DISPLACEMENT MATCHING: ONE OBJECTIVE FOR REWARD ALIGNMENT AND DISTILLATION IN FEW-STEP GENERATORS

Anonymous authors
Paper under double-blind review



Figure 1: Qualitative results on SD3.5-M using DM and other methods.

ABSTRACT

Few-step generators synthesize images with a handful of deterministic updates, but post-training them faces two difficulties: the deployed sampler lacks a stochastic transition likelihood for standard policy gradients, and reward alignment and distillation lack a unified formulation. We take a distributional view, expressing both forms of supervision as multiplicative tilts of the generator's terminal distribution. Forward noising associates clean outputs with noisy observations, yielding a clean-sample posterior at each noise level. The terminal tilt shifts these posterior means, defining a displacement in the student's clean-prediction space. We introduce *Displacement Matching* (DM), which learns this posterior displacement as an additive prediction update, unifying pure distillation, sequential reward alignment, and joint training in one regression. DM estimates the generator's posterior mean separately from re-noised terminal samples, since a few-step head represents a finite-interval jump and need not equal that mean. The resulting regression accommodates black-box rewards without requiring trajectory likelihoods or differentiation through the rollout. Our population analysis connects these updates to deterministic deployment, establishing first-order reward improvement up to an explicit finite-step residual and a unique teacher fixed point for raw distillation. At matched reward-query budgets, four-step DM improves alignment and distillation quality while preserving fidelity, outperforming SOTA methods with 0.362 vs. 0.345 HPS on SD3.5-M and 0.93 vs. 0.81 GenEval on Z-Image.

1 INTRODUCTION

Few-step distilled generators now match the quality of strong multi-step teachers in only a handful of deterministic sampling steps (Song et al., 2023; Yin et al., 2024b;a; Geng et al., 2025). However,

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.

reward alignment of these generators still faces two main problems. First, few-step generators lack a learning interface that is both faithful to deployment and practical. Their deterministic transitions provide no stochastic reverse-process likelihood for standard policy gradients. Auxiliary stochastic constructions supply a likelihood, but their relationship to deterministic deployment needs separate justification (Huang et al., 2026a; Li et al., 2026). Reward backpropagation differentiates the deployed rollout directly, but requires differentiable rewards and costly, potentially unstable differentiation through the composed sampler (Clark et al., 2023; Prabhudesai et al., 2024). Second, reward alignment and distillation lack a unified formulation. Existing joint methods derive the two mechanisms separately and combine them afterwards (Jiang et al., 2026), leaving their integration dependent on separate learning objectives and update rules.

We take a distributional view of both problems. The deployed generator induces a distribution over clean outputs, which we call its terminal distribution. Reward alignment changes this distribution by assigning more weight to higher-reward samples. Distillation instead reweights it toward the teacher distribution, using the teacher-to-student density ratio when the teacher law is absolutely continuous with respect to the student law. Both are instances of a multiplicative *tilt*, providing a common description of the desired distributional change. This distributional view leaves two questions: how to turn the desired change into a practical training objective, and whether optimizing that objective produces the intended change in the deployed generator's output distribution.

Our key observation is that forward noising turns this terminal tilt into a local prediction target. Applying a fixed noising kernel to terminal samples defines a joint distribution of clean and noisy samples, and conditioning on a noisy input yields a posterior over clean samples. Tilting the terminal distribution reweights this posterior and shifts its mean, producing a displacement in the student's clean-prediction space. We introduce *Displacement Matching* (DM) to learn this *posterior displacement* as an additive update to the student's prediction. Reward alignment and distillation share this regression objective but use different displacement estimators. For reward alignment, the first-order displacement is proportional to the conditional covariance between advantage and the clean sample, estimated using an advantage-weighted posterior residual. For distillation, the displacement is the posterior-mean difference $\mu_T - \mu_\theta$.

The subtle choice is where the student's posterior mean comes from. Prior forward-process methods identify it with the student's own prediction (Zheng et al., 2025). A few-step head, however, represents a jump across a finite interval, so its prediction need not equal the posterior mean of its own terminal distribution. DM therefore estimates this mean separately from re-noised terminal samples, while using the frozen student prediction as the baseline for the displacement update.

The remaining question is whether fitting these local targets produces the intended terminal change when the sampling steps are composed. We analyze it under two population assumptions. **(P1) Forward consistency** matches the distribution of re-noised training queries to that of the states visited by the sampler, so prediction updates are learned where they will be applied. **(P2*) Flow consistency** models each deployed transition as a finite-interval probability-flow map, allowing us to trace these prediction updates through successive sampling steps. Under these assumptions, we characterize the first-order reward response with an explicit $O(\eta^2) + O(\eta\varepsilon_K)$ residual, where ε_K measures the finite-step approximation gap. We also establish the teacher as the unique fixed point of raw population distillation.

Across four-step SD3.5-M (Esser et al., 2024) and Z-Image (Z-Image Team, 2025), DM improves reward alignment and distillation quality while preserving fidelity. Posterior-replacement ablations test the separate estimator, and local-response diagnostics measure the change realized by the deployed sampler. Our contributions are:

1. We translate terminal distribution tilts into posterior displacements through forward noising, providing a learning interface without reverse-process likelihoods (Section 2.2).
2. We derive a first-order unbiased reward estimator and an exact teacher displacement in common clean-prediction coordinates, enabling pure distillation, sequential alignment, and joint training with one regression objective (Section 2.3).
3. We connect the updates to the deployed sampler with an explicit finite-step residual and a unique teacher fixed point (theorem 1 and proposition 1), and validate posterior displacement across distillation recipes and model scales (Section 3).

2 METHOD

2.1 DEPLOYED STUDENT AND TERMINAL DISTRIBUTION

Let c be the condition and $q_t(x_t \mid x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 I)$ the Gaussian forward kernel. Fix $1 = t_0 > \cdots > t_K = 0$, with $(\alpha_{t_0}, \sigma_{t_0}) = (0, 1)$ and $(\alpha_{t_K}, \sigma_{t_K}) = (1, 0)$. A deterministic student $f_\theta(x_t, t, c) \in \mathcal{X}_0$ parameterizes each transition through a clean prediction. Starting from $x^\theta_{t_0} = Z \sim \mathcal{N}(0, I)$, its sampler uses

$$\widehat{x}^\theta_{0,k} = f_\theta(x^\theta_{t_k}, t_k, c), \quad \widehat{\epsilon}^\theta_k = \frac{x^\theta_{t_k} - \alpha_{t_k}\widehat{x}^\theta_{0,k}}{\sigma_{t_k}}, \quad x^\theta_{t_{k+1}} = F_{\theta,k}(x^\theta_{t_k}, c) := \alpha_{t_{k+1}}\widehat{x}^\theta_{0,k} + \sigma_{t_{k+1}}\widehat{\epsilon}^\theta_k. \tag{1}$$

The output $X^\theta_0 := x^\theta_{t_K}$ induces the terminal law $p_{\theta,0}(\cdot \mid c) := \mathrm{Law}(X^\theta_0 \mid c)$. We seek the increment of f_θ that realizes a prescribed change of this law, using terminal samples alone and without differentiating through the rollout.

2.2 FROM HEAD INCREMENTS TO POSTERIOR DISPLACEMENTS

Our goal is to change the student's terminal law through an update of its head. Under the flow-map structure of Section 2.4, $f_\theta = f_{p_{\theta,0}}$, where f_p is the head determined by the probability flow of p. A target law $p^r_{\theta,0}$ therefore requires the head increment $f_{p^r_{\theta,0}} - f_{p_{\theta,0}}$. Reward alignment and distillation specify this target through a nonnegative tilt r:

$$p^r_{\theta,0}(x_0 \mid c) = \frac{r(x_0, c)\, p_{\theta,0}(x_0 \mid c)}{Z_r(c)}, \qquad Z_r(c) = \mathbb{E}_{p_{\theta,0}}[r(X_0, c)], \quad 0 < Z_r(c) < \infty. \tag{2}$$

Here $r = 1 + \eta A$ gives a local reward alignment tilt, while $r = p_{T,0}/p_{\theta,0}$ specifies distillation. The tilt determines the pointwise density change $p^r_{\theta,0} - p_{\theta,0} = (r/Z_r - 1)p_{\theta,0}$. We now need to express this change as an update at each noisy input x_t for f.

Our key observation is that forward noising represents this distributional change through posterior-mean displacements. The forward noising kernel $q_t(x_t \mid x_0)$ associates each clean law with a conditional distribution over X_0 given $X_t = x_t$. For the current law, Bayes' rule gives

$$p_\theta(x_0 \mid x_t, t, c) = \frac{q_t(x_t \mid x_0)\, p_{\theta,0}(x_0 \mid c)}{p_{\theta,t}(x_t \mid c)}. \tag{3}$$

Here $p_{\theta,t}$ denotes the noisy marginal of the current law. For any clean law p, define its posterior mean in the head's clean-prediction coordinates:

$$\mu_p(x_t, t, c) := \mathbb{E}_p[X_0 \mid X_t = x_t, c]. \tag{4}$$

Subscripts θ, T, D select the current, teacher, and distilled-anchor laws. At schedule levels, write $\mu_{p,k}(x_{t_k}, c) := \mu_p(x_{t_k}, t_k, c)$. Because the forward kernel is unchanged, the tilted posterior is proportional to $r(x_0, c)p_\theta(x_0 \mid x_t, t, c)$. Its mean $\mu_r := \mu_{p^r_{\theta,0}}$ is therefore

$$\mu_r(x_t, t, c) = \frac{\mathbb{E}[r(X_0, c)\, X_0 \mid X_t = x_t, c]}{\mathbb{E}[r(X_0, c) \mid X_t = x_t, c]}, \tag{5}$$

where expectations are taken under the current posterior distribution in Eq. (3). The distributional change is thus represented at each noisy input by

$$\Delta_r(x_t, t, c) := \mu_r(x_t, t, c) - \mu_\theta(x_t, t, c). \tag{6}$$

This representation is exact at the level of fields. For $\alpha_t, \sigma_t > 0$, Tweedie's identity gives

$$\nabla_{x_t} \log \frac{p^r_{\theta,t}(x_t \mid c)}{p_{\theta,t}(x_t \mid c)} = \frac{\alpha_t}{\sigma_t^2}\Delta_r(x_t, t, c), \tag{7}$$

where $p^r_{\theta,t}$ is the noisy marginal of the target law. With the current law fixed, the full displacement field determines the target noisy law by normalization. Gaussian noising is injective when $\alpha_t > 0$, so it also determines the target clean law. Thus $\Delta_r(\cdot, t, c)$ is an **equivalent** representation of the terminal distributional change. This field can be estimated from re-noised terminal samples without evaluating either terminal density (Section 2.3).

To obtain the head update, we must account for its finite sampling interval. A posterior mean is defined at one noise level, and its mismatch with the four-step head is substantial (Fig. 4). Under the flow-map structure, a small tilt changes the head through a transported direction averaged over that interval. DM approximates this increment by Δ_r at the starting level and targets $f_\theta + \Delta_r$, with a finite-step substitution error bounded in Section 2.4. This preserves the learned sampling transition when $r = 1$, even when $f_\theta \neq \mu_\theta$.

This leads to a regression objective for the head increment. Within each round, we freeze the anchor θ and train a copy θ^+. Given queries $\widetilde{X}_{t_k} \sim p_{\theta,t_k}$ and a displacement estimator $\widehat{\Delta}_k$, Displacement Matching minimizes the objective

$$\mathcal{L}(\theta^+) = \sum_{k=0}^{K-1} \mathbb{E}\left[\left\| \Big(f_{\theta^+}(\widetilde{X}_{t_k}, t_k, c) - f_\theta(\widetilde{X}_{t_k}, t_k, c) \Big) - \widehat{\Delta}_k \right\|_2^2\right]. \tag{8}$$

Queries, anchor predictions, and targets are detached. Squared-error regression learns the target's conditional mean. Its unrestricted population minimizer is

$$f_{\theta^+}(x, t_k, c) = f_\theta(x, t_k, c) + \mathbb{E}[\widehat{\Delta}_k \mid \widetilde{X}_{t_k} = x, c], \qquad p_{\theta,t_k}\text{-a.e. } x. \tag{9}$$

Individual targets need not equal Δ_r. It suffices to match their conditional mean: $\mathbb{E}[\widehat{\Delta}_k \mid \widetilde{X}_{t_k} = x, c] = \Delta_r(x, t_k, c)$, with equality up to $O(\eta^2)$ for reward alignment. The next section constructs estimators with this property.

2.3 Displacement Estimators and Training Modes

To construct such estimators, we generate X_0^- with the anchor sampler and re-noise it independently at each level:

$$\widetilde{X}_{t_k} = \alpha_{t_k} X_0^- + \sigma_{t_k} \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, I), \quad k = 0, \ldots, K-1. \tag{10}$$

Conditioning these pairs on $\widetilde{X}_{t_k}$ gives the posterior in Eq. (3). A visited rollout state instead determines its terminal output. Its clean-sample posterior is a point mass.

Lemma 1 (Displacement estimators)**.** *With exact posterior means, the following targets satisfy the conditional-mean requirement of Eq.* (9)*.*

For scalar rewards, take a bounded advantage A and $r_\eta = 1 + \eta A \geq 0$ with sufficiently small $\eta > 0$. Then we have

$$\widehat{\Delta}_{R,k} = \eta A(X_0^-, c)\big(X_0^- - \mu_{\theta,k}(\widetilde{X}_{t_k}, c)\big), \tag{11}$$

$$\mathbb{E}[\widehat{\Delta}_{R,k} \mid \widetilde{X}_{t_k} = x, c] = \Delta_{r_\eta}(x, t_k, c) + O(\eta^2). \tag{12}$$

The conditional mean is exactly $\eta \operatorname{Cov}(A, X_0 \mid X_t = x, c)$. It matches the first-order term of Δ_{r_η} from Eq. (5)*. No tilted posterior mean is fitted.*

For vector supervision, an available target posterior mean $\mu_{r,k}$ supplies the exact displacement:

$$\widehat{\Delta}_{V,k} = \mu_{r,k}(\widetilde{X}_{t_k}, c) - \mu_{\theta,k}(\widetilde{X}_{t_k}, c) = \Delta_r(\widetilde{X}_{t_k}, t_k, c). \tag{13}$$

Set $\widehat{\Delta}_k = \widehat{\Delta}_{R,k}$ or $\widehat{\Delta}_{V,k}$ in Eq. (8). The population optimum is $f_\theta + \Delta_{r_\eta} + O(\eta^2)$ for the reward branch and $f_\theta + \Delta_r$ exactly for the vector branch. In practice, we fit a DMD-style fake-score denoiser on re-noised anchor outputs (Yin et al., 2024b) and apply Tweedie's identity to estimate μ_θ. The anchor head f_θ remains the regression baseline.

These scalar and vector estimators can be used separately or combined in Eq. (8), yielding the following training modes.

For distillation, the diffusion teacher supplies μ_T. When $p_{T,0} \ll p_{\theta,0}$, $r_D = p_{T,0}/p_{\theta,0}$ gives the vector target without evaluating the density ratio:

$$\widehat{\Delta}_{D,k} = \mu_{T,k}(\widetilde{X}_{t_k}, c) - \mu_{\theta,k}(\widetilde{X}_{t_k}, c). \tag{14}$$

For reward alignment after distillation, use the scalar target relative to the distilled anchor θ_D:

$$\widehat{\Delta}_{R|D,k} = \eta A(X_0^-, c)\big(X_0^- - \mu_{D,k}(\widetilde{X}_{t_k}, c)\big), \qquad X_0^- \sim p_{\theta_D,0}(\cdot \mid c). \tag{15}$$

Joint training adds both targets in the same clean-prediction coordinates:

$$\widehat{\Delta}_{\mathrm{joint},k} = \eta A(X_0^-, c)\big(X_0^- - \mu_{\theta,k}(\widetilde{X}_{t_k}, c)\big) + \eta\big(\mu_{T,k}(\widetilde{X}_{t_k}, c) - \mu_{\theta,k}(\widetilde{X}_{t_k}, c)\big). \tag{16}$$

Both terms are scaled by η and derive from their respective terminal tilts. The shared regression specifies their joint update.

2.4 Why the Update Moves the Deployed Law

We now ask whether learning $f_\theta + \Delta_r$ produces the intended terminal change. The regression fits local prediction updates. The deployed output depends on their composition across all sampling steps. This requires matching training queries to deployed states and bounding the gap between posterior and head responses. We analyze both under two properties of the flow-map structure in the realizable population limit.

(P1) Forward consistency. The probability flow transports the forward marginals of the terminal law. Thus $\mathrm{Law}(x^\theta_{t_k} \mid c) = p_{\theta,t_k}(\cdot \mid c)$ for all k, so re-noised queries have the deployed-state marginals.

(P2*) Flow consistency. Let w_p be the probability-flow field of the corruption family of p. Write $\Psi^w_{t\to t'}$ for its solution map. Each deployed transition realizes a finite-interval map:

$$F_{\theta,k}(x,c) = \Psi^{w_\theta}_{t_k\to t_{k+1}}(x), \qquad w_\theta = w_{p_{\theta,0}}, \quad p_{\theta,t_k}\text{-a.e. } x, \quad k < K. \tag{17}$$

This is the flow-map object targeted by MeanFlow (Geng et al., 2025), consistency models (Song et al., 2023), and trajectory distillation (Salimans & Ho, 2022; Kim et al., 2024). It constrains the transition map without identifying the head with a velocity or a posterior mean.

Under (P2*), the head's exact response averages a transported direction over each interval. DM uses the posterior displacement at the interval's start. Appendix B defines their root-mean-square gap ε_K. Under transport regularity, $\varepsilon_K = O(h_K)$, where $h_K := \max_k(t_k - t_{k+1})$. Proposition 2 tracks this substitution through the composed sampler and yields the following result.

Theorem 1 (First-order realization of a reward tilt)**.** *Assume* (P1) *and* (P2*)*, let A be bounded, and let ε_K be the finite-step gap of Appendix B. Under its regularity conditions, if θ^+ realizes*

$$f_{\theta^+}(x,t_k,c) = f_\theta(x,t_k,c) + \Delta_{r_\eta}(x,t_k,c), \qquad p_{\theta,t_k}\text{-a.e. } x, \quad k < K, \tag{18}$$

then, for bounded Lipschitz φ,

$$\mathbb{E}_{p_{\theta^+,0}}[\varphi(X_0)] = \mathbb{E}_{p^{r_\eta}_{\theta,0}}[\varphi(X_0)] + O_\varphi(\eta^2) + O_\varphi(\eta\varepsilon_K), \tag{19}$$

and, when A is also in the test class,

$$\mathbb{E}_{p_{\theta^+,0}}[A] = \mathbb{E}_{p_{\theta,0}}[A] + \eta\,\mathrm{Var}_{p_{\theta,0}}(A) + O_A(\eta^2) + O_A(\eta\varepsilon_K). \tag{20}$$

Remark 1 (From estimation to realization)**.** *The population reward update of Section 2.3 satisfies Eq.* (18) *up to $O(\eta^2)$, which is absorbed into the second-order remainder of theorem 1.*

Distillation has an exact fixed-point characterization. Its update vanishes when the posterior-mean fields agree. Their equivalence to the clean law in Section 2.2 yields the following result.

Proposition 1 (Teacher fixed point)**.** *In the population ideal, the raw distillation branch is stationary at any level with $\alpha_{t_k} > 0$ if and only if $p_{\theta,0} = p_{T,0}$. The teacher is therefore its unique fixed point.*

Theorem 1 characterizes the local first-order response, not monotone improvement throughout training. Appendix C analyzes the idealized iteration in which each tilt is realized exactly. Our experiments evaluate DM in practical few-step settings.

2.5 Practical Implementation

For a group $\mathcal{G}$ and scalar signal u^i, use

$$\mathcal{C}_\mathcal{G}[u^i] := \mathrm{clip}\left(\frac{u^i - \bar{u}_\mathcal{G}}{s_\mathcal{G}}\right), \qquad \bar{u}_\mathcal{G} = |\mathcal{G}|^{-1}\sum_{j\in\mathcal{G}} u^j, \tag{21}$$

with $s_\mathcal{G} > 0$ and optional clipping. Rewards are processed before entering the scalar residual product. The vector branch is uncentered by default. Centering also performs well (Appendix E).

The anchor follows the trainable copy by $\theta \leftarrow \tau\theta + (1-\tau)\theta^+$. A slowly increasing τ keeps samples, posterior estimates, and targets relative to a quasi-static base, and $\tau \to 1$ recovers the frozen-anchor analysis of Section 2.4.

Table 1: Pure distillation on SD3.5-M (512^2). Students use $K = 4$ CFG-free steps; the teacher uses 50 steps with CFG 4.5.

Method	Steps	GenEval ↑	DPG-Bench ↑
Teacher (CFG 4.5)	50	0.65	84.05
PCM (Wang et al., 2024)	4	0.63	82.94
DMD (Yin et al., 2024b)	4	0.63	82.63
TDM (Luo et al., 2025)	4	<u>0.64</u>	<u>83.52</u>
Ours	4	**0.66**	**84.90**

3 EXPERIMENTS

We evaluate pure distillation, sequential reward alignment after distillation, and joint training with the common loss Eq. (8); only the displacement estimator changes. We then test transfer to a 6B backbone and ablate the main design choices.

3.1 SETUP

Models and training. All experiments use 512^2 resolution, a fixed $K = 4$ schedule, and CFG-free student sampling. Student and posterior-mean denoiser are trained with LoRA adapters ($\alpha = r = 64$). SD3.5-Medium (Esser et al., 2024) is distilled from a 50-step, CFG-4.5 teacher; the 6B Z-Image (Z-Image Team, 2025) from a 50-step, CFG-4 teacher. We use group size $|\mathcal{G}| = 24$, step size $\eta = 0.15$, EMA τ warmed up from 0 to 0.5. All methods share the same training schedule.

Rewards and prompts. We use HPSv2.1 (Wu et al., 2023), CLIPScore (Hessel et al., 2021), PickScore (Kirstain et al., 2023) for SD3.5 and additionally use GenEval (Ghosh et al., 2023) for Z-Image. We train 1,000 iterations following training prompts of DiffusionNFT (Zheng et al., 2025) and RTDMD (Huang et al., 2026a). We use the same training schedule for all methods on SD3.5-M and Z-Image.

Evaluation and baselines. Distillation uses GenEval and DPG-Bench (Hu et al., 2024). Reward alignment additionally reports all training rewards, held-out ImageReward (Xu et al., 2023) and Aesthetic v2 (Schuhmann, 2022), and DPG-Bench as a fidelity sentinel. We compare with DMD (Yin et al., 2024b), TDM (Luo et al., 2025), and PCM (Wang et al., 2024) for distillation. We organize RL baselines by training regime. FlowGRPO (Liu et al., 2025), AdvantageFlow(1.1) (Kveton et al., 2026) and DiffusionNFT (Zheng et al., 2025) are sequential baselines (initialized from the same flow-distilled checkpoint as our sequential flow-base variant). TDM-R1 (Luo et al., 2026), GDMD (Dong et al., 2026), DMDR (Jiang et al., 2026), and RTDMD (Huang et al., 2026a) represent joint pipelines. These methods keep distillation active during reward optimization.

3.2 PURE DISTILLATION

Table 1 evaluates the vector branch in Eq. (14). Our four-step CFG-free student outperforms prior four-step methods and even slightly exceeds the 50-step teacher on both metrics. This supports the interpretation of distillation as a teacher-directed posterior displacement optimized by the same regression as reward alignment. Proposition 1 concerns distributional matching rather than metric dominance, so small automatic-metric gains over the teacher are not contradictory.

3.3 REWARD ALIGNMENT AFTER DISTILLATION AND JOINT TRAINING

Table 2 compares reward alignment at a matched number of reward queries. Our sequential mode starts from either our flow-distilled student or a GAN-distilled base. The latter does not guarantee the flow-map assumption. Both variants improve the trained and held-out metrics while preserving DPG. Their similar results show that the scalar branch is not tied to one distillation recipe. After reward alignment, the flow-distilled initialization is stronger on CLIP, PickScore, GenEval, ImageReward, and DPG. The GAN initialization is slightly stronger on HPS and Aesthetic.

Table 2: Reward alignment on SD3.5-M at a matched number of reward queries. GenEval, IR, and Aes are held-out metrics, and DPG measures fidelity.

Method	HPS	CLIP	Pick	GenEval	IR	Aes	DPG
Base model							
SD3.5-M	0.284	0.280	21.84	0.65	1.01	5.94	84.05
Distilled starting points							
Ours, flow-distilled base	0.285	0.277	21.82	0.66	1.11	6.12	84.90
GAN-distilled base	0.289	0.273	21.94	0.65	1.09	6.08	83.81
RL after distillation							
FlowGRPO (Liu et al., 2025)	0.303	0.288	21.91	0.66	1.28	5.76	83.10
DiffusionNFT (Zheng et al., 2025)	0.314	<u>0.292</u>	22.23	0.65	1.32	5.73	83.54
AdvantageFlow (1.1) (Kveton et al., 2026)	0.303	0.281	22.22	0.66	1.34	5.87	83.83
Joint distillation and reward alignment							
TDM-R1 (Luo et al., 2026)	0.342	0.292	22.20	<u>0.67</u>	1.36	6.02	<u>84.12</u>
GDMD (Dong et al., 2026)	0.344	0.289	22.51	0.65	1.34	6.21	83.92
MeanFlowNFT (Huang et al., 2026b)	0.327	0.285	22.50	0.66	1.36	6.27	84.07
DMDR (Jiang et al., 2026)	0.336	0.290	22.64	0.65	1.30	<u>6.32</u>	83.81
RTDMD (Huang et al., 2026a)	<u>0.345</u>	0.282	<u>22.68</u>	0.66	<u>1.39</u>	6.31	84.07
Displacement Matching							
Ours, seq. (flow base)	0.358	0.302	23.63	**0.69**	1.52	6.48	85.77
Ours, seq. (GAN base)	0.360	0.301	23.58	0.68	1.51	6.51	85.53
Ours, joint	**0.362**	**0.304**	**23.66**	**0.69**	**1.53**	**6.52**	**85.84**

Table 3: Joint reward alignment on 6B Z-Image. HPS, CLIP, Pick, GenEval are training rewards; IR and Aes are held-out metrics

Method	HPS	CLIP	Pick	GenEval	IR	Aes	DPG
Base model							
Z-Image	0.289	0.288	21.52	0.73	0.93	5.79	86.92
Joint distillation and reward alignment							
Z-Image-Turbo	0.302	0.278	21.84	0.76	1.07	5.84	85.57
TDM-R1 (Luo et al., 2026)	0.338	0.288	22.79	<u>0.81</u>	1.43	6.27	87.77
GDMD (Dong et al., 2026)	0.321	0.283	22.63	0.78	1.41	6.29	87.54
MeanFlowNFT (Huang et al., 2026b)	0.321	0.285	22.74	0.79	1.42	6.27	87.43
DMDR (Jiang et al., 2026)	0.339	0.281	22.53	0.75	1.39	6.25	87.31
RTDMD (Huang et al., 2026a)	<u>0.342</u>	<u>0.287</u>	<u>22.82</u>	0.79	<u>1.44</u>	<u>6.32</u>	<u>87.82</u>
Displacement Matching							
Ours, joint	**0.351**	**0.294**	**23.17**	**0.93**	**1.55**	**6.39**	**88.35**

The joint mode in Eq. (16) is best or tied for best in every column. It retains the teacher displacement throughout reward optimization and therefore maintains a stable distributional anchor. The scalar and vector targets are added in the same clean-prediction space and trained with one loss.

We apply the same joint recipe to the 6B Z-Image student. It improves every trained and held-out metric as well as DPG in Table 3. These results demonstrate transfer beyond SD3.5-M.

3.4 Local Response Diagnostics

We evaluate the first-order response predicted by Theorem 1 on the deployed SD3.5-M sampler. Since DM constructs its update using an estimated posterior mean, we also examine how estimation errors affect this response. Both diagnostics use the group-standardized linear advantage $A(x_0) = \langle x_0, u \rangle$, whose ideal response is 1.

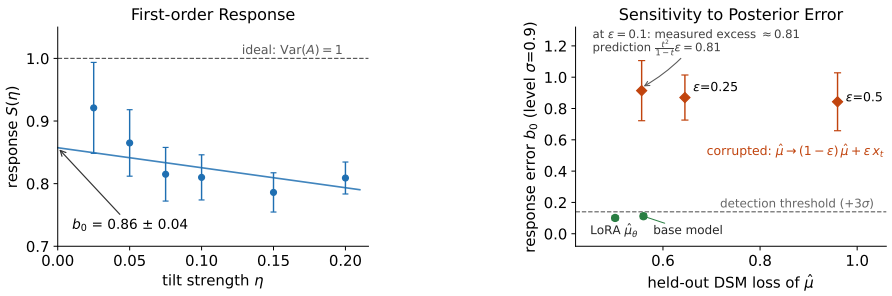


Figure 2: Local response diagnostics. (a) The deployed response extrapolates to $b_0 = 0.86 \pm 0.04$ (ideal: 1). (b) Synthetic posterior bias is detected, while swapping trained denoisers has little effect.

Table 4: Sequential design ablations. DM-self uses the frozen-model prediction in place of $\hat{\mu}_\theta$.

Metric / variant	Default	Visited queries	Mixed queries	No center	DM-self	2nd order
MNR ↑	1.568	1.451	1.540	1.553	1.441	1.576
DPG ↑	85.77	85.23	85.71	85.65	85.03	85.52
Gradient variance ↓	1.0×	–	–	1.8×	–	–

First-order response. We measure $S(\eta) = (\mathbb{E}_{p_{\theta^+_\eta,0}}[A] - \mathbb{E}_{p_{\theta,0}}[A])/\eta$ for $\eta \in [0.025, 0.2]$ and fit $S(\eta) = b_0 + b_1\eta$ to extrapolate to $\eta \to 0$. The fitted intercept is $b_0 = 0.86 \pm 0.04$, recovering 86% of the ideal response on the deployed four-step schedule (Fig. 2a).

Sensitivity to posterior error. At $\sigma_t = 0.9$, replacing the trained LoRA denoiser with the base model raises held-out denoising loss by 12% but leaves response error essentially unchanged. Injecting the bias $\hat{\mu}_\theta \to (1-\epsilon)\hat{\mu}_\theta + \epsilon x_t$ with $\epsilon = 0.1$ raises loss by a comparable 11%, yet increases response error ninefold (Fig. 2b). The excess error matches the analytical prediction (0.81 measured versus 0.81 predicted). Thus, the update remains robust to the tested denoiser swap and responds predictably to the injected posterior bias, despite similar changes in denoising loss.

Overall, DM recovers a substantial fraction (86%) of the ideal first-order response on SD3.5-M. The unit response is an idealized benchmark that neglects finite-step, posterior-estimation, and optimization errors. Recovering 86% of this response in a large pretrained generator with only four sampling steps, learned posterior estimates, and a finite training budget therefore provides direct quantitative evidence for the practical relevance of the first-order analysis. This agreement, together with robustness to the tested denoiser swap and the predicted effect of injected bias, provides converging empirical support for the analysis.

3.5 Ablations

All ablations use sequential SD3.5-M training. Mean Normalized Reward (MNR) averages trained rewards and DPG-Bench tracks fidelity. "Visited queries" use rollout states instead of re-noised queries; "Mixed queries" aggregate re-noised and rollout states; "No center" omits group-mean removal; "DM-self" replaces $\hat{\mu}_\theta$ with the stopped frozen-model prediction; and "2nd order" adds the second-order tilt correction. Table 4 tests the main estimator choices. DM-self reduces both MNR and DPG. Re-noising and centering improve stability and scale balance, while a second-order target brings negligible gain. These trends support the theoretical construction: re-noising supplies the forward-posterior coupling, and separate posterior estimation avoids identifying a finite-step head with a conditional mean. The limited benefit of second-order correction suggests that the first-order target already captures the useful displacement at the tested update scale.

Fig. 3 shows that moderate refresh rates are stable, whereas fully on-policy training raises reward but collapses DPG as denoiser fit loss reaches $2.6\times$ its quasi-static value. Performance is stable around $\eta = 0.15$; the largest step exposes first-order bias and fidelity loss.

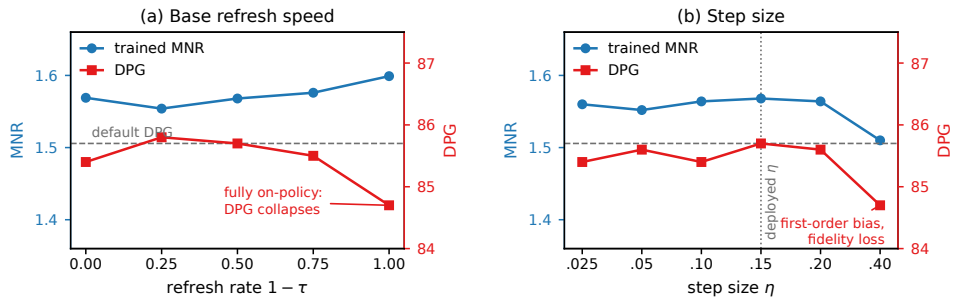


Figure 3: Anchor and step-size sweeps. Fully on-policy refresh raises trained MNR but collapses DPG; η is stable over the middle range.

4 Related Work

Diffusion RL. DDPO (Black et al., 2024), DPOK (Fan et al., 2023), and GRPO-based methods (Liu et al., 2025; Xue et al., 2025; Li et al., 2026) optimize stochastic reverse trajectories using per-step likelihoods. These updates do not directly apply to deterministic few-step samplers, which lack the corresponding stochastic transition densities. DiffusionNFT (Zheng et al., 2025) and AdvantageFlow (Kveton et al., 2026) instead learn from forward-corrupted, scored samples. DM shares AdvantageFlow's local reward tilt, also derived by RAM (Bergmeister et al., 2026) from KL-regularized optimal control. A remaining issue is output semantics: a distilled prediction head represents a finite-interval transition and need not equal a posterior mean. Reward updates derived for instantaneous predictors therefore require further justification under few-step deployment.

Few-step reward alignment and distillation. Existing few-step alignment methods impose different requirements on rewards, samplers, or model parameterizations. DMDR (Jiang et al., 2026) combines distribution matching with reward optimization, but its default ReFL implementation requires differentiable rewards. GDMD (Dong et al., 2026) scores and reweights teacher-derived distillation targets, keeping reward optimization dependent on teacher supervision. TDM-R1 (Luo et al., 2026) guides the generator with a learned surrogate reward, making its updates depend on surrogate accuracy. RTDMD (Huang et al., 2026a) introduces stochastic intermediate transitions for policy gradients, so its optimization operates on a different sampling process from native deterministic deployment. MeanFlowNFT (Huang et al., 2026b) relies on MeanFlow-specific predictor constructions, limiting applicability to that parameterization. DM instead learns reward updates directly from scalar feedback under the native deterministic sampler. It supports students obtained through flow or adversarial distillation and uses the same posterior-displacement regression for reward-only and joint training. Our analysis links these updates to the deployed finite-step response.

5 Conclusion

We introduced Displacement Matching (DM), which translates terminal distribution tilts into posterior-mean displacements, unifying reward alignment and distillation in one regression. Separately estimated posterior means accommodate few-step prediction heads and enable learning from scalar rewards without trajectory likelihoods or rollout backpropagation. Our population analysis establishes first-order unbiasedness of the reward displacement estimator, reward improvement up to an explicit finite-step residual, and a unique teacher fixed point for raw distillation. In pure distillation, the four-step CFG-free SD3.5-M student outperforms prior four-step methods and slightly exceeds its 50-step teacher, reaching DPG-Bench 84.90 versus 84.05. Reward alignment reaches HPS 0.362 on SD3.5-M and GenEval 0.93 on Z-Image, compared with 0.345 and 0.81 for the strongest respective baselines at matched reward-query budgets. Sequential and joint training both improve trained and held-out metrics while preserving fidelity, with gains also observed from adversarially distilled initializations. Local diagnostics recover 86% of the ideal response on SD3.5-M and reproduce the predicted effect of injected posterior bias. Extending the analysis to one-step models, reducing the finite-step response gap, and accommodating larger distribution shifts remain directions for future work.

AI USE STATEMENT

We used generative AI tools (large language models) in the following ways. (i) Writing: to improve the clarity, grammar, and presentation of text written by the authors; AI was not used to draft any section of the paper from scratch. (ii) Retrieval and discovery: to help identify and summarize related work during the literature search; every cited work was subsequently read and verified by the authors. (iii) Mathematical derivations: to assist in deriving and checking intermediate steps of some proofs and equations; the formulation of all mathematical claims, the overall proof strategies, and the final proofs are the authors' own, and every AI-assisted step was independently re-derived and verified by the authors. (iv) Implementation: to assist in writing the code (e.g., evaluation scripts, baseline implementations, and plotting utilities). The method itself, including its design and algorithmic details, was developed by the authors; AI tools were used to translate the authors' specifications into code. All AI-assisted code was reviewed line by line, tested against hand-verified cases, and validated by reproducing the reported results before use.

Generative AI was not used to generate or clean datasets, to develop the theoretical framework or research hypotheses, to design the experiments, to perform qualitative or thematic analysis, to interpret results, or for translation. The research questions, methodology, experimental design, results, and conclusions are the authors' own. We take full responsibility for the content of this paper, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human participants, user studies, personal or sensitive data, or newly collected datasets. Our experiments use publicly available academic datasets and pretrained models released under licenses that permit academic research. We follow the applicable terms of use and do not redistribute restricted assets. To the best of our knowledge, this work raises no additional ethical concerns.

REPRODUCIBILITY STATEMENT

The main paper specifies the proposed method, assumptions, algorithms, and evaluation protocol. The appendix provides the complete derivations and proofs, implementation and training details, hyperparameters, dataset and model information, and additional experimental results. An anonymized implementation, including configuration files and evaluation scripts, will be available at `anonymous.4open.science/r/ICLR2027-DM`. Together, these materials are intended to support reproduction of all main and supplementary results.

REFERENCES

Andreas Bergmeister, Stefanie Jegelka, Nikolas Nüsken, Carles Domingo-Enrich, and Jakiw Pidstrigach. Reinforce adjoint matching: Scaling rl post-training of diffusion and flow-matching models, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Linwei Dong, Ruoyu Guo, Ge Bai, Zehuan Yuan, Yawei Luo, and Changqing Zou. Guiding distribution matching distillation with gradient-based reinforcement learning. *arXiv preprint arXiv:2604.19009*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine*

Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: reinforcement learning for fine-tuning text-to-image diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

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

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

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

Yushi Huang, Xiangxin Zhou, Ruoyu Wang, Chi Zhang, Jun Zhang, and Tianyu Pang. Reinforcing few-step generators via reward-tilted distribution matching, 2026a. URL https://arxiv.org/abs/2605.26108.

Yushi Huang, Xiangxin Zhou, Jun Zhang, Liefeng Bo, and Tianyu Pang. MeanFlowNFT: Bringing forward-process RL to average-velocity generators. *arXiv preprint arXiv:2607.15273*, 2026b.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, Bo Zhang, Mengmeng Wang, Steven Hoi, Peng Gao, and Harry Yang. Distribution matching distillation meets reinforcement learning, 2026. URL https://arxiv.org/abs/2511.13649.

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. *ICLR*, 2024.

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 *Advances in Neural Information Processing Systems*, 2023.

Branislav Kveton, Anup Rao, Subhojyoti Mukherjee, Krishna Kumar Singh, and Viet Dac Lai. Advantageflow: Advantage-weighted least squares for rl in flow models, 2026. URL https://arxiv.org/abs/2605.26013.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2026. URL https://arxiv.org/abs/2507.21802.

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. *arXiv preprint arXiv:2505.05470*, 2025.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching, 2025. URL https://arxiv.org/abs/2503.06674.

Yihong Luo, Tianyang Hu, Weijian Luo, and Jing Tang. Tdm-r1: Reinforcing few-step diffusion models with non-differentiable reward, 2026. URL https://arxiv.org/abs/2603.07700.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients, 2024. URL https://arxiv.org/abs/2407.08737.

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

Christoph Schuhmann. LAION-Aesthetics. LAION blog, 2022. URL https://laion.ai/blog/laion-aesthetics/.

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

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *NeurIPS*, 2024.

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, 2023. URL https://arxiv.org/abs/2306.09341.

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 *NeurIPS*, pp. 15903–15935, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

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

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

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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. *arXiv preprint arXiv:2509.16117*, 2025.

A Experimental Details

All experiments share the training pipeline of Section 2.3: anchor rollouts, re-noised per-level queries (Eq. 10), and the displacement-matching regression (Eq. 8). The modes differ only in the displacement estimator $\widehat{\Delta}_k$ and in the schedules reported below. Unless stated otherwise, both the student and the fake-score posterior-mean denoiser are trained with LoRA adapters of rank $r = 64$ and scaling $\alpha = 64$, at 512^2 resolution with the fixed $K = 4$ CFG-free schedule. Table 5 summarizes all hyperparameters.

A.1 Distillation

Pure distillation uses the vector-branch estimator (Eq. 14). The anchor follows the EMA update in Section 2.5, with its coefficient warmed up linearly from $\tau = 0$ to $\tau = 0.9$ over the first 10 iterations and held at 0.9 thereafter. All remaining optimization settings follow the reward-learning configuration below.

A.2 Reward Learning on SD3.5-M

Prompts. We train on the prompt set of DiffusionNFT (Zheng et al., 2025).

Rewards. Three scalar rewards are used: CLIPScore, HPSv2.1, and PickScore. Each iteration standardizes reward group-wise (Eq. 21, $|G| = 24$) before forming the scalar-branch target. We use all rewards as non-differentiable rewards.

Optimization. We train for 1,000 iterations with learning rate 1×10^{-4} and step size $\eta = 0.15$. The posterior-mean denoiser is trained on a faster time scale (TTUR), taking 5 gradient steps per student step. The anchor EMA coefficient is warmed up linearly from $\tau = 0$ to $\tau = 0.5$ over the first 300 iterations and held constant afterwards, matching the refresh schedule of DiffusionNFT (Zheng et al., 2025).

A.3 Scale-Up on Z-Image

The 6B Z-Image experiments reuse the SD3.5-M configuration with two changes. First, the step size is reduced to $\eta = 0.1$. Second, training follows a two-stage curriculum: (i) 600 iterations on the DiffusionNFT prompt set with the same three rewards (CLIPScore, HPSv2.1, PickScore); (ii) 400 iterations on the GenEval training prompt set (Ghosh et al., 2023) with four rewards, adding the rule-based GenEval score to the previous three. All baselines use the same training schedule. The GenEval reward is non-differentiable; the scalar branch consumes it without modification, since Eq. 11 requires only scalar advantages.

A.4 Evaluation Protocol

CLIPScore, HPSv2.1, PickScore, Aesthetic v2, and ImageReward are computed on the DrawBench prompt set, following the evaluation protocol of DiffusionNFT (Zheng et al., 2025). GenEval and DPG-Bench are computed on their respective official prompt sets with their official evaluation pipelines. All evaluations sample with the deployed $K = 4$ CFG-free schedule at 512^2 resolution.

A.5 Training Cost

We compare wall-clock training costs normalized to a common workload of four update timesteps. Table 6 reports the resulting costs for DM, RTDMD, and TDM-R1. Under this normalization, DM has a similar total cost to RTDMD and requires approximately half the time of TDM-R1.

Fake-score estimation costs 6.2 seconds per update timestep for DM, compared with 5.3 seconds for RTDMD, an increase of approximately 17%. Over four update timesteps, this component accounts for 24.8 seconds, or approximately 31% of DM's total cost, making it one of the largest auxiliary costs. This additional overhead is offset by our faster deterministic rollouts, yielding a comparable overall training cost to RTDMD despite the more expensive fake-score estimation.

Table 5: Training hyperparameters. "Stage 1 / Stage 2" refers to the Z-Image curriculum in Appendix A.3. The distillation anchor schedule applies to the pure-distillation runs of Section 3.2.

	SD3.5-M	Z-Image (6B)
Resolution	512^2	512^2
Student steps K	4 (CFG-free)	4 (CFG-free)
Teacher	50 steps, CFG 4.5	50 steps, CFG 4.0
LoRA rank r / α	64 / 64	64 / 64
Learning rate	1×10^{-4}	1×10^{-4}
TTUR (denoiser : student)	5 : 1	5 : 1
Group size $\|G\|$	24	24
Step size η	0.15	0.10
Anchor EMA τ (reward learning)	$0 \to 0.5$, linear, 300 iters	$0 \to 0.5$, linear, 300 iters
Anchor EMA τ (distillation)	$0 \to 0.9$, linear, 10 iters	—
Iterations	1,000	600 + 400
Training prompts	DiffusionNFT	DiffusionNFT / GenEval
Rewards	CLIP, HPS, Pick	CLIP, HPS, Pick (+ GenEval, Stage 2)

Table 6: Wall-clock training time normalized to four update timesteps.

Method	Normalized time (s)
DM (ours)	81.3
RTDMD	84.1
TDM-R1	161.5

A.6 Training algorithm

Algorithm 1 summarizes the three training modes. Following the main text, θ^+ denotes the trainable student and θ the rollout anchor, which is held fixed within each iteration. The auxiliary denoiser has parameters ψ and estimates the anchor posterior mean $\widehat{\mu}_{\theta,k}$ through Tweedie's identity. It is fitted to independently re-noised anchor outputs using denoising score matching, with the denoiser-to-student update ratio specified in the training details. We select the mode through

$$(\lambda_R, \lambda_D) = \begin{cases} (0,1), & \text{pure distillation,} \\ (\eta, 0), & \text{reward alignment after distillation,} \\ (\eta, \eta), & \text{joint training.} \end{cases}$$

For sequential alignment, the student is initialized from the distilled checkpoint. Rewards are normalized within each prompt group before forming the scalar displacement; the vector displacement is uncentered by default. Both displacements use the same posterior-mean estimate at the same re-noised query. The regression baseline is the anchor prediction at that query. Rollout samples, queries, and regression targets are detached, and only θ^+ receives gradients from the student loss. The anchor is refreshed after the student update.

B Finite-Step Residual and Proofs for the Method Section

This appendix states and proves the exact response result used in Section 2.4 (Proposition 2), derives the transport bound $\varepsilon_K = O(h_K)$, and proves Theorem 1, Lemma 1, and Proposition 1.

B.1 Setting, induced heads, and regularity

The transition in Equation 1 is affine in the head,

$$F_{\theta,k}(x,c) = a_k x + b_k f_\theta(x, t_k, c), \qquad a_k = \frac{\sigma_{t_{k+1}}}{\sigma_{t_k}}, \qquad b_k = \alpha_{t_{k+1}} - a_k \alpha_{t_k}. \tag{22}$$

Lemma 2 (Head to map correspondence). *Assume the signal-to-noise ratio α_t/σ_t is strictly decreasing in t. Then $b_k > 0$ for every $k < K$. Consequently, for any clean law p with probability-flow*

Algorithm 1 Displacement Matching

Require: Student θ^+, auxiliary denoiser ψ
Require: Grid $\{t_k\}_{k=0}^K$, mode weights (λ_R, λ_D), EMA coefficient τ
Require: Teacher posterior mean μ_T when $\lambda_D > 0$
Initialize rollout anchor $\theta \leftarrow \theta^+$
for each training iteration **do**
  $\mathcal{B} = \{(c^i, X_0^{-,i})\}_{i=1}^N \leftarrow \mathrm{Rollout}(\theta)$
  $A^i \leftarrow 0$ for all i
  if $\lambda_R > 0$ **then**
    $\{A^i\}_{i=1}^N \leftarrow \mathrm{GroupNormalize}(\mathrm{Reward}(\mathcal{B}))$
  end if
  Update ψ by denoising score matching on re-noised samples from $\mathcal{B}$
  $\mathcal{L} \leftarrow 0$
  for $k = 0, \dots, K-1$ **do**
    Draw independent $\epsilon_k^i \sim \mathcal{N}(0, I)$ for all i
    $\widetilde{X}_{t_k}^i \leftarrow \alpha_{t_k} X_0^{-,i} + \sigma_{t_k} \epsilon_k^i$
    $\widehat{\mu}_k^i \leftarrow \widehat{\mu}_{\theta,k}(\widetilde{X}_{t_k}^i, c^i)$ using ψ
    $\widehat{\Delta}_k^i \leftarrow \lambda_R A^i (X_0^{-,i} - \widehat{\mu}_k^i)$
    if $\lambda_D > 0$ **then**
      $\widehat{\Delta}_k^i \leftarrow \widehat{\Delta}_k^i + \lambda_D \big(\mu_{T,k}(\widetilde{X}_{t_k}^i, c^i) - \widehat{\mu}_k^i\big)$
    end if
    $x_{0,k}^{\star,i} \leftarrow \mathrm{sg}\big[f_\theta(\widetilde{X}_{t_k}^i, t_k, c^i) + \widehat{\Delta}_k^i\big]$
    $\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{N} \sum_{i=1}^N \big\| f_{\theta^+}(\widetilde{X}_{t_k}^i, t_k, c^i) - x_{0,k}^{\star,i} \big\|_2^2$
  end for
  Update θ^+ to minimize $\mathcal{L}$
  $\theta \leftarrow \tau\theta + (1-\tau)\theta^+$
end for

solution map Ψ^{w_p}, the head

$$f_{p,k}(x, c) := b_k^{-1} \big(\Psi_{t_k \to t_{k+1}}^{w_p}(x) - a_k x \big) \tag{23}$$

is the unique head whose transition equation 22 reproduces the exact interval map. Attribute (P2*) *states $f_\theta(\cdot, t_k, c) = f_{p_{\theta,0},k}(\cdot, c)$ for p_{θ,t_k} almost every point.*

Proof. For $k < K-1$ both noise scales are positive, so $b_k = \sigma_{t_{k+1}}(\alpha_{t_{k+1}}/\sigma_{t_{k+1}} - \alpha_{t_k}/\sigma_{t_k}) > 0$ by strict monotonicity. For $k = K-1$ we have $\sigma_{t_K} = 0$ and $\alpha_{t_K} = 1$, so $b_{K-1} = 1$. Uniqueness follows by solving the affine relation equation 22 for the head. □

For a smooth path $(\rho_0^s)_{s \in (-1,1)}$ with $\rho_0^0 = p_{\theta,0}$, write $f_k^s := f_{\rho_0^s,k}$ and $\bar{\delta}_k := \partial_s|_0 f_k^s$ as in the main text.

Regularity (R). There are $s_0 > 0$ and constants $M_2, L, M_\delta, \Lambda_e$ such that for all $|s| \le s_0$ and $k < K$: (R1) the map $s \mapsto f_k^s(x, c)$ is twice differentiable with $\sup_x \|\partial_s^2 f_k^s(x, c)\|_2 \le M_2$. (R2) every transition considered below is L Lipschitz in x. (R3) the directions $\bar{\delta}_k$ and δ_k are bounded by M_δ, and $e_k = \bar{\delta}_k - \delta_k$ is Λ_e Lipschitz in x. We also assume the almost-everywhere identities in (P2*) hold on an open set containing the supports of the visited marginals. Continuity of both sides upgrades the almost-everywhere statement on such a set.

Lemma 3 (Composition stability)**.** *Let $G_0, \dots, G_{K-1}$ be L Lipschitz maps and let $H_0, \dots, H_{K-1}$ be arbitrary maps. Set $z_0 = z$ and $z_{j+1} = H_j(z_j)$. Then*

$$\big\| H_{K-1} \circ \cdots \circ H_0(z) - G_{K-1} \circ \cdots \circ G_0(z) \big\|_2 \le \sum_{j=0}^{K-1} L^{K-1-j} \big\| H_j(z_j) - G_j(z_j) \big\|_2. \tag{24}$$

Proof. Insert hybrid compositions whose first j maps are H and whose remaining maps are G. Consecutive hybrids differ by the G tail applied to $H_j(z_j)$ and to $G_j(z_j)$. The tail is L^{K-1-j}

Lipschitz, so each difference is at most $L^{K-1-j}\|H_j(z_j) - G_j(z_j)\|_2$. Summing over j proves the claim. □

B.2 Exact response for averaged prediction displacements

Proposition 2 (Exact response)**.** *Assume* (P2*) *and* (R)*. Let* $(\rho_0^s)_{s\in(-1,1)}$ *be smooth with* $\rho_0^0 = p_{\theta,0}$*, let* $f_k^s := f_{\rho_0^s,k}$ *induce the exact interval map* $\Psi^{w_{\rho_0^s}}_{t_k\to t_{k+1}}$*, and set* $\bar\delta_k := \partial_s|_0 f_k^s$*. If* θ_s *replaces every* $f_\theta(\cdot,t_k,c)$ *by* $f_\theta(\cdot,t_k,c) + s\bar\delta_k(\cdot,c)$*, then for every bounded Lipschitz* φ*,*

$$\mathbb{E}_{p_{\theta_s,0}}[\varphi(X_0)] = \mathbb{E}_{\rho_0^s}[\varphi(X_0)] + O_\varphi(s^2). \tag{25}$$

Moreover, replacing $s\bar\delta_k$ *by* $s\delta_k$ *with* $\delta_k := \partial_s|_0 \mu_{\rho_0^s,k}$ *changes the left side by at most* $O_\varphi(s\varepsilon_K) + O_\varphi(s^2)$*, where* $e_k := \bar\delta_k - \delta_k$ *and* $\varepsilon_K := \max_{k<K}(\mathbb{E}_{X\sim p_{\theta,t_k}}\|e_k(X,t_k,c)\|_2^2)^{1/2}$*.*

Proof. **Step 1: the path law is a student.** Since $(\alpha_{t_0},\sigma_{t_0}) = (0,1)$, the forward marginal of every clean law at t_0 is $\mathcal{N}(0,I)$. The probability-flow field of ρ_0^s transports the marginals of its own corruption family, so composing its exact interval maps carries $Z\sim\mathcal{N}(0,I)$ to a sample of ρ_0^s. By Lemma 2, the student with heads f_k^s realizes exactly this composition. Its terminal law is ρ_0^s.

Step 2: per-level gap of the perturbed student. The perturbed student uses heads $f_\theta + s\bar\delta_k$. By (P2*), $f_\theta(\cdot,t_k,c) = f_k^0$ on the visited supports. A Taylor expansion under (R1) gives

$$\sup_x \big\|f_k^0(x,c) + s\bar\delta_k(x,c) - f_k^s(x,c)\big\|_2 \le \tfrac{M_2}{2}\, s^2. \tag{26}$$

The transitions of the two students therefore differ by at most $b_k\frac{M_2}{2}s^2$ in supremum norm.

Step 3: composition. Apply Lemma 3 with the exact path transitions as the G maps, which are L Lipschitz by (R2). The terminal samples of the two students differ by at most $C_1 s^2$ pointwise, with $C_1 = \frac{M_2}{2}\sum_j L^{K-1-j}b_j$. For bounded Lipschitz φ,

$$\big|\mathbb{E}_{p_{\theta_s,0}}[\varphi(X_0)] - \mathbb{E}_{\rho_0^s}[\varphi(X_0)]\big| \le \mathrm{Lip}(\varphi)\, C_1 s^2,$$

which is Equation 25.

Residual term. Replacing $s\bar\delta_k$ by $s\delta_k$ changes the transition at level k by sb_ke_k evaluated at the visited state of that level. Let x_k^0 denote the visited state of the unperturbed student, whose law is p_{θ,t_k} by (P1). By Lemma 3 applied against the unperturbed student, the visited state z_k of either perturbed student satisfies $\|z_k - x_k^0\|_2 \le C_2 s$ with C_2 built from L, b_j, and M_δ. Using (R3) and the Cauchy-Schwarz inequality,

$$\mathbb{E}\big\|e_k(z_k)\big\|_2 \le \big(\mathbb{E}_{p_{\theta,t_k}}\|e_k\|_2^2\big)^{1/2} + \Lambda_e C_2 s \le \varepsilon_K + \Lambda_e C_2 s.$$

Taking expectations in Lemma 3 bounds the change of the terminal expectation by $\mathrm{Lip}(\varphi)\sum_k L^{K-1-k}b_k\, s(\varepsilon_K + \Lambda_e C_2 s) = O_\varphi(s\varepsilon_K) + O_\varphi(s^2)$. □

B.3 Transport formula and the bound $\varepsilon_K = O(h_K)$

Normalization. Use the main-text schedule $t_0 > \cdots > t_K$, with $\Delta t_k := t_k - t_{k+1} > 0$ and $h_K := \max_k \Delta t_k$. Denote the auxiliary flow-time coordinate by $\tilde t = \psi(t)$. Assume ψ is strictly decreasing and C_{time} Lipschitz on the full schedule interval, with C_{time} independent of grid refinement. Set $\tilde t_k := \psi(t_k)$ and $\Delta\tilde t_k := \tilde t_{k+1} - \tilde t_k > 0$. Then

$$\Delta\tilde t_k \le C_{\mathrm{time}}\Delta t_k, \qquad \tilde h_K := \max_k \Delta\tilde t_k \le C_{\mathrm{time}} h_K.$$

In these coordinates, the probability-flow ODE reads $\dot y_{\tilde t} = w(y_{\tilde t},\tilde t)$ over $[\tilde t_k,\tilde t_{k+1}]$, and the induced head is the interval average $u_k(y) = (\Delta\tilde t_k)^{-1}\big(\Psi^w_{\tilde t_k\to\tilde t_{k+1}}(y) - y\big)$. Level-wise, this head is an invertible affine function of the clean-prediction head, with the conversion back to clean-prediction coordinates uniformly bounded over the considered schedules. By Tweedie's identity, the instantaneous field $w_p(\cdot,\tilde t)$ is likewise a level-wise affine function of the posterior mean μ_p. Tangent gaps therefore transfer between the two coordinate systems up to bounded constant factors, and it suffices

to bound the defect for the field tangents. We mark these normalized tangents with a superscript w to distinguish them from the clean-prediction tangents above.

Write $w^s := w_{\rho_0^s}$ and $\dot{w}_0 := \partial_s|_0 w^s$. Differentiating the exact solution map in s gives the variation-of-constants formula

$$\bar{\delta}_k^w(y) = \frac{1}{\Delta\tilde{t}_k} \int_{\tilde{t}_k}^{\tilde{t}_{k+1}} D\Psi^{w^0}_{q\to\tilde{t}_{k+1}}(y_q)\, \dot{w}_0(y_q, q)\, dq, \qquad y_q = \Psi^{w^0}_{\tilde{t}_k\to q}(y), \tag{27}$$

while the left-endpoint tangent is $\delta_k^w(y) = \dot{w}_0(y, \tilde{t}_k)$. Write $e_k^w := \bar{\delta}_k^w - \delta_k^w$ for their gap.

Assumptions (T). Assume the normalization bounds above. On a region containing the trajectories, $|w^0| \le M_w$, the field w^0 is L_w Lipschitz in y uniformly in $\tilde{t}$, and $\dot{w}_0$ is bounded by M and Λ Lipschitz in $(y, \tilde{t})$, with constants independent of grid refinement.

Lemma 4 (Defect bound)**.** *Under* (T), $\sup_y \|e_k^w(y)\|_2 \le C_T\, \Delta\tilde{t}_k$ *with* $C_T = \Lambda(M_w + 1) + L_w M e^{L_w \tilde{h}_K}$. *Hence* $\varepsilon_K \le C\, \tilde{h}_K \le C C_{\mathrm{time}} h_K = O(h_K)$, *where* C *absorbs the uniformly bounded coordinate conversion.*

Proof. Fix $q \in [\tilde{t}_k, \tilde{t}_{k+1}]$ and write $I(q)$ for the integrand in equation 27. Three elementary estimates control $I(q) - \dot{w}_0(y, \tilde{t}_k)$. First, $\|y_q - y\|_2 \le M_w \Delta\tilde{t}_k$ because $|w^0| \le M_w$. Second, Gronwall's inequality gives $\|D\Psi^{w^0}_{q\to\tilde{t}_{k+1}} - \mathrm{Id}\| \le e^{L_w\Delta\tilde{t}_k} - 1 \le L_w \Delta\tilde{t}_k\, e^{L_w\tilde{h}_K}$. Third, the Lipschitz property of $\dot{w}_0$ gives $\|\dot{w}_0(y_q, q) - \dot{w}_0(y, \tilde{t}_k)\|_2 \le \Lambda(\|y_q - y\|_2 + |q - \tilde{t}_k|) \le \Lambda(M_w + 1)\Delta\tilde{t}_k$. Combining,

$$\|I(q) - \dot{w}_0(y, \tilde{t}_k)\|_2 \le \big(L_w M e^{L_w\tilde{h}_K} + \Lambda(M_w + 1)\big)\Delta\tilde{t}_k.$$

Averaging over q preserves the bound, so $\|\bar{\delta}_k^w(y) - \delta_k^w(y)\|_2 \le C_T \Delta\tilde{t}_k$. Applying the bounded coordinate conversion, taking the maximum over k and the root mean square over p_{θ,t_k}, and using $\tilde{h}_K \le C_{\mathrm{time}} h_K$ yields the claim. □

The three contributions have a direct reading. They are the interval transport of the query, the Jacobian response of the base flow, and the time variation of the tangent inside the interval. Each vanishes linearly with the interval length, so the defect is a genuinely finite-step object.

B.4 Proof of Theorem 1

We first record an exact identity behind Equation 6.

Lemma 5 (Exact tilt of the posterior mean)**.** *Fix* (x, t, c) *and write* $\mu = \mu_\theta(x, t, c)$, $\bar{A}(x) = \mathbb{E}[A(X_0, c) \mid X_t = x, c]$, *and*

$$C(x, t, c) = \mathbb{E}\big[A(X_0, c)\,(X_0 - \mu) \mid X_t = x, c\big],$$

with $X_0 \sim p_{\theta,0}$. *Let* $|A| \le M_A$ *and* $0 < \eta \le 1/(2M_A)$. *Then*

$$\Delta_{r_\eta}(x, t, c) = \mu_{r_\eta}(x, t, c) - \mu(x, t, c) = \frac{\eta\, C(x, t, c)}{1 + \eta\bar{A}(x)}. \tag{28}$$

In particular, for the path $\rho_0^s = p_{\theta,0}^{1+sA}$ *of Appendix B.1, the posterior tangent is* $\delta_k = \partial_s|_0 \mu_{\rho_0^s,k} = C(\cdot, t_k, c)$. *If in addition the conditional second moment* $\mathbb{E}[\|X_0 - \mu\|_2^2 \mid X_t = x, c]$ *is bounded by* V, *then* $\|\Delta_{r_\eta} - \eta C\|_2 \le 2\eta^2 M_A^2 V^{1/2}$ *uniformly in* x, *which proves Equation 6 with a uniform remainder.*

Proof. By Equation 5 with $r_\eta = 1 + \eta A$,

$$\mu_{r_\eta}(x, t, c) = \frac{\mu + \eta\, \mathbb{E}[AX_0 \mid X_t = x, c]}{1 + \eta\bar{A}(x)}.$$

Subtracting μ and simplifying the numerator to $\eta\, \mathbb{E}[A(X_0 - \mu) \mid X_t = x, c] = \eta C$ gives equation 28. Differentiating at $\eta = 0$ gives $\delta_k = C$. For the remainder, $\Delta_{r_\eta} - \eta C = -\eta^2 \bar{A}\, C/(1 + \eta\bar{A})$. The bound follows from $|\bar{A}| \le M_A$, $1 + \eta\bar{A} \ge 1/2$, and $\|C\|_2 \le M_A V^{1/2}$ by the Cauchy-Schwarz inequality. □

Proof of Theorem 1. Consider the path $\rho_0^s = p_{\theta,0}^{1+sA}$, a valid density path for $|s| \le s_0 < 1/M_A$ that is smooth in s, and let δ_k, $\bar{\delta}_k$, and ε_K be its posterior tangent, exact tangent, and finite-step gap as defined in Appendix B.1. Note that $\rho_0^\eta = p_{\theta,0}^{r_\eta}$.

For Equation 19, compare three students. Student (a) has heads $f_\theta + \Delta_{r_\eta}(\cdot, t_k, c)$ and terminal law $p_{\theta^+,0}$. Student (b) has heads $f_\theta + \eta\delta_k$. Student (c) has heads $f_\theta + \eta\bar{\delta}_k$. By Lemma 5, the heads of (a) and (b) differ by at most $2\eta^2 M_A^2 V^{1/2}$ in supremum norm, so Lemma 3 gives a terminal gap of $O_\varphi(\eta^2)$. The gap between (b) and (c) is the residual part of Proposition 2 at $s = \eta$, which contributes $O_\varphi(\eta\varepsilon_K) + O_\varphi(\eta^2)$. The gap between (c) and $\rho_0^\eta = p_{\theta,0}^{r_\eta}$ is Equation 25 of Proposition 2 at $s = \eta$, contributing $O_\varphi(\eta^2)$. Summing the three gaps proves Equation 19.

For Equation 20, the normalized tilt gives

$$\mathbb{E}_{p_{\theta,0}^{r_\eta}}[A] = \frac{\mathbb{E}_{p_{\theta,0}}[A] + \eta\,\mathbb{E}_{p_{\theta,0}}[A^2]}{1 + \eta\,\mathbb{E}_{p_{\theta,0}}[A]} = \mathbb{E}_{p_{\theta,0}}[A] + \eta\,\mathrm{Var}_{p_{\theta,0}}(A) + O(\eta^2),$$

using $|A| \le M_A$. Applying Equation 19 with $\varphi = A$ finishes the proof. □

B.5 Proof of Lemma 1

Proof of Lemma 1. **Coupling.** The anchor output satisfies $X_0^- \sim p_{\theta,0}(\cdot \mid c)$ by construction, and Equation 10 draws $\widetilde{X}_k \sim q_{t_k}(\cdot \mid X_0^-)$ with independent noise. The pair $(X_0^-, \widetilde{X}_k)$ therefore has exactly the joint law of (X_0, X_{t_k}) that defines the posterior mean in Equation 4. Conditional expectations of functions of X_0^- given $\widetilde{X}_k = x$ coincide with posterior expectations under $p_{\theta,0}$.

Scalar branch. With the coupling identity and Lemma 5,

$$\mathbb{E}[\widehat{\Delta}_{R,k} \mid \widetilde{X}_k = x, c] = \eta\, C(x, t_k, c) = \big(1 + \eta\bar{A}(x)\big)\,\Delta_{r_\eta}(x, t_k, c) = \Delta_{r_\eta}(x, t_k, c) + O(\eta^2),$$

where the remainder is uniform under the moment bound of Lemma 5. This is Equation 12.

Vector branch. The estimator in Equation 13 is a deterministic function of $\widetilde{X}_k$ and equals $\Delta_r(\widetilde{X}_k, t_k, c)$ by the definition in Equation 6. This is Equation 13.

Joint branch. Let γ_R and γ_D denote the coefficients of the scalar and vector terms in Equation 16, respectively, and define $r_R = 1 + \gamma_R A$ and $r_D = p_{T,0}/p_{\theta,0}$. By linearity of conditional expectation and the two branch identities above,

$$\begin{aligned}
&\mathbb{E}[\widehat{\Delta}_{\mathrm{joint},k} \mid \widetilde{X}_k = x, c] \\
&\quad = \mathbb{E}\left[\gamma_R A(X_0^-, c)\big(X_0^- - \mu_{\theta,k}(\widetilde{X}_k, c)\big) \mid \widetilde{X}_k = x, c\right] + \gamma_D\big(\mu_{T,k}(x,c) - \mu_{\theta,k}(x,c)\big) \\
&\quad = \Delta_{r_R}(x, t_k, c) + \gamma_D \Delta_{r_D}(x, t_k, c) + O(\gamma_R^2).
\end{aligned} \tag{29}$$

Thus the joint estimator is first-order unbiased for the combined reward and teacher displacement in Equation 16. Since $\gamma_R, \gamma_D = O(\eta)$, its remainder is $O(\eta^2)$.

Visited states. If the query were the visited state $x_{t_k}^-$ of the deterministic rollout, then X_0^- would be a measurable function of the query. The conditional law of X_0^- given the query would be a point mass, so the conditional expectation of the scalar estimator would reduce to a single weighted residual rather than the conditional covariance, and Equation 12 would fail. This justifies the renoised coupling in the main text. □

B.6 Proof of Proposition 1

Proof of Proposition 1. By Equation 9, the branch leaves the head unchanged at level k if and only if $\mathbb{E}[\widehat{\Delta}_{D,k} \mid \widetilde{X}_k = x, c] = 0$ for p_{θ,t_k} almost every x. The target equation 14 is deterministic given $\widetilde{X}_k$, so stationarity reads $\mu_{T,k} = \mu_{\theta,k}$ on the same set. Here both posterior means are exact, as stated by the population ideal.

Fix a level with $\alpha_{t_k} > 0$. Every level $k \le K - 1$ has $t_k > 0$ and hence $\sigma_{t_k} > 0$, so p_{θ,t_k} is a Gaussian mixture with a strictly positive density on all of $\mathcal{X}_0$. The identity $\mu_{T,k} = \mu_{\theta,k}$ therefore

holds Lebesgue almost everywhere. Tweedie's identity gives, for $M \in \{\theta, T\}$,

$$\nabla \log p_{M,t_k}(x \mid c) = \frac{\alpha_{t_k}\, \mu_{M,k}(x,c) - x}{\sigma_{t_k}^2},$$

so the two score fields agree almost everywhere. Both densities are smooth and positive by Gaussian smoothing, so their log ratio is constant, and normalization forces $p_{T,t_k} = p_{\theta,t_k}$. In Fourier terms, $\widehat{p}_{M,t_k}(\xi) = \widehat{p}_{M,0}(\alpha_{t_k}\xi)\exp(-\sigma_{t_k}^2\|\xi\|_2^2/2)$. The Gaussian factor never vanishes and $\alpha_{t_k} > 0$, so $\widehat{p}_{T,0} = \widehat{p}_{\theta,0}$ everywhere and $p_{T,0} = p_{\theta,0}$.

Conversely, equal clean laws give equal posterior means at every level. The target then vanishes and the branch is stationary, so the teacher is the unique fixed point. □

C Detailed Convergence of Iterated Linear Tilts

This appendix gives the population-level statement used in Section 2.4. It concerns the ideal density chain in which each round applies the linear tilt exactly. The implemented displacement regression is an inexact realization of one round.

Theorem 2 (Iterated linear tilt)**.** *Let $A(\cdot, c)$ be bounded and choose $\eta > 0$ such that $1+\eta A(x_0, c) > 0$ on the support of p_0. Define*

$$p_{n+1}(x_0 \mid c) = \frac{(1+\eta A(x_0,c))p_n(x_0 \mid c)}{\mathbb{E}_{p_n}[1+\eta A(X_0,c)]}. \tag{30}$$

Then

$$p_n(x_0 \mid c) = \frac{(1+\eta A(x_0,c))^n p_0(x_0 \mid c)}{\mathbb{E}_{p_0}[(1+\eta A(X_0,c))^n]}. \tag{31}$$

Moreover,

$$\mathbb{E}_{p_{n+1}}[A] - \mathbb{E}_{p_n}[A] = \frac{\eta \operatorname{Var}_{p_n}(A)}{1+\eta\mathbb{E}_{p_n}[A]} \geq 0. \tag{32}$$

Finally, for any $a < b$ with $p_0(A \geq b) > 0$,

$$p_n(A \leq a) \leq \frac{p_0(A \leq a)}{p_0(A \geq b)} \left(\frac{1+\eta a}{1+\eta b}\right)^n. \tag{33}$$

Consequently, the ideal chain concentrates on the essential maximizers of A over the support of p_0.

Proof. Iterating equation 30 gives equation 31. For monotonicity,

$$\mathbb{E}_{p_{n+1}}[A] = \frac{\mathbb{E}_{p_n}[A(1+\eta A)]}{\mathbb{E}_{p_n}[1+\eta A]} = \frac{\mathbb{E}_{p_n}[A] + \eta\mathbb{E}_{p_n}[A^2]}{1+\eta\mathbb{E}_{p_n}[A]}.$$

Subtracting $\mathbb{E}_{p_n}[A]$ yields equation 32. The denominator is positive because the tilt weight is positive. For concentration, the closed form gives

$$p_n(A \leq a) = \frac{\mathbb{E}_{p_0}[\mathbf{1}_{\{A\leq a\}}(1+\eta A)^n]}{\mathbb{E}_{p_0}[(1+\eta A)^n]}.$$

The numerator is at most $p_0(A \leq a)(1+\eta a)^n$, while the denominator is at least $p_0(A \geq b)(1+\eta b)^n$. This proves equation 33. Taking a below the essential supremum of A and $b > a$ proves concentration on the top-advantage region. □

Inexact rounds. The theorem describes the ideal population chain. In the algorithm, each round is realized through displacement regression on re-noised queries, so there are per-round residuals: the scalar $O(\eta^2)$ expansion error, the finite-step identification term $O_\varphi(\eta\varepsilon_K)$, possible anchor-calibration error, finite-sample error, and optimization error. The main text therefore uses Theorem 2 as the limiting population behavior, while practical training is an inexact version of equation 30.

Table 7: Full diffusion RL comparison on SD3.5-M.

Method	HPS	CLIP	Pick	GenEval	IR	Aes	DPG
FlowGRPO	0.343	0.302	23.01	0.68	1.52	6.44	85.17
DiffusionNFT	0.358	0.319	23.48	**0.69**	**1.54**	6.52	85.77
DM	**0.361**	**0.328**	**23.57**	**0.69**	**1.54**	**6.55**	**85.97**

D Reward Alignment for Full-Step Diffusion

We further evaluate DM on the original SD3.5-M model using 25 denoising steps per rollout as full step diffusion. We follow the main experimental settings and use the model's native denoising prediction in place of the separately trained posterior-mean estimator. This evaluates the same reward-alignment objective beyond distilled few-step generators.

As shown in Table 7, DM outperforms FlowGRPO and DiffusionNFT on all three training rewards. It also achieves the highest Aesthetic and DPG-Bench scores while matching DiffusionNFT on ImageReward. These results establish that DM remains effective in the full-step regime, improving reward scores while maintaining strong held-out performance.

The comparison with our four-step results further highlights the effectiveness of distillation followed by DM. The sequential variant achieves HPS 0.358 and DPG-Bench 85.77, close to 0.361 and 85.97 for 25-step DM, while obtaining a slightly higher PickScore (23.63 versus 23.57). Joint training achieves similarly competitive results. CLIPScore remains the main gap, at 0.302–0.304 for the four-step variants versus 0.328 for full-step DM. Overall, our few-step pipeline approaches the final quality of full-step reward alignment with only four sampling steps.

E Pure Distillation: Displacement Matching versus DMD

E.1 Shared posterior direction and update construction

DM and Distribution Matching Distillation (DMD) share a teacher–student posterior signal. DM places this signal within the terminal-tilt formulation used for both distillation and reward alignment. In pure distillation, the teacher-to-anchor density ratio defines the teacher law as the tilted target. At a re-noised query $\widetilde{X}_k^i$, the corresponding clean-space displacement is estimated by

$$u_k^i = \mu_{T,k}(\widetilde{X}_k^i, c) - \widehat{\mu}_{\bar{\theta},k}(\widetilde{X}_k^i, c), \tag{34}$$

where $\bar{\theta}$ denotes the frozen or EMA rollout anchor and θ^+ denotes the trainable copy. The student fits the resulting displacement relative to the anchor prediction:

$$\ell_{D,k}^i = \frac{1}{2}\left\| f_{\theta^+}(\widetilde{X}_k^i, t_k, c) - f_{\bar{\theta}}(\widetilde{X}_k^i, t_k, c) - u_k^i \right\|_2^2. \tag{35}$$

The queries, anchor predictions, and displacement targets are held fixed when differentiating with respect to θ^+. The anchor is refreshed by

$$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta^+, \tag{36}$$

so rollout samples, the fitted posterior $\widehat{\mu}_{\bar{\theta}}$, and the reference prediction are defined relative to the same anchor law. Here $\tau = 0$ gives the fully on-policy variant, whereas $\tau \to 1$ approaches a frozen anchor.

DMD derives an approximate reverse-KL gradient from the difference between a teacher score and a fake score fitted to generated samples (Yin et al., 2024b). For the same forward corruption, the score identity gives

$$s_T(x_t, t) - s_{\text{fake}}(x_t, t) = \frac{\alpha_t}{\sigma_t^2}\left(\mu_T(x_t, t) - \mu_{\text{fake}}(x_t, t)\right), \tag{37}$$

which recovers the posterior difference in Eq. 34 when the student reference law is matched. Thus the raw teacher-minus-student direction, re-noising, and the use of a separate posterior estimator are shared with DMD.

Table 8: Pure-distillation results with different centering and EMA configurations. The first two columns indicate group-mean removal from the displacement and use of an EMA rollout anchor. All rows retain DM's query and regression construction.

Group mean	EMA anchor	Variant	Training	Failures / runs	GenEval ↑	DPG-Bench ↑
✓	✓	group centering	Stable	$0/3$	0.65	84.4
–	✓	DM (ours)	Stable	$0/3$	0.66	84.9
✓	–	On-policy posterior	Collapse	$3/3$	–	–
–	–	Uncentered on-policy	Stable	$0/3$	0.61	82.3

The update construction specifies how this signal changes the student. DMD evaluates the score difference on a re-noised generated output and backpropagates it through the mapping that produced that output. In Eq. 35, DM instead fits the change in the student's prediction at the re-noised query itself, relative to the anchor prediction at that same query. The distinction concerns the sampling coupling, the student Jacobian receiving the displacement, and the reference held fixed during fitting. The anchor jointly defines the rollout law, posterior reference, and prediction baseline in DM.

Original DMD also uses paired regression for stabilization, while DMD2 introduces a two-time-scale update rule, a GAN objective, and multi-step generator training (Yin et al., 2024b;a). The uncentered on-policy configuration below retains DM's query distribution and regression objective. It shares DMD's posterior signal, while the published DMD/DMD2 systems use their respective optimization procedures. DM uses the shared signal as an anchor-relative target, providing a common displacement-regression interface for teacher distillation and scalar-reward alignment.

E.2 Practical roles of vector centering and EMA anchoring

Vector centering subtracts the component shared across a sampled group. The centered vector branch also achieves strong distillation performance, as shown in Table 8. We use the uncentered variant as the default configuration.

EMA anchoring defines a slowly changing reference for the rollout samples, the fitted posterior, and the prediction baseline. Its role in defining the update remains the same with or without vector centering. The teacher fixed-point result in the main text concerns the raw population displacement in Eq. (34), before the finite-group processing used in the implementation.

E.3 Comparison of centering and EMA configurations

Table 8 compares four configurations with the same student initialization, teacher, four-step schedule, posterior architecture, and training budget. The table evaluates the resulting configurations. We report three independent runs per configuration; failures/runs refers to the settings used for each row. A run is declared collapsed when its training loss diverges or its generated samples exhibit persistent visual corruption.

Both EMA-anchored configurations train successfully. The centered variant reaches 0.65 GenEval and 84.4 DPG-Bench, close to the default DM scores of 0.66 and 84.9. Effective distillation therefore does not depend on vector centering, while the uncentered default achieves a modest advantage in the reported comparison. Relative to the DMD baseline in Table 1 (0.63 GenEval and 82.6 DPG-Bench), both EMA-anchored variants attain higher reported scores, demonstrating that the uncentered branch is itself an effective distillation procedure.

The centered on-policy configuration fails in all three runs, whereas the uncentered on-policy configuration is stable but reaches lower scores (0.61 GenEval and 82.3 DPG-Bench). In the evaluated configurations, the uncentered EMA variant gives the strongest overall results, with the centered EMA variant providing a closely competitive alternative.

F Displacement Matching and AdvantageFlow

F.1 DM-self is exactly matched adaptive AdvantageFlow

Fix one rollout sample and suppress the prompt, noise level, and schedule index. Let

$$y = f_{\theta^+}(\widetilde{X}, t, c), \qquad g = \mathrm{sg}[f_\theta(\widetilde{X}, t, c)], \qquad x = X_0,$$

where θ is the frozen or EMA rollout model and θ^+ is its trainable copy. Replacing the separately estimated posterior mean in the reward displacement by g gives the DM-self loss

$$\ell_{\mathrm{self}} = \|y - g - \eta A(x - g)\|_2^2 . \tag{38}$$

Consider AdvantageFlow with the advantage-dependent rollout coefficient

$$\gamma_\eta(A) = \eta^{-1} - A \tag{39}$$

and no unmatched reference term. Its per-sample objective is

$$\ell_{\mathrm{AF}}^{\mathrm{adapt}} = A\|y - x\|_2^2 + (\eta^{-1} - A)\|y - g\|_2^2. \tag{40}$$

Proposition 3 (Exact sample-wise equivalence)**.** *For any x, y, g, A, and $\eta > 0$,*

$$\eta\ell_{\mathrm{AF}}^{\mathrm{adapt}} = \ell_{\mathrm{self}} + \eta A(1 - \eta A)\|x - g\|_2^2. \tag{41}$$

Consequently, the scaled adaptive AdvantageFlow loss $\widetilde{\ell}_{\mathrm{AF}}^{\mathrm{adapt}} := \eta\ell_{\mathrm{AF}}^{\mathrm{adapt}}$ and DM-self have identical gradients and Hessians with respect to θ^+ and the same minimizers. The unscaled AdvantageFlow gradient differs only by the global factor $1/\eta$.

Proof. Set $d = y - g$ and $z = x - g$. Expanding the scaled AdvantageFlow loss gives

$$\eta\ell_{\mathrm{AF}}^{\mathrm{adapt}} = \eta A\|d - z\|_2^2 + (1 - \eta A)\|d\|_2^2 \tag{42}$$

$$= \|d\|_2^2 - 2\eta A\langle d, z\rangle + \eta A\|z\|_2^2. \tag{43}$$

On the other hand,

$$\ell_{\mathrm{self}} = \|d\|_2^2 - 2\eta A\langle d, z\rangle + \eta^2 A^2\|z\|_2^2. \tag{44}$$

Subtracting the two expressions yields

$$\eta\ell_{\mathrm{AF}}^{\mathrm{adapt}} - \ell_{\mathrm{self}} = \eta A(1 - \eta A)\|z\|_2^2.$$

Because g is stopped, the right-hand side is independent of y and θ^+, which proves the gradient, Hessian, and minimizer statements. □

For $\eta = 1$, Equation 39 reduces to the published AdvantageFlow choice $\gamma(A) = 1 - A$. More generally, the matched objective can be implemented by defining $A' = \eta A$ and applying the same $1 - A'$ rule. The identity itself holds for arbitrary g and does not require g to be a posterior mean. Posterior self-consistency is instead required to interpret this shared update as the posterior displacement of the rollout-induced distribution.

Scope of the identity. The equality is pointwise and therefore holds under any common query distribution. It requires the same rollout sample x, re-noised query $\widetilde{X}$, advantage normalization, step size, per-level weight, stop-gradient anchor, and reduction convention. A reference penalty preserves the equality only when the identically scaled term is added to both losses. Different query schedules, non-stopped anchors, or the fixed AdvantageFlow(1.1) coefficient define different updates.

Table 9: Numerical verification of the exact equivalence between DM-self and scaled matched adaptive AdvantageFlow.

Precision	$E_{\mathrm{id}} \downarrow$	$E_{\mathrm{grad}} \downarrow$	Gradient cosine $\uparrow$	Output drift $\downarrow$
FP32	<1e-7	<1e-6	>0.999	<1e-6
BF16	<1e-3	<1e-2	>0.999	<1e-3

F.2 Numerical verification of the identity

We verify Proposition 3 independently of downstream training. We cache a minibatch $\{(X_0^i, \widetilde{X}_k^i, A_i, c_i)\}_{i,k}$ and evaluate both objectives from the same parameters, using identical per-level weights and detached target networks. Dropout and stochastic augmentation are disabled, and both losses use the same accumulation order. We report

$$E_{\mathrm{id}} = \frac{\left|\widetilde{L}_{\mathrm{AF}}^{\mathrm{adapt}} - L_{\mathrm{self}} - C\right|}{|L_{\mathrm{self}}| + |C| + 10^{-12}}, \tag{45}$$

$$E_{\mathrm{grad}} = \frac{\|\nabla_{\theta^+} \widetilde{L}_{\mathrm{AF}}^{\mathrm{adapt}} - \nabla_{\theta^+} L_{\mathrm{self}}\|_2}{\|\nabla_{\theta^+} L_{\mathrm{self}}\|_2 + 10^{-12}}, \tag{46}$$

$$C = \frac{1}{B} \sum_{i,k} \eta A_i (1 - \eta A_i) \|X_0^i - g_k^i\|_2^2, \tag{47}$$

together with the gradient cosine similarity. As a trajectory-level check, we additionally run 5 deterministic optimizer steps from identical parameters and optimizer states and measure prediction drift on a held-out query batch.

F.3 Posterior correction beyond the self-posterior special case

Let

$$\mu(x,t,c) = \mathbb{E}[X_0 \mid X_t = x, c], \qquad e(x,t,c) = f_\theta(x,t,c) - \mu(x,t,c),$$

and let $m(x,t,c) = \mathbb{E}[A(X_0,c) \mid X_t = x, c]$. The desired scalar posterior tangent is

$$C(x,t,c) = \mathbb{E}[A(X_0,c)(X_0 - \mu(x,t,c)) \mid X_t = x, c].$$

Replacing μ by the rollout prediction gives

$$\mathbb{E}[A(X_0,c)(X_0 - f_\theta(x,t,c)) \mid X_t = x, c] = C(x,t,c) - m(x,t,c)e(x,t,c). \tag{48}$$

Global or per-prompt centering gives $\mathbb{E}[A \mid c] = 0$, but does not generally imply $m(x,t,c) = 0$. Thus, the DM-self/AdvantageFlow special case matches the intended posterior tangent only when the rollout head is posterior self-consistent or the advantage is conditionally centered.

The sequential ablation in Table 4 measures the downstream consequence at four steps. Replacing $\hat{\mu}_\theta$ by the stopped rollout prediction reduces MNR from 1.568 to 1.441 and DPG-Bench from 85.77 to 85.0. Since this replacement is the matched adaptive AdvantageFlow special case above, the comparison isolates the separately estimated posterior target under otherwise shared couplings and optimization.

F.4 Head–posterior mismatch and its downstream consequence

AdvantageFlow's variance-reduction analysis identifies the rollout prediction with the conditional mean of the rollout-induced clean distribution. We measure this posterior-mean identification directly. For a deployed model M, let

$$X_0^M \sim p_{M,0}(\cdot \mid c), \qquad \widetilde{X}_t^M = \alpha_t X_0^M + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

and define

$$\mu_M(x,t,c) = \mathbb{E}_{p_{M,0}}[X_0^M \mid \widetilde{X}_t^M = x, c].$$

We fit an auxiliary posterior estimator $\hat{\mu}_M$ using re-noised samples from the model's own deployed output distribution. The normalized head–posterior gap is

$$g_M(t) = \frac{\mathbb{E}\Big[\|f_M(\widetilde{X}_t^M, t, c) - \hat{\mu}_M(\widetilde{X}_t^M, t, c)\|_2^2\Big]}{\mathbb{E}\Big[\|X_0^M - \hat{\mu}_M(\widetilde{X}_t^M, t, c)\|_2^2\Big]}. \tag{49}$$

All expectations use held-out prompts, rollout samples, and forward noise.

If $\hat{\mu}_M = \mu_M$, conditional-mean orthogonality gives

$$\mathbb{E}\|X_0^M - f_M\|_2^2 = \mathbb{E}\|X_0^M - \mu_M\|_2^2 + \mathbb{E}\|f_M - \mu_M\|_2^2, \tag{50}$$

and hence

$$g_M(t) = \frac{\mathcal{R}_t(f_M) - \mathcal{R}_t(\mu_M)}{\mathcal{R}_t(\mu_M)}, \qquad \mathcal{R}_t(h) := \mathbb{E}\|X_0^M - h(\widetilde{X}_t^M, t, c)\|_2^2. \tag{51}$$

Thus $g_M(t) = 0$ is posterior self-consistency, $g_M(t) = 1$ means that the head's excess risk equals the irreducible posterior risk, and $g_M(t) > 1$ indicates a still larger mismatch. Near the clean endpoint the denominator can become small; we omit levels with $\sigma_t < 0.2$, where the measured denominator falls below 0.02 per dimension (no deployed level is affected).

Estimator stability. For each model, we train two posterior estimators from independent initializations and report their normalized disagreement

$$d_M(t) = \frac{\mathbb{E}\|\hat{\mu}_M^{(1)} - \hat{\mu}_M^{(2)}\|_2^2}{\mathbb{E}\|X_0^M - \hat{\mu}_M^{(1)}\|_2^2}. \tag{52}$$

This is an estimator-stability diagnostic rather than a formal upper bound on posterior-estimation error.

Four-step student versus long-step control. We evaluate the four-step student and its 50-step teacher at the same four nonterminal noise levels from the student's deployed schedule. Each $\hat{\mu}_M$ is fitted on disjoint samples from its own model-induced clean distribution, and the rollout and prediction query use the same CFG policy within each model. Figure 4(a) reports $g_M(t)$ against the forward-noise level σ_t.

The four-step head has a substantially larger posterior-mean mismatch at every shared noise level: its normalized gap is 0.33–0.52, compared with 0.011–0.012 for the 50-step teacher. Equivalently, the mismatch is 29–$50\times$ larger across the four levels. This comparison diagnoses a difference in output semantics between the coarse student head and the teacher predictor; it does not by itself attribute the difference causally to the number of sampling steps. Posterior-estimator disagreement is 0.015 for the student and 0.008 for the teacher.

Paired-training consequence. To test whether replacing the rollout-induced posterior has a downstream effect, we train DM-full and DM-self from the same four-step checkpoint using group=24, K=4 and 1000 rounds. The two methods initially improve at similar rates, but their measured reward trajectories subsequently separate: DM-full continues to improve and stabilizes, whereas DM-self loses part of its initial gain. At the final round, their trained multi-reward averages are 1.568 and 1.441, respectively. Together with the exact equivalence above, this result shows that the separately estimated posterior target matters beyond the DM-self/matched-AF special case. Because $g_M(t)$ is not measured at every training round, we interpret panel (b) as the consequence of replacing the target, rather than assigning the precise separation time causally to posterior drift.

F.5 Additional Local Response Diagnostics

We provide the full results underlying Section 3.4, together with supplementary evaluations on alternative time grids. All diagnostics use the group-standardized linear advantage $A(x_0) = \langle x_0, u \rangle$, with ideal first-order response 1.

First-order response and extrapolation. We evaluate the deployed response $S(\eta)$ defined in Section 3.4 over $\eta \in [0.025, 0.2]$. The full-range linear fit $S(\eta) = b_0 + b_1\eta$ gives $b_0 = 0.86 \pm 0.04$,

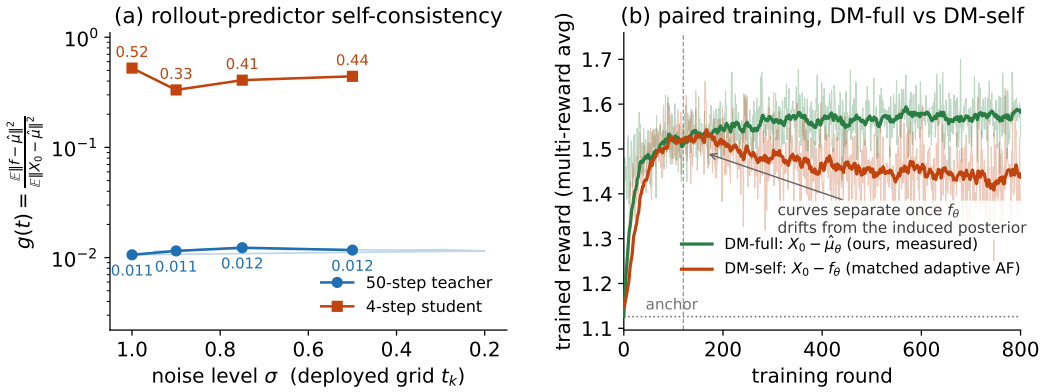


Figure 4: Posterior-mean mismatch and its downstream consequence. (a) Normalized head–posterior gap $g_M(t)$ at the four-step student's deployed nonterminal noise levels. The four-step student has $g_4(t) \in [0.33, 0.52]$, whereas its 50-step teacher has $g_{50}(t) \in [0.011, 0.012]$, a 30–50$\times$ separation across levels. Both posterior estimators are trained on disjoint rollout data and evaluated on held-out prompts, rollouts, and forward noise; per model we use 48 prompts (36 fit, 12 held out) with 32 rollouts each, and every level is evaluated with 4 independent forward-noise draws over the 384 held-out rollouts. (b) Paired reward training from the same four-step checkpoint. DM-full uses the measured posterior residual $X_0 - \hat{\mu}_\theta$, whereas DM-self uses $X_0 - f_\theta$ and is exactly the matched adaptive AdvantageFlow control under Proposition 3. Light curves are per-round rewards and dark curves are exponential moving averages with a 25-round span. The horizontal dotted line is the initial reward, and the vertical dashed line marks the round at which the two trajectories begin to separate (round 120).

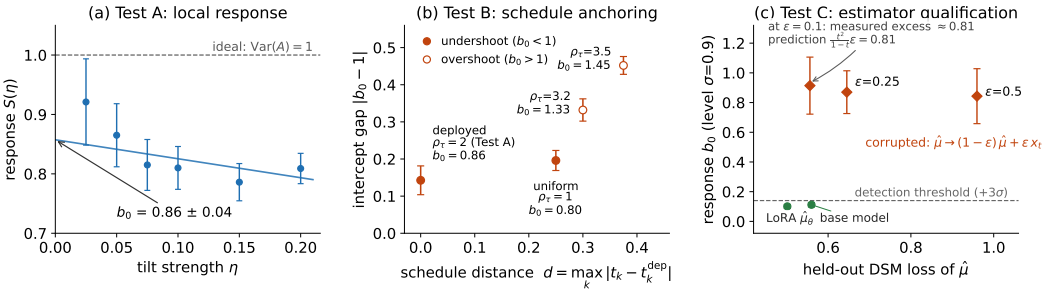


Figure 5: Local response diagnostics. (a) The deployed response extrapolates to $b_0 = 0.86 \pm 0.04$ (ideal: 1). (b) Response error grows under large schedule shifts. (c) Synthetic posterior bias is detected, while swapping trained denoisers has little effect.

corresponding to a response gap $|b_0 - 1| = 0.14 \pm 0.04$ (Fig. 5a). Restricting the fit to smaller η or allowing curvature gives intercepts up to 0.97. The main text reports the full-range linear fit, which provides the more conservative estimate of response recovery.

Off-schedule evaluations. We additionally evaluate the same student on alternative time grids. Grid deviations are measured by $d = \max_k |t_k - t_k^{\text{dep}}|$. We compare the deployed shifted grid ($\rho_\tau = 2$), a uniform grid ($\rho_\tau = 1$), and two heavier shifts ($\rho_\tau = 3.2, 3.5$), where ρ_τ is the max-to-mean interval ratio. The uniform grid gives a comparable response gap (0.20 ± 0.03 versus 0.14 ± 0.04 on the deployed grid). The heavier shifts increase the gap to 0.33 ± 0.03 and 0.45 ± 0.02, both more than 3σ above the deployed result, and produce overshoot ($b_0 > 1$; Fig. 5b). These evaluations characterize behavior outside the student's fixed training schedule. Since changing the grid also changes the query states and transition intervals, the resulting degradation does not isolate the finite-step residual in Theorem 1.

Sensitivity to posterior error. At $\sigma_t = 0.9$, we compare a denoiser swap with a controlled posterior bias at similar increases in held-out denoising loss. Replacing the trained LoRA denoiser with the

base model increases loss by 12%, while response errors remain comparable (0.11 ± 0.02 versus 0.10 ± 0.01). We then inject $\hat{\mu}_\theta \to (1-\epsilon)\hat{\mu}_\theta + \epsilon x_t$ with $\epsilon = 0.1$. This increases denoising loss by 11% but yields a ninefold response error, with a 4σ separation from baseline (Fig. 5c). The excess error agrees with the closed-form prediction: 0.81 measured and 0.81 predicted. The similar loss increases therefore have substantially different effects on the reward response. In this setting, denoising loss alone does not determine the impact of posterior estimation error, whereas the controlled bias produces the quantitatively predicted change.