RL-NATIVE DISTILLATION: EXPLOITING SCORED TRAJECTORIES FOR FEW-STEP IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient text-to-image generation requires both reinforcement-learning (RL)-based reward alignment and few-step distillation, yet these procedures are typically performed sequentially, increasing training cost and risking the loss of reward gains during compression. We take an *RL-native* perspective: diffusion RL already generates reward-scored finite-step trajectories, whose intermediate states provide distillation supervision. Based on this insight, we propose **REST** (Reward-Enhanced Scored-Trajectory Distillation), a single-stage co-training framework in which a decoupled student learns from the evolving RL teacher's trajectories without changing teacher optimization. Advantage-Modulated Distillation (AMD) transforms rollout advantages into signed weights, strengthening imitation of preferred trajectories and aligning distillation priorities with task value. The resulting framework is general and lightweight, requires no extra image rollouts, no separate distillation dataset, and no adversarial training. Experiments on compositional generation, visual text rendering, and human-preference alignment demonstrate competitive few-step, CFG-free generation with RAM or DiffusionNFT teachers. With only four sampling steps, REST-RAM achieves a DrawBench PickScore of 23.97, outperforming both the 40-step RAM teacher (23.95) and RTDMD (23.71).

1 INTRODUCTION

Recent progress in diffusion and flow-matching models (forest labs, 2024; Wu et al., 2025) has substantially improved the visual quality of text-to-image generation. In real practice, however, two post-training procedures are often required before such models become truly useful (Jiang et al., 2025; Fan et al., 2026): few-step sampling and classifier-free guidance (CFG) (Ho & Salimans, 2022) distillation (Luo et al., 2023; Yin et al., 2024b) for efficient inference, and reinforcement-learning-based alignment for human preference (Shao et al., 2024). These two procedures are traditionally executed sequentially, which is cumbersome and often suboptimal: distillation may wash out the reward gains obtained from RL, while subsequent RL can disturb a previously distilled few-step generator and degrade its structure or few-step inference stability (Jiang et al., 2025).

A natural question is whether reward alignment and few-step distillation can be performed in a unified post-training stage. As shown in Fig. 1, recent works such as DMDR (Jiang et al., 2025) and RTDMD (Huang et al., 2026) make progress in this direction by introducing a warm-up distribution-matching distillation (DMD) (Yin et al., 2024b) before RL, followed by RL+DMD parallel optimization. The multi-stage system often involves balancing stage-specific hyperparameters and using more data and training iterations. They also suffer from complicated pipelines compared to RL algorithms, such as extra rollouts, and parallel training. More importantly, the frozen RL-agnostic teacher distribution used by the distillation objective may pull against the continuously evolving reward-optimized policy, creating an inherent tension between imitation and reward improvement.

In this work, we take an *RL-native* perspective: distillation can be built directly into RL training by reusing its reward-scored rollouts and intermediate states, without a separate pre-distillation stage or external distillation data. We observe that RL training already improves direct few-step generation (Fig. 1), but its reward objective does not explicitly train the model to preserve full-step generation quality under a shorter compressed sampling. To learn this compression while preserving the original RL optimization, we attach a decoupled student branch that performs segment-wise imitation

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.

along the teacher's existing reward-scored trajectories. The teacher continues to improve through RL, while the student learns a few-step, CFG-free generator from the evolving teacher.

Few-step distillation is inherently lossy, leading to structural ambiguity and fine-grained artifacts, and we aim to minimize the loss of semantically important content. Inspired by Xue et al. (2025a), we further reuse the rollout reward in student distillation through *Advantage-Modulated Distillation* (AMD). AMD applies an affine transformation to the reward advantage of each teacher rollout, yielding a signed modulation coefficient for the base distillation loss. AMD incorporates reward feedback to encourage the student to preserve task-relevant generation capabilities during compression. This mechanism induces a contrastive effect analogous to CFG (Ho & Salimans, 2022) and NFT (Chen et al., 2026): high-reward trajectories exert stronger attractive supervision toward desirable teacher behaviors, whereas sufficiently low-reward trajectories provide a mild repulsive gradient that discourages the student from reproducing undesirable ones. Compared with uniform distillation, this signed reward modulation improves both visual clarity and reward alignment. Because AMD only reweights the per-sample objective without changing its target, it also serves as a general reward-aware wrapper for different distillation losses.

Methods	Data	Stages	Teacher	Train Steps	Task Feedback
DMDR	2M data	2	Frozen Base	1.5k+1.5k	Partial
RTDMD	2M data	2	Frozen Base	1.5k+1.0k	No
RL then Distill	25k RL data	2	After-RL	1k+1.5k	No
Distill then RL	Both	2	Frozen Base	1.5k+1k	No
Ours	25k RL data	1	RL-evolving	500/1k	Yes



Figure 1: REST achieves few-step, reward-aligned generation with simple, efficient training (top). Few-step CFG-free samples on PickScore, OCR, and GenEval (bottom) show that REST improves on RL-trained models' few-step capabilities, yielding higher quality and preference alignment.

In summary, we present **REST** (Reward-Enhanced Scored Trajectory Distillation), a unified RL-distillation co-training framework for efficient and reward-aligned diffusion generation. To our knowledge, it is the first single-stage framework for RL-distillation collaboration. First, it delivers high-quality few-step generation, matching or even surpassing the full-step inference RL-trained teacher: the general AMD reward modulation compensates for the quality loss introduced by pure distillation. Second, it is general: the decoupled teacher–student design and the generic AMD formulation make the method independent of a specific teacher RL algorithm or a particular distillation loss, as shown in Sec. 4.2 and Appendix B. Third, it is simple and efficient: the framework avoids warm-up scheduling and delicate coordination among multiple competing objectives; it also requires no extra image rollouts and no separate distillation dataset as in DMDR and RTDMD.

- We propose REST, a decoupled teacher–student co-training framework that attaches a lightweight student branch to an arbitrary RL teacher branch. The student reuses the teacher's existing rollout trajectory and reward scores, introducing no extra sampling cost while avoiding interference with teacher RL optimization.
- We introduce AMD, a general reward-aware modulation mechanism for arbitrary trajectory distillation losses. The scale-and-shift transformation of advantages helps align distillation priorities with task value, improving the quality of the few-step student.

- We conduct extensive experiments with RAM and DiffusionNFT, demonstrating competitive 4/8 step CFG-free few-step inference that approaches the quality of the full-inference RL teacher. Compared with two-stage pipelines, it requires much fewer training iterations and obtains substantially better alignment performance.

2 Related Work

Reinforcement learning for diffusion alignment. Reinforcement learning has become a major paradigm for aligning diffusion and flow-matching models with human preferences. Policy-gradient methods such as Flow-GRPO (Liu et al., 2025), DanceGRPO (Xue et al., 2025b), MixGRPO (Li et al., 2025), AWM (Xue et al., 2025a) and DiffusionNFT (Zheng et al., 2025) optimize terminal rewards over denoising trajectories, while preference-based methods (Rafailov et al., 2023; Wallace et al., 2024; Liang et al., 2025) avoid explicit reward modeling through pairwise supervision. Direct-gradient approaches such as ReFL (Xu et al., 2023), DRaFT (Clark et al., 2024), DRTune (Wu et al., 2024), and LeapAlign (Liang et al., 2026) propagate reward gradients through differentiable sampling paths, which require differentiable rewards. Despite their success in improving reward-aligned sampling, these methods largely overlook the potential of integrating distillation into RL, and thus the possibility of achieving reward-aligned few-step generation remains underexplored.

Few-step diffusion distillation. Few-step distillation aims to compress multi-step diffusion samplers into generators requiring only one or a few inference steps. Early trajectory-distillation methods directly regress the student toward the outputs of the teacher's ODE trajectory: Knowledge Distillation (Luhman & Luhman, 2021) matches the teacher's full sampling result in a single step, and Progressive Distillation (Salimans & Ho, 2022) iteratively halves the number of sampling steps by training the student to match two teacher steps at a time. While consistency-based methods, including Consistency Models (Song et al., 2023), Latent Consistency Models (Luo et al., 2023), and Phased Consistency Models (Wang et al., 2024), enforce trajectory-level consistency, score-based methods such as SwiftBrush (Nguyen & Tran, 2024) and Score Implicit Matching (Luo et al., 2024) pursue fast generation through score distillation. Distribution Matching Distillation (DMD) (Yin et al., 2024b) and DMD2 (Yin et al., 2024a) improve few-step quality by matching student and teacher distributions with auxiliary fake-score or discriminator-like networks. However, these distillation objectives are usually reward-agnostic and often require a frozen teacher, a separate distillation stage, or additional auxiliary models. REST is orthogonal to the choice of base distillation loss: it modulates common distillation losses with reward-derived advantages, making distillation preference-aware without redesigning the underlying distillation algorithm.

Unified RL-distillation training. Sequentially applying RL alignment and few-step distillation is cumbersome and can cause objective interference: distillation may wash out RL gains, while later RL can destabilize a previously distilled few-step generator. Recent works such as DMDR (Jiang et al., 2025) and RTDMD (Huang et al., 2026) integrate DMD-style distillation with RL, but their fake-score-centered designs involve multi-network training, warm-up, and staged optimization, and their frozen reference distributions may lag behind the continuously evolving reward-optimized policy. REST instead uses a decoupled dual-branch design: the teacher continues its original RL optimization, while the student simultaneously learns a few-step policy from the same reward-scored teacher rollouts through REST. This yields a unified RL-distillation co-training framework that requires no extra image rollouts, no separate distillation dataset, and no multi-stage model training.

3 Method

3.1 Preliminary: Forward-Process Diffusion RL

Our teacher branch can in principle be optimized by any diffusion RL algorithm, but it pairs especially well with forward-process RL algorithms. The representative instances include DiffusionNFT (Zheng et al., 2025), AWM (Xue et al., 2025a) and RAM (Bergmeister et al., 2026), which all cast reward alignment as a reward-weighted velocity regression on the forward noising process $x_t = (1-t)x_0 + t\epsilon$, where x_0 is the rollout sample and $\epsilon \sim \mathcal{N}(0, I)$. DiffusionNFT converts a scalar reward into an optimality probability $r \in [0, 1]$ and implicitly parameterizes a positive and a

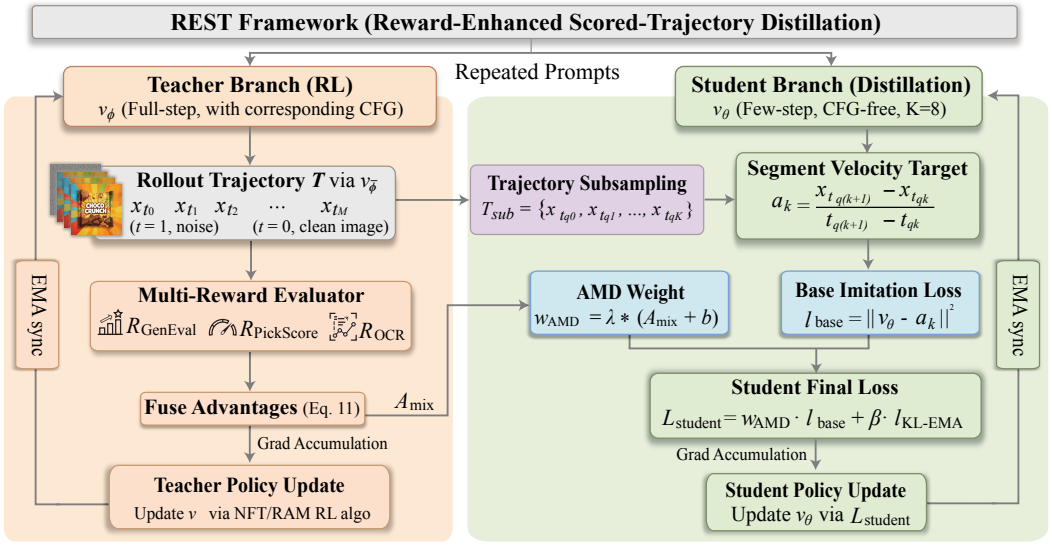

Figure 2: Overview of REST. The decoupled student reuses reward-scored teacher rollouts, while AMD modulates trajectory distillation without altering teacher RL optimization.

negative policy around the old velocity v^{old},

$$v_\phi^+ = (1-\beta)v^{\text{old}} + \beta v_\phi(x_t), \qquad v_\phi^- = (1+\beta)v^{\text{old}} - \beta v_\phi(x_t), \tag{1}$$

and trains the policy v_ϕ with

$$\mathcal{L}_{\text{NFT}}(\phi) = \mathbb{E}\left[r \left\| v_\phi^+(x_t) - v^{gt} \right\|^2 + (1-r) \left\| v_\phi^-(x_t) - v^{gt} \right\|^2 \right], \tag{2}$$

where $v^{gt} = \epsilon - x_0$ is the forward flow-matching (Lipman et al., 2022) target. RAM instead directly regresses v_ϕ toward a reward-shifted version of the reference velocity,

$$\mathcal{L}_{\text{RAM}}(\phi) = \mathbb{E}_t \left[\left\| v_\phi(x_t) - \text{sg}\big(v^{\text{base}}(x_t) + r(x_0)\big(v^{gt} - v^{\text{old}}(x_t)\big)\big) \right\|^2 \right], \tag{3}$$

where $\text{sg}(\cdot)$ is the stop-gradient operator, $v^{\text{base}}(\cdot)$ denotes the frozen base model, and $v^{\text{old}}(\cdot)$ denotes the lagged policy used for sampling. In our framework, the teacher branch is optimized by any forward-process algorithm, e.g., Eq. 2 or Eq. 3, producing an evolving teacher policy v_ϕ together with reward-scored rollouts that the student branch reuses.

3.2 Decoupled Teacher–Student Co-training

Given the teacher policy v_ϕ in Sec. 3.1, REST attaches a decoupled student branch v_θ that learns a few-step, CFG-free generator from the teacher's rollouts, without altering teacher optimization (Fig. 2). For each prompt c, the teacher branch samples with its ODE solver (Song et al., 2020) on an M-step schedule (e.g., $M = 20$) with classifier-free guidance (CFG), producing a trajectory

$$\mathcal{T} = \{x_{t_0}, x_{t_1}, \ldots, x_{t_M}\}, \qquad 1 = t_0 > t_1 > \cdots > t_M = 0, \tag{4}$$

and a terminal reward $r = R(x_0, c)$. The student, in contrast, is designed to run with only $K \ll M$ steps (e.g., $K = 4$ or 8) at inference time. Rather than sampling a separate trajectory for the student, we select a K-step subset of the teacher's schedule and re-index it as an imitation trajectory

$$\mathcal{T}_{\text{sub}} = \{x_{t_{q0}}, x_{t_{q1}}, \ldots, x_{t_{qK}}\} \subset \mathcal{T}, \qquad \{t_{q0}, \ldots, t_{qK}\} \subset \{t_0, \ldots, t_M\}, \tag{5}$$

so that the student can be trained directly on the teacher's existing rollout, with no additional sampling required. For each student step k, the *piecewise trajectory velocity* of the corresponding segment is directly computed from the teacher's rollout,

$$v_k^{\text{gt}} = \frac{x_{t_{q(k+1)}} - x_{t_{qk}}}{t_{q(k+1)} - t_{qk}}. \tag{6}$$

In this decoupled framework, the student θ learns this piecewise trajectory (Sec. 3.3), inheriting the teacher's generation quality while enabling few-step CFG-free sampling.

This dual-branch design brings two benefits. First, it is a one-stage, end-to-end training framework: the student is trained by reusing the teacher's existing rollout trajectory, making training more efficient than a two-stage pipeline. Second, the teacher policy v_ϕ continues to evolve under RL training, avoiding a frozen teacher that could hold back the student branch's RL progress.

REST requires only complete, time-indexed teacher trajectories and their terminal rewards, supporting different RL teachers without changing their objectives. We validate this with RAM and DiffusionNFT (Tab. 1 and Fig. 4); Appendix C analyzes their shared velocity-regression structure.

3.3 Advantage-Weighted Regression for Teacher Tracking

Advantage-weighted regression (Peters & Schaal, 2007; Peng et al., 2019; Kostrikov et al., 2021), *i.e.* AWR, recasts policy improvement as a reward-weighted supervised regression problem: it treats each observed state-action pair as a demonstration and re-weights its log-likelihood under the current policy by the advantage of that action, so that high-advantage demonstrations receive greater weight. We adopt this view to let the student track the teacher's reward-scored rollout: each teacher trajectory segment is treated as a state-action demonstration, whose imitation strength is modulated by the reward of the rollout it comes from. For prompt condition c, the state is $s_k = \big(x_{t_{qk}}, t_{qk}, c\big)$ and the demonstrated action $a_k = v_k^{\mathrm{gt}}$ is given by the teacher's segment velocity (Eq. 6). Following Liu et al. (2025), we interpret the student's velocity prediction as the mean of an implicit, fixed-variance Gaussian policy over teacher segment velocities,

$$\pi_\theta(a_k \mid s_k) = \mathcal{N}\big(a_k;\, v_\theta(s_k), \sigma^2 I\big)\,. \tag{7}$$

The demonstrated action's negative log-likelihood is, up to a θ-independent constant,

$$-\log \pi_\theta(a_k \mid s_k) = \frac{1}{2\sigma^2}\|a_k - v_\theta(s_k)\|^2 + \text{const.} \tag{8}$$

Dropping the constant, this yields the plain *per-step imitation loss*

$$\ell_{\text{base}}(\theta; k) = \|v_\theta(s_k) - a_k\|^2 \;\propto\; -\log \pi_\theta(a_k \mid s_k), \tag{9}$$

which treats every teacher trajectory segment as equally reliable supervision. AWR improves a policy by imitating demonstrated actions with weights given by their advantages,

$$\max_\theta \; \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[A(s,a)\log \pi_\theta(a \mid s)\right]. \tag{10}$$

Using Eq. 8, this objective induces an advantage-weighted regression loss over teacher trajectory segments, where the log-likelihood term is replaced by the base imitation loss ℓ_{base}. Here, $\mathcal{D}$ is the reward-scored teacher trajectory dataset formed by all segments (s_k, a_k). For each reward source i, its advantage $A^{(i)}(s_k, a_k) \in [-1, 1]$ is the clipped group-normalized advantage of the rollout associated with prompt c, and is shared by all segments from the same rollout, as in Flow-GRPO (Liu et al., 2025) and AWM (Xue et al., 2025a). We fuse rewards through their normalized advantages,

$$A_{\text{mix}} = \frac{\sum_{i=1}^{N_R} \alpha_i A^{(i)}}{\sum_{i=1}^{N_R} \alpha_i}, \qquad \alpha_i \geq 0, \tag{11}$$

where N_R is the number of reward sources and α_i controls their contribution. We then introduce a global scale $\lambda = 1$ and a positive shift $b = 0.5$ to obtain a signed modulation coefficient, leading to our Advantage-Modulated Distillation (AMD) objective for teacher trajectory tracking:

$$\mathcal{L}_{\text{AMD}}(\theta) = \mathbb{E}_{(s_k,a_k)\sim\mathcal{D}}\left[\lambda\,(A_{\text{mix}} + b)\;\ell_{\text{base}}(\theta; k)\right]. \tag{12}$$

Since AMD only reweights the base loss, it can also modulate other distillation objectives, as validated with PCM-style phase consistency in Appendix B.

In practice, we further regularize the student against an EMA copy of itself, denoted by $v_{\theta_{\text{ema}}}$, using a fixed-variance Gaussian KL surrogate

$$\mathcal{L}_{\text{KL-EMA}}(\theta) = \mathbb{E}_k\left[\|v_\theta(s_k) - v_{\theta_{ema}}(s_k)\|^2\right], \tag{13}$$

which stabilizes optimization and reduces the training jitter caused by the continuously changing teacher policy. The final student objective is therefore

$$\mathcal{L}_{\text{student}}(\theta) = \mathcal{L}_{\text{AMD}}(\theta) + \beta\mathcal{L}_{\text{KL-EMA}}(\theta), \tag{14}$$

where β controls the strength of the EMA-student regularization.

3.4 Understanding Advantage-Modulated Distillation

Decomposing the AMD objective. Equation 12 can be decomposed into two additive terms,

$$\mathcal{L}_{\mathrm{AMD}}(\theta) = \underbrace{\lambda b\, \mathbb{E}_{(s_k,a_k)\sim\mathcal{D}}[\ell_{\mathrm{base}}(\theta;k)]}_{\text{imitation prior}} + \underbrace{\lambda\, \mathbb{E}_{(s_k,a_k)\sim\mathcal{D}}[A_{\mathrm{mix}}\, \ell_{\mathrm{base}}(\theta;k)]}_{\text{reward-driven correction}}. \tag{15}$$

The first term is a constant-weighted imitation prior. A positive shift b is crucial in our setting: the initial student branch lacks reliable image-generation capability, and needs a stable positive imitation signal to bootstrap its few-step generator. From another perspective, even trajectories with relatively low advantages still contain useful teacher dynamics for the student, especially at early training stages, as shown in Fig. 6 in Appendix E. The second term is the reward-driven correction inherited from reward-weighted supervised regression: it biases the student toward teacher segments with high combined reward advantages, allowing the few-step student to concentrate on the most task-level important contents of the teacher's scored rollouts and potentially surpass the average behavior of the RL-trained full-step teacher. This reward-aware distillation effect is unavailable to conventional distillation methods (Salimans & Ho, 2022) that imitate teacher rollouts uniformly.

Aligning distillation priorities with task value. Few-step distillation introduces approximation errors, but their numerical magnitude does not necessarily reflect their impact on generation quality. For example, a small stroke that turns an "O" into a "Q" can make the requested text incorrect, whereas a comparable change in background texture may have little effect on task quality. AMD incorporates terminal reward feedback to prioritize imitation of high-quality teacher trajectories, allowing task value to influence distillation beyond uniform regression. This helps align distillation priorities with task value, consistent with the gains from advantage modulation observed in Sec. 4.3.

Auxiliary DMD for 4-step generation. We find DMD helps improve image sharpness and explore it as an auxiliary loss to complement AMD in 4-step REST. For each sample, we uniformly select one of four 4 teacher states aligned to the student's target input times $\{1.0, 0.9, 0.75, 0.5\}$. From the selected state x_t, the student predicts a clean latent in one step: $\widehat{x}_0 = x_t - tv_\theta(x_t, t, c)$. After re-noising this prediction, the evolving EMA teacher and a learned fake-score model provide the DMD update direction. We set the DMD-to-AMD loss weight ratio to $2:1$.

4 Experiments

4.1 Experimental Setup

Models and training. We use Stable Diffusion 3.5 Medium (SD3.5M) (Esser et al., 2024) with LoRA (Hu et al., 2022) training. REST-RAM (REST by default) and REST-NFT use RAM and DiffusionNFT teacher branches, respectively, following their official training and evaluation protocols. We evaluate EMA students with 4/8 CFG-free inference steps. Auxiliary DMD loss is used in 4-step distillation. PickScore training uses 500/1000 iterations for 8/4-step REST-RAM and 1500 for REST-NFT; GenEval and OCR training uses 300 iterations, with details in Appendix D.

Tasks and rewards. Following RAM and Flow-GRPO (Liu et al., 2025), we evaluate compositional correctness (GenEval), visual text rendering (OCR), and human preference alignment (PickScore). We use PickScore alone for preference alignment and OCR+PickScore or GenEval+PickScore for the other tasks to preserve general image quality. This design is motivated by **reward hacking and mode collapse observed under prior single-reward optimization**, such as oversized text covering the entire image (Fig. 8 in Appendix F).

Evaluation. Task rewards are evaluated on held-out test prompts. Following DiffusionNFT (Zheng et al., 2025), we evaluate generic image quality on DrawBench (Saharia et al., 2022) prompts disjoint from the reward-training and reward-test prompts, reporting Aesthetic (Schuhmann & Beaumont, 2022), DeQA (You et al., 2025), ImageReward (Xu et al., 2023), HPSv2 (Wu et al., 2023), and PickScore (Kirstain et al., 2023).

4.2 Main Results

Tab. 1 evaluates how well REST retains the gains of its corresponding 40-step RL teacher at 4/8 CFG-free steps. For GenEval and OCR, the gray-shaded rows report single-reward baselines as diagnostic references: these settings can achieve high task rewards while degrading generic image



Figure 3: Qualitative comparison on DrawBench. REST achieves high-quality outputs with superior preference alignment under few-step inference, while matching full-step inference RL teacher.



Figure 4: Generation across RL teachers. (a) 4-step REST-RAM is comparable in visual quality to the 40-step RAM teacher, with better structural coherence and prompt adherence than 4-step RT-DMD. (b) 4/8-step REST-NFT has visual quality comparable to the 40-step DiffusionNFT teacher.

quality. We therefore focus on the rows marked by †, which use the task reward together with PickScore and compare 8-step REST-RAM with its RAM teacher. The human-preference block additionally reports 4-step REST-RAM and both REST-NFT variants.

As shown in Tab. 1, REST achieves performance comparable to the 40-step RAM/DiffusionNFT teacher on both training rewards and out-of-domain DrawBench metrics, while using 4/8 CFG-

Table 1: SD3.5M post-training on GenEval, OCR, and PickScore. REST-RAM/NFT denote RAM/DiffusionNFT teachers. Gray rows use a single task reward; † adds PickScore.

Model	Steps	Training Reward			Image Quality Metrics				
		GenEval	OCR	PickScore	Aesthetic	DeQA	ImgRwd	HPSv2	PickScore
SD3.5M	40	0.64	0.63	21.78	5.39	4.08	0.85	0.28	22.40
SD3.5M	8	0.28	0.12	19.50	5.15	2.38	-0.96	0.18	20.39
Compositional Image Generation									
Flow-GRPO	40	0.95			5.25	4.01	1.03	0.27	22.37
AWM	40	0.83			5.14	3.75	0.67	0.24	22.04
DiffusionNFT	40	0.95			4.98	4.10	0.30	0.24	21.59
RAM	40	0.97			5.38	4.09	1.19	0.29	22.52
RAM†	40	**0.95**			5.42	4.12	**1.12**	0.28	22.76
RAM†	8	0.65			5.46	2.27	-0.19	0.20	21.36
REST-RAM†	8	0.94			**5.67**	**4.15**	1.05	**0.30**	**22.79**
Visual Text Rendering									
Flow-GRPO	40		0.92		5.32	4.06	0.95	0.28	22.44
AWM	40		0.97		5.01	2.83	-0.85	0.18	20.56
DiffusionNFT	40		0.96		4.87	3.01	-0.97	0.18	20.26
RAM	40		0.97		5.23	3.90	0.44	0.26	21.83
RAM†	40		**0.96**		5.58	**4.16**	**1.07**	**0.30**	22.81
RAM†	8		0.68		5.59	2.60	-0.21	0.22	21.45
REST-RAM†	8		**0.96**		**5.65**	**4.16**	1.00	**0.30**	**22.83**
Human Preference Alignment									
Flow-GRPO	40			23.31	5.92	**4.22**	1.28	0.32	23.53
AWM	40			23.39	6.31	4.10	1.27	0.31	23.76
DiffusionNFT	40			23.45	6.18	4.09	1.29	0.30	23.73
REST-NFT	8			23.66	6.26	4.15	1.33	0.30	23.87
REST-NFT	4			23.48	6.23	4.11	1.28	0.30	23.69
RAM	40			23.70	6.17	4.21	1.35	0.32	23.95
RAM	8			23.30	**6.42**	3.82	1.11	0.31	23.55
REST-RAM	8			**23.73**	6.30	4.15	1.35	**0.33**	23.96
REST-RAM	4			23.72	6.20	4.10	**1.38**	0.31	**23.97**

Table 2: 4-step DrawBench comparison with RTDMD (1,500 warm-up + 1,000 parallel iterations). †: official weights; ‡: reproduced with official code, PickScore reward, and Pick-a-Pic data.

Methods	Steps	Train Iters	Aesthetic	ImgRwd	HPSv2	PickScore	CLIPScore
RTDMD†	4	1500+1000	6.09	1.26	**0.326**	23.30	0.930
RTDMD‡	4	1500+1000	6.18	1.27	0.310	23.71	0.885
REST-RAM	4	1000	**6.20**	**1.38**	0.311	**23.97**	**0.932**

free inference steps. It also obviously surpasses full-step RL baselines such as Flow-GRPO (Liu et al., 2025), AWM (Xue et al., 2025a) and DiffusionNFT (Zheng et al., 2025). We note that 8-step RAM can occasionally obtain a high Aesthetic score, but this is often associated with fragmented structures and noisy details that hack the metric rather than reflect better perceptual quality. Overall, by continuously exposing the student to the evolving reward-optimized teacher rollouts and using AMD to select and amplify high-value trajectory segments, REST enables the student to go beyond plain trajectory compression. Qualitative comparisons in Figs. 3, 7, and 8 (see Appendix F for the latter two) further show that REST preserves generic image quality, human preference, and visual text rendering under few-step CFG-free sampling.

Compared with Few-step RL Methods. DMDR and RTDMD (Jiang et al., 2025; Huang et al., 2026) combine a DMD warm-up stage with subsequent RL–DMD optimization. However, they use larger training datasets (Schuhmann et al., 2021) and longer training of roughly 2,500–3,000 steps. More importantly, neither work directly compares its distilled generator against the corresponding full-step RL model trained under the same setting, leaving unclear how much of the RL policy's performance is actually preserved after distillation. REST instead targets the quality of full-step RL models under the popular GenEval, OCR, PickScore, and DrawBench protocols in diffusion RL

Table 3: Progressive REST ablation: PickScore training, 8-step CFG-free evaluation on DrawBench.

Variant	Teacher RL	Dual Branch	AMD	Aesthetic	DeQA	ImgRwd	HPSv2	PickScore
Base model	–	–	–	5.15	2.38	-0.96	0.18	20.39
RL	Yes	–	–	**6.42**	3.82	1.11	0.31	23.55
Distill then RL	–	–	–	5.83	3.54	1.19	0.28	23.14
RL then Distill	Yes	–	–	6.22	4.14	1.32	0.32	23.83
RL & Distill	Yes	Yes	–	6.20	4.09	1.32	0.32	23.74
REST (full)	Yes	Yes	Yes	6.30	**4.15**	**1.35**	**0.33**	**23.96**

Table 4: AMD/auxiliary DMD ablation: 4-step REST-RAM, PickScore training, DrawBench test.

Variant	Aesthetic	DeQA	ImgRwd	HPSv2	PickScore
DMD only	6.10	3.99	1.38	0.30	23.58
AMD only	6.16	4.04	1.33	0.31	23.93
On Student traj.	6.26	4.08	1.35	0.33	23.95
REST (full)	6.20	4.10	1.38	0.31	23.97

fields, while requiring far fewer training steps. For an additional comparison, we evaluate the official RTDMD weights and reproduced RTDMD using official code on the same rewards and prompts as REST, at the training-independent DrawBench benchmark. As shown in Tab. 2, REST achieves stronger aesthetic and preference scores over both RTDMD versions despite using substantially fewer training steps.

Training efficiency. Since REST reuses the teacher's existing rollout trajectories and reward evaluations, it requires no additional rollout sampling, which is usually the most expensive stage in RL. Four-step REST-NFT and REST-RAM incur only **32.8%** and **26.1%** training-time overhead relative to standard DiffusionNFT and RAM training without distillation, respectively. Compared with sequential RL-then-distill training, REST achieves better performance with substantially fewer training iterations, as shown in Fig. 1.

4.3 Ablation Study

Decoupled co-training and AMD. Tab. 3 studies each component in 8-step REST-RAM PickScore training. We start from direct 8-step CFG-free inference with the base model, then add standard RL (RAM), a sequential RL-then-distill or distill-then-RL two-stage pipeline, simultaneous dual-branch RL with distillation, and finally the full REST model with AMD. The results lead to three observations. First, full-step standard RL itself improves 8-step CFG-free sampling, showing that reward optimization can partially compensate for the low-step generation gap. Second, simultaneous dual-branch RL-distillation in REST retains most of the gains of the RL-then-distill pipeline with far fewer training iterations, since the student reuses the teacher rollouts instead of requiring a separate distillation stage. Third, AMD further strengthens reward preference and improves perceptual quality, indicating that reward-aware trajectory weighting is more effective than uniform imitation.

AMD and auxiliary DMD for four-step generation. Tab. 4 isolates the contributions of AMD and auxiliary DMD in 4-step REST-RAM under PickScore training. AMD strengthens reward alignment, yielding higher PickScore, while DMD improves visual sharpness. We also compare DMD using student-rollout versus teacher-rollout states. Both yield similar results and are viable choices, while reusing teacher states avoids additional student rollouts and reduces sampling overhead.

5 Conclusion

We present **REST**, a unified RL-distillation co-training framework that turns reward-scored teacher trajectories into supervision for a few-step, CFG-free student. Its decoupled design preserves teacher RL optimization, while **Advantage-Modulated Distillation** (AMD) strengthens imitation of preferred trajectories and suppresses or repels low-reward ones. An auxiliary DMD loss using stored teacher states supports four-step generation. Experiments on SD3.5M demonstrate competitive reward alignment and image quality at four and eight inference steps. The ablation and analysis studies further validate the importance of decoupled co-training, advantage modulation, KL-EMA stabilization, and the generality of REST across teacher RL algorithms and distillation losses.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, literature and reference checking, and code implementation for ablation studies and the Auxiliary DMD loss. The AI-assisted code was reviewed and tested for correctness by one author. AI tools also help provide theoretical support for extending REST to a broader class of ODE-based diffusion RL teachers by comparing RAM and DiffusionNFT in the Appendix. We reviewed and revised the AI-assisted manuscript content and take responsibility for the final text, claims, code, data, and other artifacts in this work.

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. *arXiv preprint arXiv:2605.10759*, 2026.

Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. Nft: Bridging supervised learning and reinforcement learning in math reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 124025–124042, 2026.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Linqian Fan, Peiqin Sun, Tiancheng Wen, Shun Lu, and Chengru Song. Rdm: Re-conceptualizing distribution matching as a reward for diffusion distillation. *arXiv preprint arXiv:2603.28460*, 2026.

Black forest labs. Flux.1-dev. `https://github.com/black-forest-labs/flux`, 2024.

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

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

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Zhen Li, Bo Zhang, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

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*, volume 36, 2023.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13208, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23238–23248, 2026.

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

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.

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

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

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024.

Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7807–7816, 2024.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, pp. 745–750, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed 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 *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

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

Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. laion.ai, 2022.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32211–32252, 2023.

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*, pp. 8228–8238, 2024.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

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.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 108–124. Springer, 2024.

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. *NeurIPS*, 2023.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025a.

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.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 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. *arXiv preprint arXiv:2509.16117*, 2025.