香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Reinforcing Diffusion Models by Instantaneous Trajectory Distribution Matching

Yihong Luo[1]
[1]Hong Kong University of Science and Technology

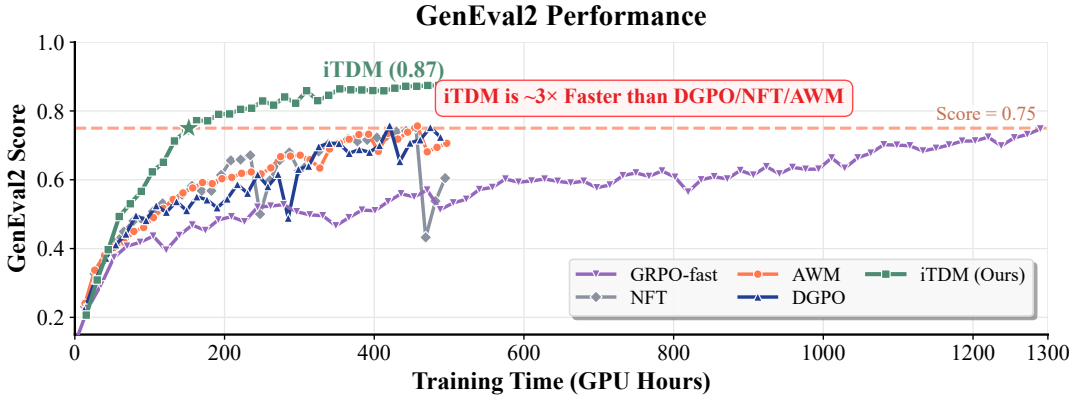

Figure 1 | **GenEval2 performance on SD3.5-M.** Our **iTDM** is nearly 3× faster than DGPO, DiffusionNFT, and AWM, while continuously scaling to **0.87** without training collapse.

Forward diffusion reinforcement learning (RL)—enabling deterministic ODE sampling without trajectory backpropagation—has grown increasingly popular for efficiently aligning diffusion models. However, we reveal that existing methods essentially rely on signed reward-weighted regression (RWR), regressing the *trainable policy* velocity onto rollouts from a *behavior distribution*. Unbiased baseline subtraction implicitly assumes that the policy velocity equals the behavior marginal velocity—an idealized assumption violated under classifier-free guidance (CFG) rollouts, offline/off-policy updates, and imperfect online self-consistency. This *behavior–policy mismatch* breaks baseline invariance and injects *baseline-induced gradient drift*, destabilizing training. To resolve this issue, we propose Instantaneous Trajectory Distribution Matching (iTDM), taking the infinitesimal limit of Trajectory Distribution Matching toward a time-local reward-guided target via the posterior-score identity to yield a sample-level gradient anchored at the behavior velocity without learning an additional reward-tilted target score. Tracking this velocity with an online auxiliary denoiser guarantees exact baseline invariance at the auxiliary optimum, effectively suppressing baseline drift across both online and offline RL. Extensive experiments across compositional generation, text rendering, and preference alignment show that iTDM achieves state-of-the-art in-domain and out-of-domain performance. Notably, on the challenging GenEval2 benchmark, iTDM converges nearly 3× faster than competing methods while scaling smoothly to 0.87 without collapse. Code is available at `https://github.com/Luo-Yihong/iTDM`.

1. Introduction

Reinforcement learning (RL) has become a cornerstone for post-training generative models, advancing large language models (Bai et al., 2022; DeepSeek-AI, 2025; Shao et al., 2024; Ziegler et al., 2019) and empowering diffusion and flow-matching models in compositional generation, text rendering, and preference alignment (Black et al., 2023; Liu et al., 2025; Wallace et al., 2024; Xue et al., 2025b). Early online diffusion RL methods (Liu et al., 2025; Xue et al., 2025b) apply trajectory-level



Figure 2 | **Qualitative comparisons on human preference alignment (PickScore) with SD3.5-M.** All models are trained and evaluated under CFG = 1.0. Compared with competing forward diffusion RL baselines, **iTDM** faithfully follows intricate prompt details while delivering sharper textures and superior aesthetic quality, **avoiding blurry artifacts**.

policy gradients via stochastic differential equation (SDE) samplers, sacrificing the efficiency and fidelity of deterministic ordinary differential equation (ODE) solvers while incurring heavy trajectory-backpropagation overhead. Consequently, *forward diffusion RL* (Luo et al., 2025d; Xue et al., 2025a; Zheng et al., 2025)—which collects clean rollouts via deterministic ODE solvers and optimizes policies along the forward noising path—has gained popularity for its training efficiency.

In GRPO-style post-training (Shao et al., 2024), subtracting a prompt-level baseline $b(\mathbf{c})$ to form centered advantages ($A = R - b(\mathbf{c})$) is essential for variance reduction and relative preference learning. However, we reveal that existing forward diffusion RL methods (Luo et al., 2025d; Xue et al., 2025a; Zheng et al., 2025) essentially rely on signed reward-weighted regression (RWR): the rollout sampler defines a *behavior distribution* q_0 with marginal velocity $v_q(\mathbf{x}_t, t)$, whereas RWR evaluates its denoising residual at the *trainable policy* velocity $v_\theta(\mathbf{x}_t, t)$. *Unbiased baseline subtraction therefore implicitly rests on an idealized assumption:* $v_\theta = v_q$. Yet this assumption fails in practice under classifier-free guidance (CFG) rollouts, offline/off-policy updates, and imperfect online self-consistency, creating pervasive *behavior–policy mismatch* ($v_\theta \neq v_q$) with two severe side effects: (1) *Baseline-induced gradient drift*: non-zero residual mean ($v_\theta - v_q \neq 0$) breaks baseline invariance and injects a spurious drift proportional to $b(\mathbf{c})(v_\theta - v_q)$, causing unanchored RWR to collapse and existing methods to stagnate or diverge under CFG-guided and offline RL; and (2) *Coarse rollout-artifact fitting*: RWR's inherent pull toward q_t drives v_θ to fit few-step ODE discretization artifacts, degrading textures and out-of-domain fidelity (Figs. 2 and 4).

We argue that these failures stem not from centered advantages, but from forcing distributional policy improvement into a signed regression tied to the fragile behavior-policy consistency assumption ($v_\theta = v_q$). Based on this insight, an ideal forward diffusion RL method should tolerate the effects of behavior–policy mismatch without drifting toward coarse few-step rollouts.

To this end, we propose **I**nstantaneous **T**rajectory **D**istribution **M**atching (**iTDM**), a principled distribution-matching framework for forward diffusion RL. Building on Trajectory Distribution Matching (TDM) (Luo et al., 2025c), iTDM takes its *infinitesimal limit* to match a reward-guided target distribution $\pi_t^R \propto q_t \exp(m_t)$ at each noise level. Via the Gaussian posterior-score identity, iTDM

converts the intractable target score into an exact sample-level RL gradient anchored at the behavior velocity v_q, without learning an additional reward-tilted target score. Theoretically, iTDM ensures *exact baseline invariance* even when $v_\theta \neq v_q$. Practically, we track v_q online with an auxiliary denoiser v_ϕ, guaranteeing exact baseline invariance at the auxiliary optimum and effectively suppressing baseline drift. Combined with pretrained reference regularization v_{ref}, iTDM mitigates rollout-artifact fitting and yields a unique stationary policy independent of the relationship between q_t and v_θ, naturally unifying online and offline RL.

This design brings two key benefits: (1) **Principled Drift and Artifact Resistance**, guaranteeing baseline invariance at the auxiliary optimum while decoupling regularization from coarse rollouts; (2) **Stable Online and Offline Reinforcement Post-Training** across guided, unguided, group-wise, and pairwise setups. Extensive experiments on SD3.5-M (Esser et al., 2024) across compositional generation, text rendering, and preference alignment validate iTDM's effectiveness: it converges nearly 3× faster to **0.87** on GenEval2 without collapse (Fig. 1), unlocks stable CFG-guided online RL with leading DrawBench out-of-domain quality, and overcomes offline collapse while outperforming Diffusion-DPO (Wallace et al., 2024) in the pairwise offline setting ($G = 2$).

2. Background

Diffusion Models and Flow Matching. Given a clean latent distribution q_0 (conditioned on a prompt $\mathbf{c}$, omitted when unambiguous) and independent Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the linear forward interpolation path at noise level $t \in (0, 1)$ and its sample-wise conditional velocity are

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\epsilon, \quad v(\mathbf{x}_t \mid \mathbf{x}_0) := \frac{d\mathbf{x}_t}{dt} = \epsilon - \mathbf{x}_0. \tag{1}$$

Let q_t denote the marginal distribution of $\mathbf{x}_t$. Regressing onto $v(\mathbf{x}_t \mid \mathbf{x}_0)$ yields the marginal velocity $v_q(\mathbf{x}, t) = \mathbb{E}[v(\mathbf{x}_t \mid \mathbf{x}_0) \mid \mathbf{x}_t = \mathbf{x}]$, linked to the marginal score via:

$$s_{q_t}(\mathbf{x}) = -\frac{\mathbf{x} + (1-t)v_q(\mathbf{x}, t)}{t}. \tag{2}$$

Policy Gradient and Baseline Invariance. In policy-based RL such as REINFORCE, PPO (Schulman et al., 2017), and GRPO (Shao et al., 2024), a tractable stochastic policy $\pi_\theta(\mathbf{x}_0 \mid \mathbf{c})$ maximizes expected reward $\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim \pi_\theta(\cdot|\mathbf{c})}[r(\mathbf{x}_0, \mathbf{c})]$. A foundational property of policy gradients is *baseline invariance*: subtracting any prompt-dependent baseline $b(\mathbf{c})$ to form centered advantages $A(\mathbf{x}_0, \mathbf{c}) = r(\mathbf{x}_0, \mathbf{c}) - b(\mathbf{c})$ leaves the population gradient strictly unchanged:

$$\mathbb{E}_{\mathbf{x}_0 \sim \pi_\theta(\cdot|\mathbf{c})}\Big[\big(r(\mathbf{x}_0, \mathbf{c}) - b(\mathbf{c})\big)\nabla_\theta \log \pi_\theta(\mathbf{x}_0 \mid \mathbf{c})\Big] = \nabla_\theta \mathcal{J}(\theta). \tag{3}$$

This exact baseline invariance legitimizes advantage centering in PPO and GRPO as a pure variance-reduction mechanism without biasing the optimization direction.

Forward Diffusion RL. To bypass costly trajectory-level policy gradients, forward diffusion RL methods (Luo et al., 2025d; Xue et al., 2025a; Zheng et al., 2025) collect clean rollouts $\mathbf{x}_0 \sim q_0$ from a *behavior distribution* (e.g., an online rollout policy p_{θ^-} or frozen generator) via deterministic ODE solvers and optimize the *trainable policy* velocity v_θ directly along the forward noising path. Up to gradient equivalence, their objectives unify as

$$\mathcal{L}_{\text{FwdRL}}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim q_0,\, t, \epsilon}\Big[\, \text{sg}(\omega)\big(r(\mathbf{x}_0, \mathbf{c}) - b(\mathbf{c})\big)\|v(\mathbf{x}_t \mid \mathbf{x}_0) - v_\theta\|_2^2 + \lambda\|v_\theta - v_{\text{ref}}\|_2^2 + \beta\|v_\theta - v_{\text{old}}\|_2^2\Big], \tag{4}$$

where $b(\mathbf{c})$ is a prompt-level baseline, v_{ref} is the frozen pretrained model, and method-specific choices of $\omega = \omega(\mathbf{x}_0, \mathbf{x}_t)$, v_{old}, and (λ, β) are summarized in Tab. 1 (with v_{θ^-} denoting rollout-generating

Table 1 | Unification of existing forward diffusion RL methods.

Method	Weight ω	Anchor v_{old}	Default Regularization
DGPO (Luo et al., 2025d)	group-level sigmoid gate	—	$\lambda \geq 0,\ \beta = 0$
DiffusionNFT (Zheng et al., 2025)	1	v_{θ^-}	$\lambda \geq 0,\ \beta = (1 - r + b)$
AWM (Xue et al., 2025a)	Approx. importance ratio	$v_{\theta_{\text{EMA}}}$	$\lambda \geq 0,\ \beta = 1$

parameters). At their core, existing methods rely on *signed reward-weighted regression* (RWR), $(r-b)\|v(\mathbf{x}_t \mid \mathbf{x}_0) - v_\theta\|_2^2$, paired with different regularizers. While subtracting $b(\mathbf{c})$ is indispensable for variance reduction, ***we reveal that baseline subtraction in signed RWR is generally ill-posed under behavior–policy mismatch ($v_\theta \neq v_q$), violating baseline invariance and biasing the population gradient***.

3. Method

Existing forward diffusion RL methods rely on signed RWR (Eq. (4)), which ties unbiased baseline subtraction to the fragile consistency assumption $v_\theta = v_q$ while pulling v_θ toward coarse rollout marginals. To tolerate behavior–policy mismatch without drifting toward rollout artifacts, we propose **I**nstantaneous **T**rajectory **D**istribution **M**atching (**iTDM**), which formulates forward diffusion RL by taking the *infinitesimal limit* of Trajectory Distribution Matching (TDM) (Luo et al., 2025c). Below, we derive the general instantaneous TDM gradient (Sec. 3.1), construct a time-local reward-guided target that reduces to an exact sample-level update without learning an additional reward-tilted target score (Sec. 3.2), establish exact baseline invariance alongside practical velocity estimation (Sec. 3.3; Algorithm 1), and theoretically compare iTDM with signed RWR (Sec. 3.4), with proofs and scope discussions in Sec. A.

Problem Setup. At each RL step, a rollout sampler (e.g., an online policy with or without CFG, or an offline generator) produces clean samples $\mathbf{x}_0$ under prompt $\mathbf{c}$, inducing a *behavior distribution* q_0 (held fixed with respect to θ within each update) with bounded nonnegative reward $R = r(\mathbf{x}_0, \mathbf{c})$. Along the forward path in Eq. (1), $v_q(\mathbf{x}, t) = \mathbb{E}[v(\mathbf{x}_t \mid \mathbf{x}_0) \mid \mathbf{x}_t = \mathbf{x}]$ is the behavior marginal velocity and $m_t(\mathbf{x}) = \mathbb{E}[R \mid \mathbf{x}_t = \mathbf{x}]$ is the conditional expected reward. Crucially, the *trainable policy* velocity $v_\theta(\mathbf{x}, t)$ generally differs from $v_q(\mathbf{x}, t)$ ($v_\theta \neq v_q$)—a *behavior–policy mismatch* arising under CFG-guided rollouts, offline or off-policy updates, and imperfect online self-consistency.

3.1. Instantaneous Trajectory Distribution Matching

To perform forward diffusion RL without backpropagating through multi-step solvers, we consider a local reverse transition from t to $s = t - \delta$ ($\delta > 0$). Following TDM (Luo et al., 2025c), for fixed (t, δ) we define the one-step transport map T_θ and its pushforward marginal q_s^θ:

$$T_\theta(\mathbf{x}) = \mathbf{x} - \delta v_\theta(\mathbf{x}, t), \qquad q_s^\theta = (T_\theta)_\# q_t. \tag{5}$$

For any smooth target marginal $p_s^\star$ held fixed during the update, minimizing the $1/\delta$-scaled reverse KL objective $\mathcal{K}_\delta(\theta) := \frac{1}{\delta}\,\mathrm{KL}(q_s^\theta \| p_s^\star)$ with $\mathbf{x}_s = T_\theta(\mathbf{x}_t)$ yields the pathwise transport gradient:

$$\nabla_\theta \mathcal{K}_\delta(\theta) = \frac{1}{\delta}\mathbb{E}_{q_t}\left[\left(\frac{\partial \mathbf{x}_s}{\partial \theta}\right)^\top \left(s_{q_s^\theta}(\mathbf{x}_s) - s_{p_s^\star}(\mathbf{x}_s)\right)\right]. \tag{6}$$

Since $\frac{\partial \mathbf{x}_s}{\partial \theta} = -\delta J_\theta$ at fixed $\mathbf{x}_t$, where $J_\theta := \frac{\partial v_\theta(\mathbf{x}_t,t)}{\partial \theta}$ denotes the policy Jacobian, taking the infinitesimal limit $\delta \downarrow 0$ yields the ***instantaneous TDM*** gradient:

$$\boxed{g_t^\star := \lim_{\delta \downarrow 0} \nabla_\theta \mathcal{K}_\delta(\theta) = \mathbb{E}_{q_t}\big[J_\theta^\top (s_{p_t^\star} - s_{q_t})\big].} \tag{7}$$

Here arguments $(\mathbf{x}_t, t)$ are suppressed when unambiguous. With Eq. (7) adapting TDM to standard multi-step diffusion and flow-matching models, the key remaining challenge for RL is specifying a tractable reward-guided target score $s_{p_t^\star}$.

3.2. Reward-Guided Instantaneous Matching

We now instantiate $p_t^\star$ for reinforcement learning. At each noise level t, we set $p_t^\star$ to the time-local reward-guided target π_t^R with score s_t^R:

$$\begin{gathered} \pi_t^R(\mathbf{x}) = \frac{q_t(\mathbf{x}) e^{m_t(\mathbf{x})}}{Z_t}, \quad Z_t = \mathbb{E}_{q_t}[e^{m_t(\mathbf{x}_t)}], \\ \boxed{s_t^R(\mathbf{x}) := s_{\pi_t^R}(\mathbf{x}) = s_{q_t}(\mathbf{x}) + \nabla_{\mathbf{x}} m_t(\mathbf{x}).} \end{gathered} \tag{8}$$

With bounded rewards, π_t^R uniquely solves the time-local KL-regularized variational problem $\max_p \{\mathbb{E}_p[m_t] - \mathrm{KL}(p\|q_t)\}$, where $\nabla_{\mathbf{x}} m_t$ serves as the reward-driven score correction.

Sample-Level RL Gradient. Directly evaluating $s_t^R(\mathbf{x})$ is intractable, as $m_t(\mathbf{x}) = \mathbb{E}[R \mid \mathbf{x}_t = \mathbf{x}]$ is an unknown conditional expectation over clean latents that cannot be differentiated directly. Fortunately, $\nabla_{\mathbf{x}} m_t(\mathbf{x})$ admits an exact conditional expectation form:

Lemma 1 (Posterior-score identity for reward correction)**.** *Let $q_t(\mathbf{x} \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}; (1-t)\mathbf{x}_0, t^2\mathbf{I})$ denote the Gaussian transition kernel of Eq. (1). The reward-driven score correction $\nabla_{\mathbf{x}} m_t(\mathbf{x})$ satisfies*

$$\begin{aligned} \nabla_{\mathbf{x}} m_t(\mathbf{x}) &= \mathbb{E}\big[R\big(\nabla_{\mathbf{x}} \log q_t(\mathbf{x} \mid \mathbf{x}_0) - s_{q_t}(\mathbf{x})\big) \mid \mathbf{x}_t = \mathbf{x}\big] \\ &= \frac{1-t}{t} \mathbb{E}\big[R\big(v_q(\mathbf{x}_t, t) - v(\mathbf{x}_t \mid \mathbf{x}_0)\big) \mid \mathbf{x}_t = \mathbf{x}\big]. \end{aligned} \tag{9}$$

Yet Lemma 1 alone is not immediately practical: substituting $s_t^R(\mathbf{x}_t)$ into Eq. (7) still involves a conditional expectation given $\mathbf{x}_t$, which would ordinarily require learning an additional reward-tilted target score network onto Eq. (9). Crucially, in the instantaneous limit, the iTDM gradient reduces directly to a tractable sample-level update without learning an additional reward-tilted target score (proof in Sec. A.1):

Proposition 1 (Sample-level gradient of reward-guided instantaneous matching)**.** *Under the regularity assumptions in Sec. A.1, the time-weighted instantaneous gradient for the reward-guided target π_t^R in Eq. (8) satisfies*

$$\begin{aligned} \boxed{g_t^R &:= \frac{2t}{1-t} \lim_{\delta \downarrow 0} \nabla_\theta \mathcal{K}_\delta(\theta) \\ &= 2\mathbb{E}\big[R\, J_\theta^\top (v_q - v(\mathbf{x}_t \mid \mathbf{x}_0))\big].} \end{aligned} \tag{10}$$

Remark 1 (Sample-level target estimation does not extend to few-step TDM). Because Proposition 1 avoids learning an additional reward-tilted target score via a simple sample-level update, one might ask whether few-step TDM (Luo et al., 2025c, 2026a,b) admits the same trick. Crucially, this reduction holds *only* in the instantaneous limit: with $\mathbf{x}_t$ held fixed, J_θ depends solely on $\mathbf{x}_t$, allowing the outer expectation over q_t to absorb the conditional expectation given $\mathbf{x}_t$ via the tower property. In few-step generation, however, the generator Jacobian at a re-noised state remains coupled with the clean sample prediction, invalidating single-sample substitution (see Sec. A.1).

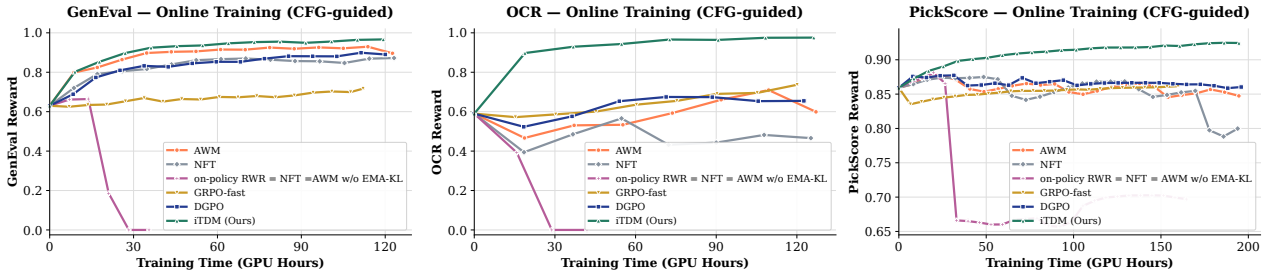

Figure 3 | **Online reinforcement post-training efficiency with CFG-Guided policy.** Note that **on-policy RWR** (equal to on-policy AWM without EMA anchors and on-policy NFT) rapidly stagnates or degrades under baseline-induced drift, whereas **iTDM** converges substantially faster and reaches higher asymptotic rewards without late-stage collapse.

Regularized iTDM Objective and Optimal Solution. For reinforcement post-training, we combine g_t^R with reference-velocity regularization weighted by $\beta > 0$ to form the iTDM-RL gradient:

$$g_t^{\text{iTDM}} = g_t^R + \beta \nabla_\theta \mathbb{E}_{q_t}\left[\left\| v_\theta(\mathbf{x}_t, t) - v_{\text{ref}}(\mathbf{x}_t, t)\right\|_2^2\right]. \tag{11}$$

Lemma 2 (Optimal policy and time-local policy improvement)**.** *For any behavior marginal q_t with full support, frozen reference marginal $p_{\text{ref},t}$ with score s_t^{ref}, and $\beta > 0$, the unique pointwise stationary velocity $v^\star$ of Eq. (11) and its score $s_t^\star$ are*

$$\begin{aligned} v^\star(\mathbf{x}, t) &= v_{\text{ref}}(\mathbf{x}, t) - \frac{t}{\beta(1-t)} \nabla_{\mathbf{x}} m_t(\mathbf{x}), \\ s_t^\star(\mathbf{x}) &= s_t^{\text{ref}}(\mathbf{x}) + \frac{1}{\beta} \nabla_{\mathbf{x}} m_t(\mathbf{x}). \end{aligned} \tag{12}$$

Moreover, $s_t^\star = \nabla_{\mathbf{x}} \log \pi_t^+$ for $\pi_t^+(\mathbf{x}) = p_{\text{ref},t}(\mathbf{x}) e^{m_t(\mathbf{x})/\beta} / Z_t^\beta$ with $Z_t^\beta = \mathbb{E}_{p_{\text{ref},t}}[e^{m_t(\mathbf{x}_t)/\beta}]$. Defining $\mathcal{J}_t(p) := \mathbb{E}_p[m_t] - \beta\, \mathrm{KL}(p \| p_{\text{ref},t})$ with m_t held fixed, if $\mathrm{KL}(q_t \| p_{\text{ref},t}) < \infty$, then π_t^+ achieves the exact improvement

$$\mathcal{J}_t(\pi_t^+) - \mathcal{J}_t(q_t) = \beta\, \mathrm{KL}(q_t \| \pi_t^+) \geq 0, \tag{13}$$

which is strict whenever $q_t \neq \pi_t^+$.

Lemma 2 reveals two structural advantages of regularized iTDM: first, the stationary target π_t^+ strictly improves the time-local regularized reward over q_t whenever $q_t \neq \pi_t^+$; second, $s_t^\star$ is anchored at s_t^{ref} and its form is independent of the relationship between the behavior distribution q_t and trainable policy v_θ. Consequently, ***regularized iTDM naturally unifies online and offline RL***.

3.3. Baseline Invariance and Practical Learning

Following GRPO (Liu et al., 2025; Shao et al., 2024), for each prompt $\mathbf{c}$ we sample a group of G clean rollouts $\{\mathbf{x}_{0,i}\}_{i=1}^G \sim q_0$ and subtract a prompt-dependent baseline $b(\mathbf{c})$ to form centered advantages $A = R - b(\mathbf{c})$ for variance reduction and relative preference learning. A key question is whether subtracting $b(\mathbf{c})$ alters the population iTDM gradient.

Lemma 3 (Exact baseline invariance)**.** *For any prompt-dependent baseline $b(\mathbf{c})$, subtracting $b(\mathbf{c})$ leaves the population iTDM gradient strictly invariant:*

$$\boxed{g_t^A := 2\mathbb{E}\left[A\, J_\theta^\top \left(v_q - v(\mathbf{x}_t \mid \mathbf{x}_0)\right)\right] = g_t^R.} \tag{14}$$

Algorithm 1 iTDM for Forward Diffusion RL

Require: Policy v_θ, auxiliary v_ϕ, reference v_{ref}, reward r, prompts $\mathcal{D}_c$, group size G, weight β.

for $n \leftarrow 1$ **to** N **do**
  Sample $\mathbf{c} \sim \mathcal{D}_c$ and $\{\mathbf{x}_{0,i}\}_{i=1}^{G} \sim q_0(\cdot \mid \mathbf{c})$; set $R_i \leftarrow r(\mathbf{x}_{0,i}, \mathbf{c})$, $A_i \leftarrow \text{clip}(\widehat{\text{Adv}}_i, -5, 5)/5$
  Sample $t \sim \mu$, $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_{t,i} \leftarrow (1-t)\mathbf{x}_{0,i} + t\epsilon_i$
  Update ϕ by minimizing $\mathcal{L}_{\text{aux}}(\phi)$ in Eq. (15)
  Update θ along $-g_t^{\text{iTDM}}$ (Eq. (11)) with $g_t^R \leftarrow \widehat{g}_t^A$ using $\text{sg}(v_\phi)$
end for

Crucially, ***exact baseline invariance holds even under behavior–policy mismatch*** ($v_\theta \neq v_q$). This yields two practical benefits: 1) **Stable Offline and Off-Policy Training**: it natively supports unbiased advantage centering under static datasets, EMA-lagged or multi-epoch replay rollouts, and CFG-guided sampling; and 2) **Robustness to Imperfect Online Self-Consistency**: even in purely online rollouts generated by v_θ, the forward marginal velocity v_q generally differs from v_θ before full convergence to the optimal velocity field (Karras et al., 2022; Liu et al., 2023; Song et al., 2021).

Learning Auxiliary Velocity v_q. Since the ground-truth behavior velocity v_q is not analytically available, we introduce an online auxiliary denoiser v_ϕ to estimate v_q, following score distillation (Luo et al., 2025c, 2026b; Yin et al., 2024a). Noisy states $\mathbf{x}_t$ are constructed by re-noising terminal rollouts $\mathbf{x}_0 \sim q_0$ via Eq. (1) without storing intermediate solver trajectories, and v_ϕ (parameterized as a lightweight LoRA adapter) minimizes the standard denoising loss:

$$\mathcal{L}_{\text{aux}}(\phi) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\|v_\phi(\mathbf{x}_t, t) - v(\mathbf{x}_t \mid \mathbf{x}_0)\right\|_2^2\right], \tag{15}$$

whose unique minimizer is $v_\phi = v_q$. In practice, we substitute the stop-gradient prediction v_ϕ for v_q, yielding the practical gradient $\widehat{g}_t^A = 2\mathbb{E}[A\, J_\theta^\top (v_\phi - v(\mathbf{x}_t \mid \mathbf{x}_0))]$, and center rewards with the group-mean baseline $b(\mathbf{c}) = \frac{1}{G}\sum_{i=1}^{G} R_i$, combining $\widehat{g}_t^A$ with the reference regularization in Eq. (11) to optimize v_θ (Algorithm 1).

3.4. Theoretical Analysis

To isolate the core mechanism shared by existing forward diffusion RL methods (Eq. (4)), we consider a *fully online* setting ($\omega = 1$) and analyze the core RWR objective at noise level t with weight $w \in \{R, A\}$:

$$\mathcal{L}_{\text{RWR}}^{w}(\theta) = \mathbb{E}\left[w \left\|v(\mathbf{x}_t \mid \mathbf{x}_0) - v_\theta(\mathbf{x}_t, t)\right\|_2^2\right]. \tag{16}$$

Whereas nonnegative weighting ($w = R \geq 0$) admits a well-defined pointwise minimizer $\mathbb{E}[R\, v(\mathbf{x}_t \mid \mathbf{x}_0) \mid \mathbf{x}_t = \mathbf{x}]/m_t(\mathbf{x})$, centering rewards into signed advantages $w = A = R - b(\mathbf{c})$ renders $\mathcal{L}_{\text{RWR}}^{A}$ unbounded below in general. To examine whether baseline subtraction remains valid in RWR and clarify what signed RWR optimizes, we characterize its population gradient:

Proposition 2 (Baseline non-invariance and gradient decomposition of RWR)**.** *For any prompt-dependent baseline $b(\mathbf{c})$ and centered advantage $A = R - b(\mathbf{c})$, the population gradient of RWR in Eq. (16) satisfies:*

*1) **Violation of baseline invariance:** Subtracting $b(\mathbf{c})$ shifts the population gradient of RWR by*

$$\boxed{\nabla_\theta \mathcal{L}_{\text{RWR}}^{A} - \nabla_\theta \mathcal{L}_{\text{RWR}}^{R} = -\Delta_t^b}, \tag{17}$$

where $\Delta_t^b := 2b(\mathbf{c})\, \mathbb{E}_{q_t}\left[J_\theta^\top (v_\theta - v_q)\right]$ is the baseline-induced drift.

2) ***Gradient decomposition:*** *The population gradient of* $\mathcal{L}^A_{\text{RWR}}$ *decomposes into*

$$\nabla_\theta \mathcal{L}^A_{\text{RWR}} = \underbrace{g^A_t}_{\textit{iTDM gradient}} + \underbrace{2\mathbb{E}_{q_t}\big[m_t\, J_\theta^\top (v_\theta - v_q)\big]}_{\Delta^q_t\textit{: drift toward } q_t} - \Delta^b_t. \tag{18}$$

Side Effects of Signed RWR. Proposition 2 isolates two extraneous drift terms in signed RWR beyond g^A_t: 1) ***Baseline-induced gradient drift*** ***($-\Delta^b_t$)***: whenever $v_\theta \neq v_q$, subtracting $b(\mathbf{c}) > 0$ pushes v_θ *away* from v_q during gradient descent, causing unanchored on-policy RWR to collapse (Figs. 3 and 5) and anchored variants to stagnate or diverge under CFG-guided and offline rollouts (Tab. 2 and Fig. 6); and 2) ***Coarse rollout-artifact fitting*** ***(Δ^q_t)***: since $m_t \geq 0$, Δ^q_t pulls v_θ toward v_q, forcing online policies to fit coarse few-step ODE artifacts (Liu et al., 2025; Luo et al., 2025d) and degrading visual and out-of-domain quality even at CFG = 1.0 (Fig. 2 and Tab. 2).

Advantages of iTDM. At the auxiliary optimum ($v_\phi = v_q$), anchoring the residual at v_q while regularizing toward v_{ref} instead of q_t eliminates both $-\Delta^b_t$ and Δ^q_t (Sec. A). This simultaneously insulates v_θ from few-step rollout artifacts for superior visual fidelity (Fig. 2 and Tab. 2) and preserves unbiased advantage centering (Lemma 3) for stable optimization across guided, unguided, and offline regimes without restrictive proximity anchors.

Remark 2. With the learned denoiser v_ϕ, practical baseline invariance holds up to $v_\phi - v_q$ (Sec. A.2). Yet unlike signed RWR—which passively assumes $v_\theta = v_q$ even as mismatch and RL updates violate it—iTDM decouples velocity estimation from policy improvement and *actively minimizes* $\|v_\phi - v_q\|_2^2$ (Eq. (15)). Approximating intractable generator scores with an online auxiliary network has been extensively validated across score distillation (Luo et al., 2025b,c, 2026b; Wang et al., 2023; Yin et al., 2024a,b; Zhou et al., 2024). Despite maintaining an auxiliary adapter, iTDM achieves markedly greater stability, less distributional drift, and faster wall-clock convergence than RWR-based methods (Figs. 1, 3 and 5).

4. Experiments

4.1. Experimental Setup

Evaluation Tasks. We post-train SD3.5-M (Esser et al., 2024) using LoRA adapters (Hu et al., 2022) for both v_θ and v_ϕ across three benchmarks (Liu et al., 2025; Luo et al., 2025d): 1) *Compositional Image Generation* on GenEval (Ghosh et al., 2023) and the harder GenEval2 (Kamath et al., 2025) (Fig. 1); 2) *Visual Text Rendering* (OCR) (Chen et al., 2023; Gong et al., 2025) for character-level accuracy; and 3) *Human Preference Alignment* with PickScore (Kirstain et al., 2023).

Training Setups and Baselines. We evaluate **online RL** under both *CFG-guided* (CFG = 4.5) and *conditional* (CFG = 1.0) policies, alongside frozen-rollout **offline RL** in *group-wise* ($G > 2$) and *pairwise* ($G = 2$) settings. We compare against forward diffusion RL baselines—DGPO (Luo et al., 2025d), DiffusionNFT (NFT) (Zheng et al., 2025), and AWM (Xue et al., 2025a) (plus unanchored **on-policy RWR**)—Diffusion-DPO (DPO) (Wallace et al., 2024) at $G = 2$, and trajectory-based GRPO-fast (Liu et al., 2025) for efficiency reference.

Out-of-Domain Evaluation Metrics. To guard against reward hacking, we evaluate seven unseen quality and preference metrics on DrawBench (Saharia et al., 2022) (see Sec. B for details).

4.2. Online Reinforcement Post-Training

Necessity of CFG-Guided Policy RL. Classifier-free guidance is vital for exploration and visual fidelity: removing CFG sharply degrades base SD3.5-M (0.63 → 0.24 GenEval, 0.59 → 0.12 OCR in Tab. 2),



Figure 4 | **Visual text rendering under conditional vs. CFG-guided online RL.** Without CFG (left), baselines overfit text accuracy by collapsing onto plain, context-free backgrounds; with CFG (right), they preserve scene context but fail to render legible text due to velocity-mismatch drift. **iTDM** excels in both regimes, uniting rich scene fidelity with sharp, accurate text under CFG-guided RL.

and unguided RL (CFG = 1.0) easily sacrifices scene context despite high task rewards (left panel of Fig. 4). Yet reinforcing a CFG-guided policy stalls RWR baselines due to persistent behavior–policy mismatch (Proposition 2), yielding blurry or illegible text (0.70–0.76 OCR; right panel of Fig. 4). Tracking v_q via v_ϕ resolves this dilemma and unlocks stable CFG-guided RL.

Quantitative Results. In Tab. 2, under the **CFG-guided policy** (CFG = 4.5, ✓), iTDM leads all benchmarks (**0.97** GenEval, **0.96** OCR, **24.08** PickScore) while staying on par with base SD3.5-M on DrawBench and sweeping all seven unseen metrics on PickScore (6.11 Aesthetic, 32.73 HPSv2.1, 1.46 ImgRwd). In contrast, RWR baselines trade out-of-domain quality for task reward: on GenEval, NFT and DGPO fall to 22.62 and 23.87 HPSv2.1 (vs. 27.90 for the base model and **28.19** for iTDM), and on OCR, DGPO drops to 19.73 HPSv2.1 and −0.45 ImgRwd despite longer training while still trailing iTDM by 0.20 in OCR accuracy. On PickScore, iTDM surpasses the strongest baseline AWM by 1.28 in-domain (24.08 vs. 22.80) and by 2.95 HPSv2.1 out-of-domain (32.73 vs. 29.78), indicating that its reward gains reflect genuine preference improvement rather than reward hacking. Consistent with Proposition 2, this gap reflects the baseline-induced drift $-\Delta_t^b$ under CFG-guided rollouts.

Under the **conditional policy** (CFG = 1.0, ✗), coarse rollout-artifact fitting (Δ_t^q in Proposition 2) causes RWR baselines on OCR to lose perceptual quality despite 0.97 accuracy (3.52 Aesthetic for NFT; 2.13 DeQA for DGPO), whereas iTDM matches or tops all three tasks (**0.98**, **0.97**, **24.06**) with markedly higher DrawBench fidelity, e.g., **18.25** vs. at most 14.86 HPSv2.1 on OCR and the best DeQA on every task.

Training Cost and Scaling. The lightweight auxiliary LoRA v_ϕ adds negligible overhead, enabling iTDM to converge substantially faster in GPU hours than GRPO-fast and RWR baselines (Figs. 3 and 5). On the harder **GenEval2** benchmark (Fig. 1), ***iTDM reaches*** 0.75 ***nearly*** 3× ***faster than***

Table 2 | **Online reinforcement post-training with guided (CFG = 4.5, ✓) and unguided (CFG = 1.0, ✗) policy.** All methods use the same strength of regularization toward the frozen reference. Best and second-best results in each setting are highlighted in **bold** and underlined, respectively.

CFG	Model	Task Metric			Image Quality		Alignment & Preference				
		GenEval	OCR	PickScore	Aesthetic	DeQA	CLIPScore	HPSv2.1	ImgRwd	PickScore	UniRwd
✓	SD3.5-M†	0.63	0.59	21.72	5.36	—	28.50	27.90	0.85	22.34	3.03
✗	SD3.5-M†	0.24	0.12	19.82	5.13	—	23.70	20.40	-0.58	20.51	2.02
Compositional Image Generation:											
	AWM	0.93	—	—	5.12	3.43	28.96	24.54	0.91	22.14	3.07
	NFT	0.89	—	—	4.92	2.97	28.74	22.62	0.64	21.89	2.97
✓	DGPO	0.90	—	—	5.09	3.17	28.98	23.87	0.75	22.05	2.98
	iTDM (Ours)	**0.97**	—	—	**5.26**	**4.12**	**29.00**	**28.19**	**0.98**	**22.49**	**3.13**
	AWM	0.92	—	—	4.71	3.94	27.50	21.41	0.15	21.35	2.62
	NFT	0.96	—	—	4.88	4.05	27.34	22.90	0.30	21.51	2.82
✗	DGPO	0.97	—	—	**5.15**	4.06	27.64	24.03	**0.71**	**22.04**	2.96
	iTDM (Ours)	**0.98**	—	—	5.08	**4.23**	**27.69**	**24.73**	0.65	21.74	**3.04**
Visual Text Rendering:											
	AWM	—	0.75	—	**5.28**	3.62	28.67	24.77	0.73	22.05	2.93
	NFT	—	0.70	—	5.19	3.67	28.75	25.43	0.68	22.10	2.91
✓	DGPO	—	0.76	—	4.95	2.56	26.82	19.73	-0.45	20.76	2.18
	iTDM (Ours)	—	**0.96**	—	5.23	**3.97**	**28.77**	**27.98**	**0.80**	**22.40**	**3.09**
	AWM	—	0.85	—	4.16	3.25	21.02	13.70	-1.69	19.69	1.52
	NFT	—	**0.97**	—	3.52	3.52	22.13	14.86	-1.49	19.86	1.50
✗	DGPO	—	**0.97**	—	4.19	2.13	21.36	14.16	-1.70	19.69	1.53
	iTDM (Ours)	—	**0.97**	—	**4.48**	**3.88**	**22.80**	**18.25**	**-1.15**	**20.14**	**1.82**
Human Preference Alignment:											
	AWM	—	—	22.80	6.02	3.99	27.73	29.78	1.18	23.23	3.27
	NFT	—	—	22.75	5.96	3.86	27.50	28.72	1.14	23.14	3.25
✓	DGPO	—	—	22.37	5.65	3.98	28.12	28.87	1.10	22.89	3.17
	iTDM (Ours)	—	—	**24.08**	**6.11**	**4.12**	**28.35**	**32.73**	**1.46**	**24.26**	**3.52**
	AWM	—	—	23.94	**6.45**	3.97	26.59	29.99	1.29	23.95	3.34
	NFT	—	—	23.89	6.40	3.81	26.36	29.05	1.24	23.86	3.32
✗	DGPO	—	—	23.85	6.18	3.60	**27.32**	29.21	1.23	23.54	3.35
	iTDM (Ours)	—	—	**24.06**	6.22	**4.26**	26.71	**31.85**	**1.33**	**24.00**	**3.36**

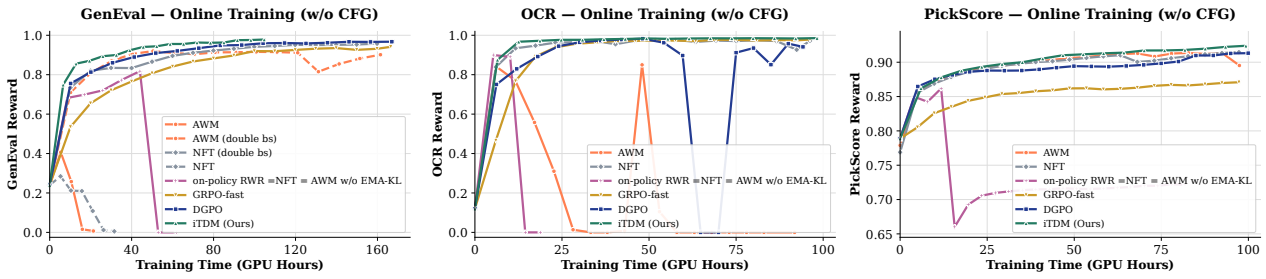

Figure 5 | **Online reinforcement post-training efficiency without guidance (CFG = 1.0).** Note that **on-policy RWR** is equal to on-policy AWM without EMA anchors and on-policy NFT. We double the batch size for AWM and NFT to stabilize training.

DGPO, NFT, and AWM, scaling smoothly to **0.87** without collapse.

Qualitative Comparison. Qualitative comparisons on PickScore (Fig. 2) and OCR (Fig. 4) show that RWR baselines suffer from blurry textures, collapsed backgrounds, or distorted text, whereas iTDM delivers crisp details, rich scene context, and accurate text rendering.

Table 3 | **Offline training across pairwise ($G = 2$) and group-wise settings.** Rollouts are generated by the frozen pretrained SD3.5-M (CFG = 4.5); all methods use the same strength of regularization toward the frozen reference. Best results in each setting are in **bold**, and second-best results in group-wise settings are underlined.

Model	Task Metric			Image Quality		Alignment & Preference				
	GenEval	OCR	PickScore	Aesthetic	DeQA	CLIPScore	HPSv2.1	ImgRwd	PickScore	UniRwd
SD3.5-M†	0.63	0.59	—	5.36	—	28.50	27.90	0.85	22.34	3.03
Compositional Image Generation:										
DPO ($G = 2$)	0.69	—	—	**5.41**	4.03	28.72	**28.55**	0.97	22.35	**3.20**
iTDM (Ours, $G = 2$)	**0.77**	—	—	5.38	**4.06**	**29.11**	28.26	**1.06**	**22.43**	3.18
AWM	0.75	—	—	**5.36**	3.72	29.10	27.66	0.93	21.49	3.14
NFT	0.52	—	—	5.33	3.76	28.30	27.13	0.72	22.00	3.00
DGPO	0.78	—	—	5.33	**3.99**	**29.11**	**28.01**	1.01	22.28	3.21
iTDM (Ours)	**0.87**	—	—	5.32	3.82	28.93	27.98	**1.03**	**22.38**	**3.23**
Visual Text Rendering:										
DPO ($G = 2$)	—	0.78	—	5.23	3.90	**29.44**	**27.99**	0.91	22.13	3.08
iTDM (Ours, $G = 2$)	—	**0.87**	—	**5.29**	**3.95**	29.26	27.57	**1.01**	**22.23**	**3.13**
AWM	—	0.85	—	**5.34**	3.48	28.92	27.64	0.90	21.82	3.01
NFT	—	0.53	—	5.21	3.52	28.33	25.97	0.62	21.74	2.95
DGPO	—	0.87	—	5.18	3.33	**29.41**	26.37	**0.99**	21.90	2.97
iTDM (Ours)	—	**0.92**	—	5.22	**3.68**	29.27	**27.69**	0.95	**22.27**	**3.12**
Human Preference Alignment:										
DPO ($G = 2$)	—	—	21.87	5.46	4.06	28.44	28.77	0.98	22.51	3.16
iTDM (Ours, $G = 2$)	—	—	**22.11**	**5.56**	**4.12**	**28.67**	**30.25**	**1.17**	**22.77**	**3.25**
AWM	—	—	21.71	**5.67**	3.73	27.02	27.55	0.83	22.20	3.09
NFT	—	—	20.17	5.29	3.74	28.23	26.69	0.65	21.96	3.02
DGPO	—	—	22.13	5.59	**4.18**	28.36	30.39	1.16	22.78	3.27
iTDM (Ours)	—	—	**22.26**	5.62	4.15	**28.71**	**30.75**	**1.24**	**22.94**	**3.34**

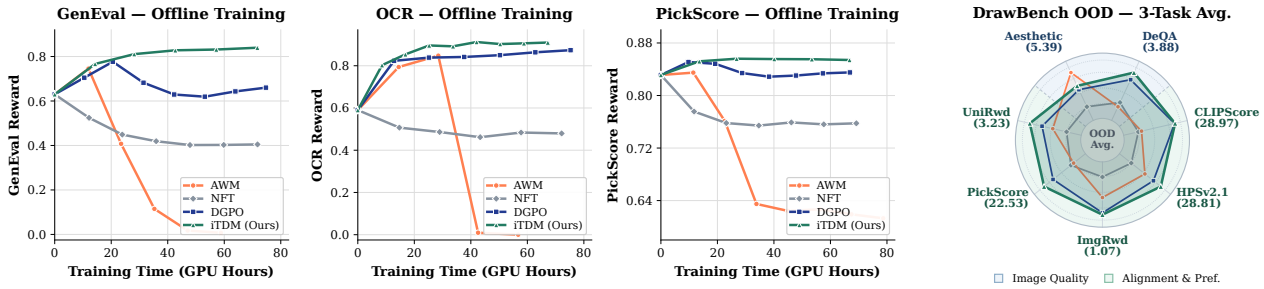

Figure 6 | **Group-wise offline RL with frozen rollouts (CFG = 4.5).** *Left:* Task reward vs. GPU hours. *Right:* Seven unseen DrawBench metrics averaged over the three tasks (parentheses denote **iTDM**). **iTDM** improves monotonically without collapse and leads out-of-domain generalization.

4.3. Offline RL and Diagnostic Analysis

Group-Wise Frozen-Rollout RL. With q_0 frozen to pretrained SD3.5-M (CFG = 4.5), using the initial velocity v_{θ_0} as a static proxy in offline AWM and NFT leaves persistent mismatch ($v_{\theta_0} \neq v_q$; Fig. 6 and Tab. 3). Consequently, NFT collapses below the base model (0.52 vs. 0.63 GenEval; 0.53 vs. 0.59 OCR) and AWM diverges after peaking at 0.75 GenEval. Supported by its decoupled stationary policy (Lemma 2) and baseline invariance at the auxiliary optimum (Lemma 3), iTDM improves monotonically to **0.87** GenEval (+9% over DGPO), **0.92** OCR, and **22.26** PickScore while leading DrawBench generalization.

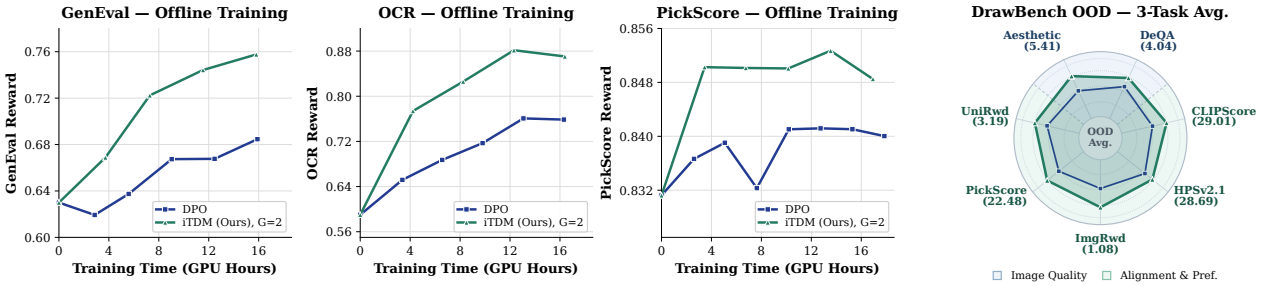

Figure 7 | **Pairwise offline RL ($G = 2$, CFG $= 4.5$) vs. Diffusion-DPO.** *Left:* Task reward vs. GPU hours. *Right:* Three-task average DrawBench radar (parentheses denote **iTDM**). Even at $G = 2$, **iTDM** converges faster and surpasses DPO across all tasks and unseen metrics.

Pairwise RL ($G = 2$) and Group Scaling. Unlike pairwise-only DPO (Wallace et al., 2024), iTDM supports arbitrary group sizes. Under identical pairwise rollouts ($G = 2$, Fig. 7), iTDM converges faster and outperforms DPO on all tasks (**0.77** vs. 0.69 GenEval, **0.87** vs. 0.78 OCR, **22.11** vs. 21.87 PickScore) and all seven unseen metrics. Scaling to larger groups (Fig. 6 and Tab. 3) further boosts performance (0.77 → 0.87 GenEval, 0.87 → 0.92 OCR).

Diagnostic Controls. Figs. 3 and 5 verify the crucial control: removing proximity anchors collapses **on-policy RWR**, confirming that NFT and AWM rely on heuristic anchors to damp baseline drift, whereas iTDM suppresses it via v_ϕ with better rewards and training stability.

5. Related Work

Aligning diffusion and flow-matching models with human preferences or downstream rewards generally follows three paradigms: supervised fine-tuning on curated image–prompt pairs (Dai et al., 2023; Podell et al., 2023), explicit reward maximization via multi-step output backpropagation (Clark et al., 2023; Lee et al., 2023; Luo et al., 2025a,b; Prabhudesai et al., 2023), control/score matching (Domingo-Enrich et al., 2025; Lee et al., 2026; Psenka et al., 2023), or policy gradients (Black et al., 2023; Fan et al., 2023; Ren et al., 2024; Ye et al., 2024), and implicit preference learning such as Diffusion-DPO (Wallace et al., 2024), Diffusion-KTO (Li et al., 2024), and group-enumerated variants (Chen et al., 2025). Recently, Flow-GRPO (Liu et al., 2025) and DanceGRPO (Xue et al., 2025b) adapted GRPO (Shao et al., 2024) to visual generation, inspiring refinements in hybrid sampling, branching, scheduling, and video alignment (He et al., 2025; Li et al., 2025a,b, 2026; Liang et al., 2026; Wang et al., 2026, 2025a,b). However, these trajectory-based methods remain tied to stochastic transition likelihoods and costly multi-step trajectory training.

To enable fast deterministic ODE rollouts without trajectory backpropagation, recent forward diffusion RL methods optimize policies directly on independently re-noised endpoints. Beyond group-level preference optimization in DGPO (Luo et al., 2025d), DiffusionNFT (Zheng et al., 2025), AWM (Xue et al., 2025a), EPG/PEPG (Choi et al., 2026b), and RAM (Bergmeister et al., 2026) (alongside (Xu et al., 2026; Ye et al., 2025)) formulate velocity-space updates, which RVM (Choi et al., 2026a) unifies at the gradient level as reward-weighted velocity regression with anchor regularization. Yet, because these RWR-style losses evaluate the residual at v_θ, under CFG rollouts, offline/off-policy updates, or imperfect online self-consistency ($v_\theta \neq v_q$), subtracting a reward baseline violates baseline invariance and induces population gradient drift (Sec. 3.4). By taking the step-normalized infinitesimal limit of Trajectory Distribution Matching (Luo et al., 2025c), **iTDM** anchors the sample-level residual at the behavior velocity via a lightweight auxiliary denoiser, guaranteeing exact baseline invariance and drift-free optimization at the auxiliary optimum to enable stable online and offline RL.

6. Conclusion

In this work, we revealed that existing forward diffusion RL methods essentially rely on signed RWR, whose implicit $v_\theta = v_q$ assumption breaks down under behavior–policy mismatch to induce baseline gradient drift while pulling policies toward coarse few-step rollout artifacts. To resolve both issues, we introduced **Instantaneous Trajectory Distribution Matching** (**iTDM**), which takes the infinitesimal limit of TDM toward a time-local reward-guided target via the posterior-score identity, yielding an exact sample-level RL gradient anchored at the behavior velocity without learning an additional reward-tilted target score. Paired with a lightweight online auxiliary LoRA denoiser and reference regularization, iTDM guarantees exact baseline invariance at the auxiliary optimum, suppresses drift and artifact fitting, and unifies online and offline RL. Extensive experiments demonstrate up to 3× faster convergence alongside state-of-the-art in-domain and out-of-domain performance across diverse tasks and post-training regimes.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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. *arXiv preprint arXiv:2605.10759*, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Renjie Chen, Wenfeng Lin, Yichen Zhang, Jiangchuan Wei, Boyuan Liu, Chao Feng, Jiao Ran, and Mingyu Guo. Towards self-improvement of diffusion models via group preference optimization, 2025. URL https://arxiv.org/abs/2505.11070.

Jaemoo Choi, Wei Guo, Yuchen Zhu, Arash Vahdat, Molei Tao, Julius Berner, and Yongxin Chen. Scaling reinforcement learning for diffusion models via velocity matching. *arXiv preprint arXiv:2608.23664*, 2026a.

Jaemoo Choi, Yuchen Zhu, Wei Guo, Petr Molodyk, Bo Yuan, Jinbin Bai, Yi Xin, Molei Tao, and Yongxin Chen. Rethinking the design space of reinforcement learning for diffusion models: On the importance of likelihood estimation beyond loss design. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026b.

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

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

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. *Advances in Neural Information Processing Systems*, 36, 2023.

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.

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. TempFlow-GRPO: When timing matters for GRPO in flow models. *arXiv preprint arXiv:2508.04324*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

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. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Jeongjae Lee, Jinho Chang, Jeongsol Kim, and Jong Chul Ye. Reward score matching: Unifying reward-based fine-tuning for flow and diffusion models. *arXiv preprint arXiv:2604.17415*, 2026.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2025a. URL https://arxiv.org/abs/2507.21802v1.

Rui Li, Yuanzhi Liang, Ziqi Ni, Haibin Huang, Chi Zhang, and Xuelong Li. Rethinking reward signals in video GRPO: When scores become targets. *arXiv preprint arXiv:2511.19356*, 2025b.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *Advances in Neural Information Processing Systems*, 37, 2024.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branch-GRPO: Stable and efficient GRPO with structured branching in diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Zhanhao Liang, Tao Yang, Jie Wu, Chengjian Feng, and Liang Zheng. LeapAlign: Post-training flow matching models at any generation step by building two-step trajectories. *arXiv preprint arXiv:2604.15311*, 2026.

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.

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

Yihong Luo, Tianyang Hu, Weijian Luo, Kenji Kawaguchi, and Jing Tang. Reward-instruct: A reward-centric approach to fast photo-realistic image generation. *arXiv preprint arXiv:2503.13070*, 2025a.

Yihong Luo, Tianyang Hu, Yifan Song, Jiacheng Sun, Zhenguo Li, and Jing Tang. Adding additional control to one-step diffusion with joint distribution matching. *arXiv preprint arXiv:2503.06652*, 2025b.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025c. URL https://arxiv.org/abs/2503.06674.

Yihong Luo, Tianyang Hu, and Jing Tang. Reinforcing diffusion models by direct group preference optimization, 2025d. URL https://arxiv.org/abs/2510.08425.

Yihong Luo, Tianyang Hu, Weijian Luo, and Jing Tang. TDM-R1: Reinforcing few-step diffusion models with non-differentiable reward. *arXiv preprint arXiv:2603.07700*, 2026a.

Yihong Luo, Tianyu Pang, and Tianyang Hu. VA-TDM: A unified view for improving few-step diffusion models by velocity adversarial distillation. *arXiv preprint arXiv:2611.06796*, 2026b.

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. *arXiv preprint arXiv:2307.01952*, 2023.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. *arXiv preprint arXiv:2312.11752*, 2023.

Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

Jin Wang, Jianxiang Lu, Guangzheng Xu, Comi Chen, Haoyu Yang, Linqing Wang, Peng Chen, Mingtao Chen, Zhichao Hu, Longhuang Wu, Shuai Shao, Qinglin Lu, and Ping Luo. TAGRPO: Boosting GRPO on image-to-video generation with direct trajectory alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, Meng Wang, Pengfei Wan, and Xiaodan Liang. GRPO-Guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-GRPO: Pairwise preference reward-based GRPO for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023.

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

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023.

Yixian Xu, Yuanrui Zhang, Shengjie Luo, Kaiyuan Gao, Liwei Wang, and Di He. Designing reinforcement learning for diffusion models: A unified path-space view. *arXiv preprint arXiv:2608.14430*, 2026.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models, 2025a. URL https://arxiv.org/abs/2509.25050.

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*, 2025b.

Haotian Ye, Kaiwen Zheng, Jiashu Xu, Puheng Li, Huayu Chen, Jiaqi Han, Sheng Liu, Qinsheng Zhang, Hanzi Mao, Zekun Hao, Prithvijit Chattopadhyay, Dinghao Yang, Liang Feng, Maosheng Liao, Junjie Bai, Ming-Yu Liu, James Zou, and Stefano Ermon. Data-regularized reinforcement learning for diffusion models at scale. *arXiv preprint arXiv:2512.04332*, 2025.

Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation. *arXiv preprint arXiv:2412.01243*, 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6613–6623, 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, 2024b. URL https://arxiv.org/abs/2405.14867.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14483–14494, 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, 2025. URL https://arxiv.org/abs/2509.16117.

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

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Detailed Proofs and Scope Discussion

A.1. Regularity Assumptions and Proofs

Regularity Assumptions. We consider interior times $t \in (0,1)$ with $\mathbb{E}\|\mathbf{x}_0\|_2^2 < \infty$ and bounded rewards, holding q_0, $b(\mathbf{c})$, and ϕ fixed with respect to θ during each update. We assume $v_\theta(\mathbf{x}, t)$ is continuously differentiable in $(\mathbf{x}, \theta)$ such that T_θ in Eq. (5) is a diffeomorphism for sufficiently small $\delta > 0$, that q_s^θ and $p_s^\star$ are positive and smooth with scores converging to s_{q_t} and $s_{p_t^\star}$ as $\delta \downarrow 0$, and that vanishing boundary terms and integrable domination allow interchanging differentiation and limits with expectations.

Proof of Lemma 1, Proposition 1, and Remark 1. The $1/\delta$-weighted reverse-KL transport derivative is given by Eq. (6). Substituting $\frac{\partial \mathbf{x}_s}{\partial \theta} = -\delta J_\theta$ and taking the dominated limit as $\delta \downarrow 0$ proves Eq. (7). Next, for the Gaussian kernel $q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; (1-t)\mathbf{x}_0, t^2\mathbf{I})$ in Eq. (1), its conditional score satisfies

$$\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{\epsilon}{t} = -\frac{\mathbf{x}_t + (1-t)v(\mathbf{x}_t \mid \mathbf{x}_0)}{t}. \tag{19}$$

Differentiating the numerator and denominator of the Bayes representation of $m_t(\mathbf{x}) = \mathbb{E}[R \mid \mathbf{x}_t = \mathbf{x}]$ yields the first line of Eq. (9). Subtracting the marginal score–velocity relation in Sec. 2 from $\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0)$ gives the second line of Eq. (9).

To establish the single-sample reduction in Eq. (10) (Remark 1), note that with $\mathbf{x}_t$ held fixed in the instantaneous limit $\delta \downarrow 0$, the normalized derivative $\delta^{-1}\partial \mathbf{x}_s / \partial \theta = -J_\theta$ is measurable with respect to $\mathbf{x}_t$. By the tower property of conditional expectation, for any integrable $f(\mathbf{x}_t, \mathbf{x}_0)$,

$$\mathbb{E}_{\mathbf{x}_t}\left[J_\theta^\top \mathbb{E}[f(\mathbf{x}_t, \mathbf{x}_0) \mid \mathbf{x}_t]\right] = \mathbb{E}_{\mathbf{x}_t, \mathbf{x}_0}\left[J_\theta^\top f(\mathbf{x}_t, \mathbf{x}_0)\right]. \tag{20}$$

Setting $f(\mathbf{x}_t, \mathbf{x}_0) = R(v_q(\mathbf{x}_t, t) - v(\mathbf{x}_t \mid \mathbf{x}_0))$ and multiplying by $\frac{2t}{1-t}$ yields Eq. (10). In contrast, for a finite-step generator $\mathbf{y} = G_\theta(\mathbf{z}) + \sigma\epsilon$ ($\sigma > 0$) with conditional score $s_\theta(\mathbf{y} \mid \mathbf{z}) = -(\mathbf{y} - G_\theta(\mathbf{z}))/\sigma^2$, the Jacobian $\partial \mathbf{y}/\partial \theta$ retains dependence on $\mathbf{z}$ given $\mathbf{y}$, so $\mathbb{E}[(\frac{\partial \mathbf{y}}{\partial \theta})^\top \mathbb{E}[s_\theta(\mathbf{y} \mid \mathbf{z}) \mid \mathbf{y}]] \neq \mathbb{E}[(\frac{\partial \mathbf{y}}{\partial \theta})^\top s_\theta(\mathbf{y} \mid \mathbf{z})]$ in general. □

Proof of Lemma 2. Using Eqs. (9) and (10), conditioning on $\mathbf{x}_t = \mathbf{x}$ inside g_t^{iTDM} (Eq. (11)) yields the pointwise integrand $2J_\theta^\top[\frac{t}{1-t}\nabla_{\mathbf{x}} m_t(\mathbf{x}) + \beta(v_\theta(\mathbf{x}, t) - v_{\mathrm{ref}}(\mathbf{x}, t))]$. For any q_t with full support, setting the bracketed vector to zero yields $v^\star(\mathbf{x}, t) = v_{\mathrm{ref}}(\mathbf{x}, t) - \frac{t}{\beta(1-t)}\nabla_{\mathbf{x}} m_t(\mathbf{x})$, and applying the score–velocity relation $s_t(\mathbf{x}) = -\frac{\mathbf{x} + (1-t)v(\mathbf{x}, t)}{t}$ (Sec. 2) gives $s_t^\star(\mathbf{x}) = s_t^{\mathrm{ref}}(\mathbf{x}) + \frac{1}{\beta}\nabla_{\mathbf{x}} m_t(\mathbf{x})$ in Eq. (12). Hence $s_t^\star(\mathbf{x}) = \nabla_{\mathbf{x}} \log \pi_t^+(\mathbf{x})$. Moreover, any density p with $\mathrm{KL}(p \| p_{\mathrm{ref},t}) < \infty$ satisfies $\mathcal{J}_t(p) = \beta \log Z_t^\beta - \beta\,\mathrm{KL}(p \| \pi_t^+)$; subtracting this identity at $p = q_t$ from $p = \pi_t^+$ yields Eq. (13), with strictness from the positivity of KL divergence. □

Proof of Lemma 3 and Proposition 2. Conditioning on $(\mathbf{x}_t, \mathbf{c})$ fixes the prompt baseline $b(\mathbf{c})$ and the Jacobian J_θ, while by definition $\mathbb{E}[v_q(\mathbf{x}_t, t) - v(\mathbf{x}_t \mid \mathbf{x}_0) \mid \mathbf{x}_t, \mathbf{c}] = 0$. The baseline contribution to g_t^A therefore vanishes identically, proving $g_t^A = g_t^R$ in Lemma 3 (Eq. (14)). For RWR, conditioning its residual gives $\mathbb{E}[v_\theta - v(\mathbf{x}_t \mid \mathbf{x}_0) \mid \mathbf{x}_t, \mathbf{c}] = v_\theta - v_q$; multiplying by $-2b(\mathbf{c})$ and $J_\theta^\top$ proves the baseline shift $-\Delta_t^b$ in Proposition 2 (Eq. (17)). Finally, adding and subtracting v_q inside $v_\theta - v(\mathbf{x}_t \mid \mathbf{x}_0)$, conditioning $(R - b(\mathbf{c}))(v_\theta - v_q)$ on $(\mathbf{x}_t, \mathbf{c})$ to obtain $(m_t - b(\mathbf{c}))(v_\theta - v_q)$, and separating m_t from $-b(\mathbf{c})$ prove the decomposition in Eq. (18). □

A.2. Approximation and Implementation Scope

Auxiliary Approximation Error and Baseline Dependence. In practice, v_ϕ approximates v_q with residual error $e_\phi = v_\phi - v_q$, so $\widehat{g}_t^A = 2\mathbb{E}[A\, J_\theta^\top (v_\phi - v(\mathbf{x}_t \mid \mathbf{x}_0))]$ satisfies $\widehat{g}_t^A - g_t^A = 2\mathbb{E}[A\, J_\theta^\top e_\phi]$. By the Cauchy–Schwarz inequality, the gradient approximation error is bounded by

$$\|\widehat{g}_t^A - g_t^A\| \le 2\sqrt{\mathbb{E}\left[A^2 \|J_\theta\|_{\mathrm{op}}^2\right]}\sqrt{\mathbb{E}\left[\|e_\phi\|_2^2\right]}, \tag{21}$$

and baseline invariance holds up to e_ϕ (with $\widehat{g}_t^R$ denoting $\widehat{g}_t^A$ at $A = R$):

$$\widehat{g}_t^A - \widehat{g}_t^R = -2b(\mathbf{c})\, \mathbb{E}\left[J_\theta^\top e_\phi\right]. \tag{22}$$

Minimizing $\mathcal{L}_{\mathrm{aux}}(\phi)$ in Eq. (15) directly drives $\mathbb{E}[\|e_\phi\|_2^2]$ toward zero, simultaneously suppressing both the gradient bias and residual baseline shift.

Group Baselines, Normalization, and Clipping. Whereas Lemma 3 considers a fixed prompt baseline $b(\mathbf{c})$, drawing G independent rollouts $\{\mathbf{x}_{0,i}\}_{i=1}^G$ for prompt $\mathbf{c}$ defines the per-sample zero-mean vector

$$h_i = \left(\frac{\partial v_\theta(\mathbf{x}_{t,i}, t)}{\partial \theta}\right)^\top \big(v_q(\mathbf{x}_{t,i}, t) - v(\mathbf{x}_{t,i} \mid \mathbf{x}_{0,i})\big), \tag{23}$$

with $\mathbb{E}[h_i \mid \mathbf{c}] = 0$. By conditional independence across $j \neq i$, centering with the empirical group mean $\overline{R} = \frac{1}{G}\sum_{j=1}^G R_j$ yields

$$\mathbb{E}[(R_i - \overline{R})h_i \mid \mathbf{c}] = \left(1 - \frac{1}{G}\right)\mathbb{E}[R_i h_i \mid \mathbf{c}], \tag{24}$$

scaling the unbiased gradient by the constant factor $(1 - 1/G)$ (absorbable into the learning rate, or removed via the leave-one-out mean $\overline{R}_{-i} = \frac{1}{G-1}\sum_{j\neq i} R_j$). Moreover, group-level standard-deviation normalization and clipping (Algorithm 1) act as bounded reward shaping: for any induced joint law of $(\mathbf{x}_0, A)$ under independent forward noising, the posterior-score identity applies identically toward the shaped target $\pi_t^A \propto q_t e^{\mathbb{E}[A|\mathbf{x}_t,\mathbf{c}]}$.

B. Experimental Details

Task Descriptions and Reward Signals. Following Flow-GRPO (Liu et al., 2025) and DGPO (Luo et al., 2025d), we evaluate post-training on SD3.5-M (Esser et al., 2024) across three tasks:

- *Compositional Image Generation*: GenEval (Ghosh et al., 2023) assesses six compositional skills (single/two-object generation, counting, color attribution, spatial relations, and attribute binding), while GenEval2 (Kamath et al., 2025) (Fig. 1) tests scalability under denser multi-attribute constraints.
- *Visual Text Rendering (OCR)*: Following TextDiffuser (Chen et al., 2023) and Seedream (Gong et al., 2025), prompts use the template "`A sign that says 'text'`", scored by normalized edit distance:
$$r = \max\big(1 - N_e/N_{\mathrm{ref}},\, 0\big), \tag{25}$$
where N_e is the character-level edit distance between the recognized text and target string of length N_{ref}.
- *Human Preference Alignment*: PickScore (Kirstain et al., 2023), trained on large-scale side-by-side user preferences, measures visual appeal and prompt–image alignment.

Implementation and Evaluation Protocol. All experiments are conducted at 512×512 resolution on SD3.5-M (Esser et al., 2024). Both the trainable policy velocity network v_θ and the online auxiliary velocity network v_ϕ are parameterized with lightweight LoRA adapters (Hu et al., 2022), while the pretrained base model serves as the frozen reference velocity v_{ref}. During each iteration (Algorithm 1), we sample G rollouts per prompt using deterministic ODE sampling, compute group-standardized rewards centered by the group mean baseline $b(\mathbf{c}) = \frac{1}{G}\sum_{i=1}^{G} R_i$, and apply advantage clipping $A_i = \text{clip}(\widehat{\text{Adv}}_i, -5, 5)/5$. By default, each optimizer update uses 24 prompts with a group size of $G = 24$; when a baseline exhibits training instability (e.g., AWM and NFT in Fig. 5), we double the number of prompts per update (48 prompts) to stabilize training. Noisy states $\mathbf{x}_t$ are constructed by forward re-noising terminal clean latents via Eq. (1). In the frozen-rollout offline regime (Tab. 3), behavior rollouts are generated solely from the frozen pretrained SD3.5-M (CFG = 4.5), with $G = 2$ used for fair head-to-head comparison against Diffusion-DPO (Wallace et al., 2024). For evaluation, both task metrics and out-of-domain DrawBench (Saharia et al., 2022) metrics are computed using 40-step ODE sampling of the selected checkpoints under the corresponding guidance scale (CFG = 4.5 or CFG = 1.0).

Details of Out-of-Domain Evaluation Metrics. To detect reward hacking and verify general visual fidelity, we report seven unseen metrics on DrawBench: 1) **Aesthetic Score** (Schuhmann et al., 2022), a CLIP-based linear estimator trained on LAION aesthetic ratings; 2) **DeQA** (You et al., 2025), a multimodal LLM assessor sensitive to low-level distortions, blurriness, and synthesis artifacts; 3) **CLIPScore** (Hessel et al., 2021), measuring semantic cosine similarity between prompt and image embeddings; 4) **HPSv2.1** (Wu et al., 2023), a fine-tuned human preference model for stylistic and compositional alignment; 5) **ImageReward** (Xu et al., 2023) (**ImgRwd**), evaluating text–image coherence, structural fidelity, and harmlessness; 6) **PickScore** (Kirstain et al., 2023) on unseen DrawBench prompts (testing out-of-domain preference for GenEval/OCR models and unseen-prompt generalization for PickScore models); and 7) **UnifiedReward** (Wang et al., 2025c) (**UniRwd**), a unified multimodal human-preference reward model.