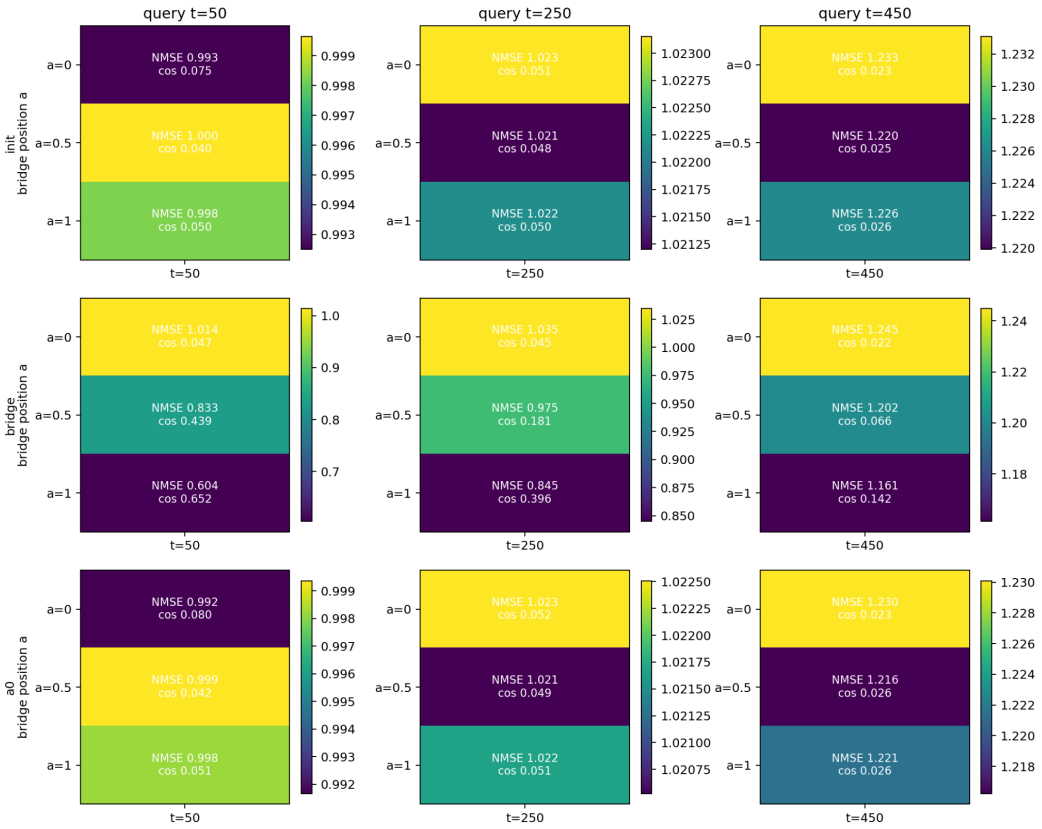


Figure 4: Held-out paired-displacement regression across bridge positions and query timesteps. Rows compare the correction adapter at initialization (top), bridge-trained adapter (middle), and reconstruction-only adapter trained at $a = 0$ (bottom). Each cell reports NMSE and cosine similarity to the target displacement.

For random guidance, Eq. 16 introduces an additional reference term and possible conditional-law shift.

Bridge sampling broadens the collection of training states between paired endpoints, but its benefit depends on how those states overlap with generator probes. This motivates the source-only versus bridge experiment. The bridge coordinate is not an input to the implemented adapter; diffusion query time t does not substitute for that coordinate. Furthermore, the target is the same full displacement $\mathbf{\Delta}$ at every bridge position, including $a = 1$. It is not a remaining displacement $(1 - a)\mathbf{\Delta}$ and need not vanish at the real endpoint. No attraction-to-manifold equilibrium is implied.

A.4 Score identities and the implemented target

At a common state, condition, and noising convention, ideal scores satisfy

$$s_G - s_{\text{data}} = (s_G - s_T) + (s_T - s_{\text{data}}). \tag{20}$$

This algebraic identity makes explicit that teacher matching and data matching are different objectives. Likewise, $s_p - s_q = \nabla \log(p/q)$ is the gradient of an ideal density ratio. Neither identity identifies the paired clean-latent displacement in Eq. 4 with a data score. The source of that pair is a teacher reconstruction of a noised real observation, not an independent sample from the teacher's generation distribution. Consequently, our interpolation does not by itself define transport from the teacher distribution to the data distribution.

A.5 GUIDANCE DEPENDENCE AND MATCHED-QUERY COMPARISON

For a common query $X = (\boldsymbol{y}_t, t, c)$, define the learned increment $\delta_\psi(X) = D^1_R(X) - D^1_T(X)$ and the reference difference $g_w(X) = D^1_T(X) - D^w_T(X)$. Adding and subtracting $D^1_T(X)$ in Eq. 9 yields the equivalent decomposition $r^w_\psi = \delta_\psi + g_w$. The paired loss trains their sum toward $\boldsymbol{\Delta}$. At fixed guidance, Proposition 1 gives $\delta^*(X) = m_\mu(X) - g_w(X)$ and therefore $r^*(X) = m_\mu(X)$. Thus, the reference difference is compensated in the unrestricted population optimum, rather than remaining an unavoidable extra bias. With random unobserved guidance, $r^*(X, w) = m_\mu(X) + \overline{B}_\mu(X) - B_w(X)$ retains the reference-averaging term; finite fitting and query shift add further limitations. The implemented T2D signal is the complete residual, whose learned and reference contributions are not separately quantified by the experiments.

For bridge-trained and reconstruction-only adapters evaluated at identical X and w, the common reference cancels exactly:

$$r^w_{\text{bridge}}(X) - r^w_{a=0}(X) = D^1_{R,\text{bridge}}(X) - D^1_{R,a=0}(X) = \delta_{\text{bridge}}(X) - \delta_{a=0}(X). \tag{21}$$

This identity makes the matched-query comparison informative: the explicit reference term is shared, so the residual difference reflects the learned predictors. The generation ablation then tests the downstream effect of this query-design choice under a common guidance convention. Student states evolve separately in those training runs, so the comparison evaluates the complete T2D procedure.

The evaluation follows this distinction: held-out paired regression measures target prediction, shared student queries measure response magnitude, and component ablations measure generation quality. Residual norms summarize the complete correction; their differences are not the norm of the predictor difference in Eq. 21.

A.6 EMPIRICAL DIAGNOSTICS OF BRIDGE CORRECTION TRANSFER

The transfer analysis in Eq. 19 separates learning the paired correction from evaluating it on student queries. We examine these two aspects with complementary diagnostics: held-out displacement regression (Experiment A) and correction magnitudes on student queries (Experiment B). Both diagnostics use the trained adapters at the SDXL 4.5k checkpoint, corresponding to $4{,}500$ optimizer updates for each of R_g and R_r and 900 generator updates; the initialization baseline in Experiment A is evaluated before training. Experiment A uses 256 held-out reconstruction–observation pairs, and Experiment B uses 64 student samples. In Experiment B, the compared adapters are evaluated on the same fixed student-query set.

Experiment A: held-out displacement regression. We compare the correction adapter at initialization, the bridge-trained adapter, and an adapter trained only at the reconstruction endpoint ($a = 0$). For held-out reconstruction–observation pairs, we retain the full target $\boldsymbol{\Delta} = \boldsymbol{z}_r - \boldsymbol{z}_T$ at each bridge position $a \in \{0, 0.5, 1\}$ and query timestep $t \in \{50, 250, 450\}$. We report normalized mean-squared error

$$\text{NMSE} = \frac{\sum_i \|r_\psi(X_i) - \boldsymbol{\Delta}_i\|_2^2}{\sum_i \|\boldsymbol{\Delta}_i\|_2^2} \tag{22}$$

and cosine similarity to $\boldsymbol{\Delta}_i$. NMSE measures prediction error relative to target energy, whereas cosine similarity measures directional alignment.

Fig. 4 shows that bridge training improves regression away from the reconstruction endpoint. At $t = 50$, the bridge-trained adapter attains NMSE 0.833 and cosine similarity 0.439 at $a = 0.5$, improving to 0.604 and 0.652 at $a = 1$. At $t = 250$, the real-endpoint query still gives NMSE 0.845 and cosine similarity 0.396. In contrast, reconstruction-only training remains close to initialization, with weak directional alignment. The improvement is not uniform: alignment remains weak at $a = 0$, and at $t = 450$ even the best bridge position has NMSE 1.161, above the zero-correction baseline of one. The low-noise interior and real-endpoint results meet the finite-predictor criterion in Eq. 18; the high-noise results do not. Thus, query broadening improves displacement prediction in identifiable regimes, while leaving its high-noise reliability unresolved.

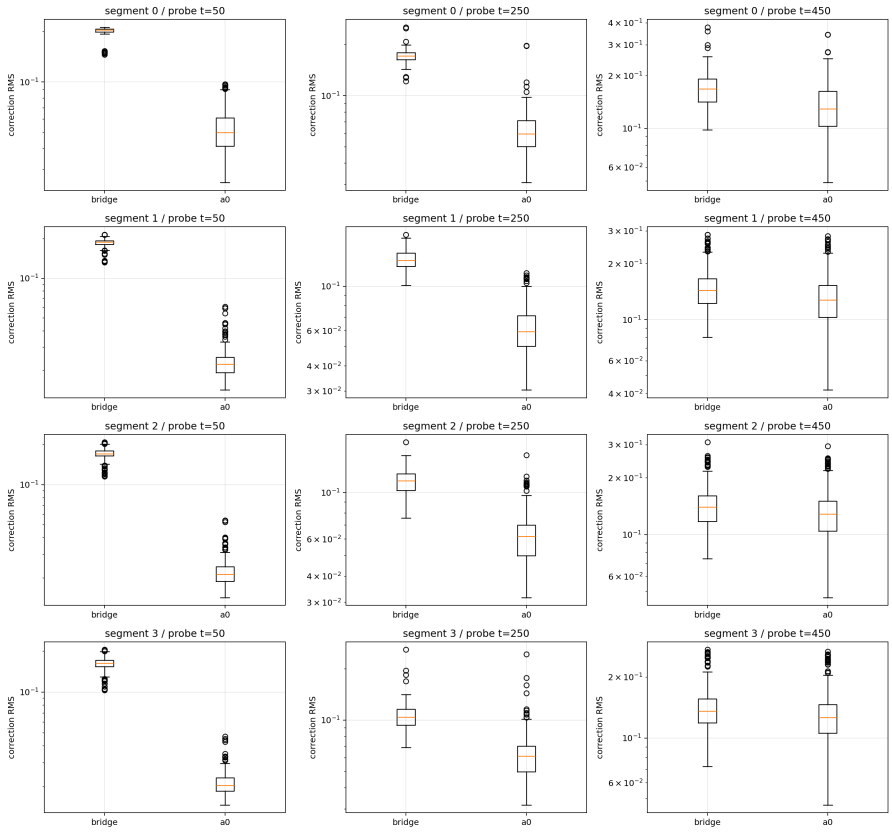


Figure 5: Correction RMS on student queries for bridge-trained and reconstruction-only ($a = 0$) adapters. Rows correspond to student segments 0–3; columns use probe timesteps 50, 250, and 450. Bridge training produces higher median correction magnitudes throughout, with the clearest separation at low query noise. The logarithmic vertical scales differ across panels. RMS characterizes response magnitude, not alignment with a ground-truth correction.

Experiment B: correction magnitude on student queries. We next examine the bridge-trained and reconstruction-only adapters on noised clean predictions from the four student segments. These are student-query evaluations, distinct from the paired queries in Experiment A. For a correction tensor with d entries, we summarize its magnitude by

$$\mathrm{RMS}(r_\psi) = \sqrt{\frac{1}{d}\|r_\psi\|_2^2}. \tag{23}$$

Fig. 5 reports the distributions separately for segments 0–3 and probe timesteps $t \in \{50, 250, 450\}$. Bridge training yields a higher median correction RMS in all twelve settings. The separation is most pronounced at $t = 50$, remains visible at $t = 250$, and narrows substantially at $t = 450$, where the distributions overlap. The effect therefore extends across the sampled student segments rather than appearing only on paired bridge queries. This is evidence of a stronger correction response on student inputs, not a measurement of its directional accuracy; larger RMS alone is not a quality criterion.

Joint interpretation and scope. Experiment A establishes improved paired-displacement prediction in specific query regimes, while Experiment B shows that bridge training also changes the correction response on student inputs. The generation ablation in Tab. 2 supplies the downstream evidence: with FD-SWD held fixed, replacing reconstruction-only training with bridge sampling increases ImageReward from 0.3163 to 0.6517 and PickScore from 21.31 to 21.77. The three experiments provide complementary evidence for query design: improved prediction on held-out pairs, a changed response on shared student inputs, and better downstream generation under the same guid-

ance convention. The formal coverage condition and guidance dependence remain as specified in Appendices A and A.5.

B Implementation of the Paired Correction

The implementation follows the paired objective in Section 4. Its active target and injection use the full displacement. Teacher guidance is sampled by the DMD guidance function (Ge et al., 2026), and the correction adapter is evaluated at guidance one.

Quantity	Reported settings
T2D source and query timesteps	uniform integers in $[20, 500]$
DMD timesteps	uniform integers in $[20, 980]$
Correction-adapter guidance	1
Bridge coordinate	$(1-p)\,\mathcal{U}[0,1] + \frac{p}{2}\delta_0 + \frac{p}{2}\delta_1$, $p = 0.5$
T2D weight warmup	300 completed correction-adapter updates
FD-SWD patch sizes / scale weights	$(8, 4)$ / $(1.0, 0.3)$
FD-SWD low / mid / high proxy weights	0.5 / 0.7 / 1.0
FD-SWD num projections	1024
λ_D (DMD)	1.0 for all backbones
λ_I (ISG)	0.2 for SDXL; 1.0 for SD3.5 Large and FLUX.1-dev
λ_{T2D}	0.25 for SDXL, 0.5 for SD 3.5 L and FLUX.1 dev
λ_{FD-SWD}	0.25 for SDXL, 0.4 for SD 3.5 L and FLUX.1 dev

Table 3: Configuration used for the reported RINFD.

Training uses batches of generator states and independent batches of real observations. Generator updates minimize the weighted sum of DMD, ISG, T2D, and FD-SWD evaluated on the student clean prediction. For every generator optimizer update, R_g receives five optimizer updates on detached generated predictions and R_r receives five optimizer updates on teacher-constructed real pairs. Optimizer-step accounting is detailed in Appendix C.2. Warmup is $\lambda_T = \lambda_T^{\max} \min(1, n_r/U)$, where n_r counts correction-adapter optimizer steps and U is the warmup length. The weight reaches its maximum after warmup and is not subsequently annealed.

C Experimental Details

C.1 Details of Datasets

Training uses an aesthetically and resolution-filtered subset of LAION-5B: image–text pairs have an aesthetic score of at least 5.0 and a minimum side length of 1024 pixels. COCO-5K uses the 5,000 images from the COCO 2017 validation split; for each image, we select the annotation with the highest CLIP ViT-B/32 similarity. GenEval follows its official prompts and category definitions. The recorded COCO evaluation uses 1024×1024 images and seed 42, with one image per selected caption. Students use their four-step configurations; full-step teachers use the NFE listed in the table. The intended within-backbone protocol aligns prompt order, preprocessing, seed lists, and decoder settings.

C.2 Implementation Details

Hardware and initialization. Training uses four NVIDIA H20 96GB GPUs, a per-device batch size of one, and two gradient-accumulation steps, giving an effective batch size of eight. The generator and both adapter branches, R_g and R_r, are initialized from the original pretrained teacher for each backbone. We train rank-64 LoRA parameters and use Fully Sharded Data Parallel (FSDP) for large-model training.

Optimizer and learning rates. All three branches use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Within each backbone, the generator, R_g, and R_r use the same learning rate: 1×10^{-4} for SDXL, 2×10^{-5} for SD3.5 Large, and 5×10^{-5} for FLUX.1-dev.

Update schedule and step accounting. We use a TTUR ratio of five on all three backbones: for every generator optimizer update, R_g and R_r each receive five optimizer updates. The reported training-step budgets count completed optimizer updates of each refiner branch, rather than generator updates or gradient-accumulation microsteps. These budgets are 4.5k for SDXL, 27k for SD3.5 Large, and 12k for FLUX.1-dev. In particular, the SDXL 4.5k checkpoint contains $4{,}500$ updates of R_g, $4{,}500$ updates of R_r, and 900 actual generator updates. The two refiner counts are separate; 4.5k is not their combined update count.

Generator loss weights. In Eq. 14, $\lambda_D = 1.0$ for all three backbones. The ISG weight is $\lambda_I = 0.2$ for SDXL and $\lambda_I = 1.0$ for SD3.5 Large and FLUX.1-dev. The T2D and FD-SWD weights and the correction warmup are listed in Tab. 3.

Timestep settings. Following Ge et al. (2026), we adopt different coarse timestep schedules depending on the model architecture. For SDXL, we follow the 1000-step discrete DDPM schedule used in DMD2 (Yin et al., 2024b), selecting step indices $\{249, 499, 749, 999\}$. For SD 3.5 Large, we switch to continuous timestep values $\{0.25, 0.5, 0.75, 1.0\}$, which are more suitable for flow-based models. In the case of FLUX.1 dev, which adopts a shifted σ inference strategy, we directly use the corresponding sigmas $\{0.512, 0.759, 0.904, 1.0\}$ as coarse anchors.

C.3 Per-category GenEval results

Tab. 4 expands the aggregate GenEval scores in Tab. 1. On SDXL, RINFD leads all four-step students on two-object composition (78.54%), while counting is its largest category deficit relative to SwD and DMD2 (37.50% versus 52.50% and 50.31%). These are cross-method comparisons, distinct from the aggregate component trade-off in Tab. 2. On SD3.5 Large and FLUX.1-dev, the complete framework attains the best overall scores (0.7336 and 0.7024); FLUX.1-dev counting reaches 78.44%, above its teacher (71.56%). These results characterize the full models and do not isolate component effects on the flow backbones.

Method	NFE ↓	Overall ↑	Single Obj. ↑	Two Obj. ↑	Counting ↑	Colors ↑	Position ↑	Color Attr. ↑
Stable Diffusion XL								
SDXL	80	0.5362	96.88%	70.96%	36.25%	85.64%	11.00%	21.00%
LCM-SDXL	4	0.5002	99.38%	55.05%	37.81%	85.11%	7.25%	15.50%
Flash-SDXL	4	0.4702	98.44%	50.00%	34.69%	79.52%	5.75%	13.75%
SDXL-Lightning	4	0.5328	98.12%	60.86%	44.38%	84.04%	9.75%	22.50%
Hyper-SDXL	4	0.5354	98.44%	65.15%	38.12%	**87.77%**	**12.50%**	19.25%
DMD2-SDXL	4	**0.5839**	**100.00%**	76.26%	50.31%	87.50%	9.75%	**26.50%**
SwD-SDXL	4	0.5650	99.06%	72.73%	**52.50%**	82.98%	10.25%	21.50%
SenseFlow	4	0.5463	99.38%	69.95%	41.56%	85.90%	6.75%	24.25%
RINFD	4	0.5551	99.69%	**78.54%**	37.50%	84.31%	8.75%	24.25%
Stable Diffusion 3.5 Large								
SD3.5 Large	80	0.7069	98.75%	90.66%	72.19%	79.79%	24.75%	58.00%
SD3.5 Large Turbo	4	0.6845	99.06%	88.13%	67.81%	77.93%	22.50%	55.25%
SwD-SD3.5	4	0.7215	99.06%	**93.94%**	**69.69%**	82.71%	27.00%	60.50%
SenseFlow	4	0.6988	**99.69%**	92.68%	65.00%	82.18%	22.25%	57.50%
RINFD	4	**0.7336**	99.38%	93.69%	67.50%	**85.11%**	**28.00%**	**66.50%**
FLUX.1-dev								
FLUX.1-dev	50	0.6658	98.12%	81.31%	71.56%	78.46%	22.50%	47.50%
FLUX.1-schnell	4	0.6436	**100.00%**	84.85%	57.50%	79.79%	20.00%	44.00%
Hyper-FLUX	4	0.6699	99.38%	79.55%	75.94%	76.60%	20.75%	49.75%
FLUX-Turbo-Alpha	4	0.6162	98.44%	72.98%	67.81%	75.27%	20.50%	34.75%
SwD-FLUX	4	0.6737	99.69%	84.85%	52.81%	**81.12%**	**28.75%**	**57.00%**
SenseFlow	4	0.6420	99.06%	71.21%	**80.00%**	77.93%	16.50%	40.50%
RINFD	4	**0.7024**	99.69%	**87.63%**	78.44%	77.93%	25.25%	52.50%

Table 4: GenEval benchmark results. We report overall score and per-attribute pass rates (%). **Bold**/Underline: best/second best in distilling the same teacher.

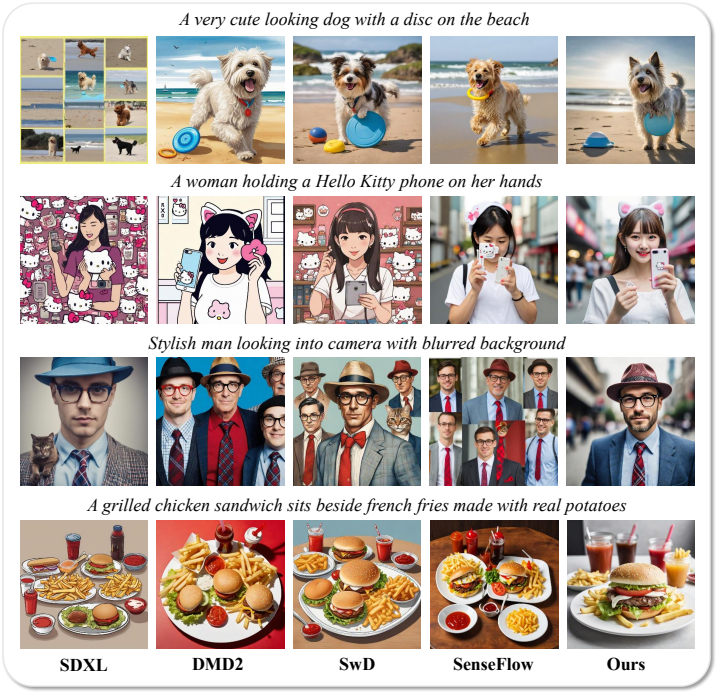


Figure 6: SDXL qualitative comparison with the full-step teacher, DMD2, SwD, and SenseFlow at four NFE, using a shared prompt and seed per row. Differences are most visible in color saturation and local contrast, consistent with the CFM and preference results in Tab. 1.



Figure 7: Additional qualitative comparison with the full-step teacher, SwD, and SenseFlow at four NFE on SD35 L and FLUX.1 dev.

C.4 Additional Visualizations

Fig. 6 and Fig. 7 compare RINFD with the full-step teacher and three four-step baselines on SDXL, SD 3.5 Large and FLUX.1 dev. All images come from the checkpoints reported in Tab. 1.

D Limitations

Teacher reconstructions can change content and their residuals can contain ordinary denoising uncertainty, not only systematic teacher bias. Squared regression learns an average correction and can suppress ambiguous detail. Training on bridges does not guarantee coverage of generator queries, particularly under changing generator distributions or guidance-dependent conditional shifts. The correction need not vanish on real samples and is not a proved manifold projection. A detached correction target does not guarantee compatible parameter-space updates across objectives. FD-SWD measures overlapping local descriptors with finite patches and projections. It does not enforce all spatial, conditional, or cross-descriptor dependencies. Real-data supervision can inherit unwanted training-data attributes, and neither objective guarantees improved aesthetics.

On the two flow backbones, some outputs appear softer in fine texture than those of adversarially trained students such as SwD, despite competitive aggregate metrics. Conditional-mean regression may average uncertain displacement components, while local statistical matching does not fully constrain instance-level detail. These are possible explanations, not a demonstrated frequency-specific mechanism or a causal attribution to removing GAN supervision. The controlled component evidence is limited to SDXL; the flow-backbone comparisons evaluate only the complete framework. Reported differences are point estimates without training-seed uncertainty estimates. Separate guidance and learned-increment effects, and their robustness across seeds, remain unquantified.

E Data, Models, and Release

The work uses LAION-5B, COCO 2017, GenEval, and pretrained text-to-image models under their respective access conditions and licenses. The intended anonymous release comprises training and evaluation code, resolved configurations, checkpoint provenance, and reproduction instructions. Source images and third-party weights are not redistributed. Any trained-checkpoint release remains subject to the applicable model and data conditions.