STABLE SIGNED FLOW REINFORCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward-based finetuning of flow models requires promoting high-reward samples and suppressing low-reward ones while maintaining training stability. Reward-weighted regression provides a simple starting point, but nonnegative weights give no explicit unfitting signal, and repeated reweighting can concentrate the distribution and cause collapse. We propose **Stable Signed Flow Reinforcement (SIFR)**, which uses signed advantages to explicitly reinforce high-reward samples and unfit low-reward ones, while constraining velocity drift from a reference model to stabilize the updates. Alternating updates of the model and a Lagrange multiplier adapt the regularization strength to a prescribed drift budget. Experiments show that SIFR achieves higher rewards than existing methods in fewer training rounds on SD3.5 Medium (2.5B), FLUX.2 Klein (9B), and Cosmos 3 Super Text2Image (64B), using preference rewards as well as rule-based rewards. These gains further transfer to held-out rewards and prompt sets. Ablations show that signed advantages and the drift constraint each contribute to stable training, and that together they make flow RL faster and raise its reward ceiling.

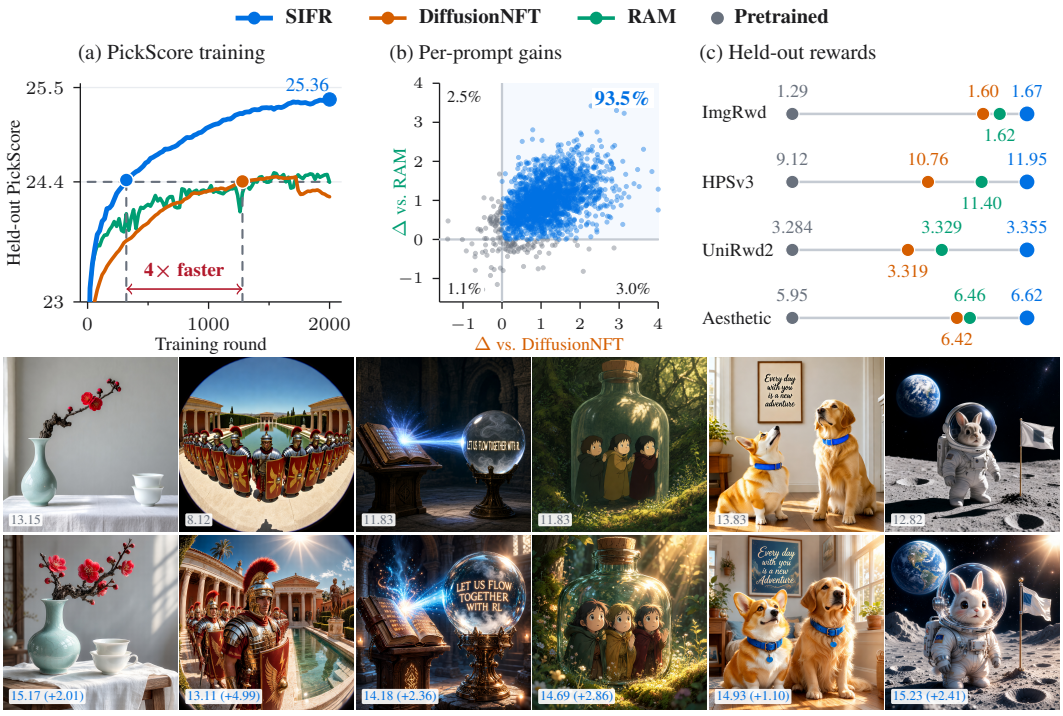


Figure 1: RL training results for Cosmos 3 Super Text2Image with SIFR. (a) PickScore evaluated on held-out prompts throughout training. (b) Per-prompt PickScore gains over DiffusionNFT and RAM. (c) Scores on held-out reward models after 500 rounds of multi-reward training. Images compare the pretrained model (top) with SIFR (bottom); labels show HPSv3 scores and gains.

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.

1 OVERVIEW

Reward-based post-training has become a standard component of language-model training and is increasingly being adopted for image and video generation with flow- and diffusion-based models. Despite substantial recent progress, post-training methods for these generative models have yet to converge on a standard paradigm like the case of LLMs. A central challenge is how to effectively promote high-reward samples and suppress low-reward ones while maintaining training stability.

A basic building block of reinforcement learning (RL) is iterative reward fine-tuning. For a nonnegative reward function r, in the case of rectified flow (RF), reward-weighted tuning works by

$$\theta^{(k+1)} = \arg\min_{\theta} \mathbb{E}_{\theta^{(k)}}\big[r(X_1)\|v_\theta(X_t,t) - (X_1 - X_0)\|^2\big].$$

Here v_θ is the model's velocity field, X_1 is a rollout from the current model $\theta^{(k)}$, $X_0 \sim \pi_0$ is base noise, $X_t = (1-t)X_0 + tX_1$, and $\mathbb{E}_{\theta^{(k)}}$ averages over these draws. However, this approach fails to work in practice: 1) repeated reward weighting progressively concentrates the model distribution and leads to collapse; and 2) nonnegative weights only encourage fitting high-reward samples and provide no explicit signal to unfit low-reward ones.

A natural remedy is to replace the reward weighting term $r(X_1)$ by advantage $A(X_1)$, so that below-average samples receive negative weights and are explicitly suppressed. However, an MSE loss with negative weights can be unbounded below and lead to training instability. We find that existing flow RL methods implicitly use this signed loss, although their stated losses are nonnegative, and that they remain stable because they also implicitly regularize it toward a reference model (section 2). The strength of this regularization is delicate: too little leaves training unstable, whereas too much anchors the model to its reference, which slows training and lowers the performance ceiling.

In this work, we propose **Stable Signed Flow Reinforcement** (**SIFR**), an online RL method for RF models that formalizes flow RL as a constrained optimization problem. At each round k, it solves

$$\min_{\theta} \mathbb{E}_{\theta^{(k)}}\big[A(X_1)\|v_\theta(X_t,t) - (X_1 - X_0)\|^2\big] \quad \text{s.t.} \quad \mathbb{E}_{\theta^{(k)}}\big[\|v_\theta(X_t,t) - v_{\text{ref}}^{(k)}(X_t,t)\|^2\big] \le \delta.$$

Here $v_{\text{ref}}^{(k)}$ is the reference velocity field in round k, and $\delta \ge 0$ is the drift budget. In practice, we solve this problem by alternating updates of the model parameters and the Lagrange multiplier, which keeps the regularization only as strong as the budget requires, so that the performance ceiling of RL is as high as possible while the optimization remains stable (section 3).

Empirical Results We evaluate SIFR on SD3.5 Medium (2.5B), FLUX.2 Klein (9B), and Cosmos 3 Super Text2Image (64B) with four training objectives. Across all 12 backbone–objective pairs, SIFR outperforms or matches DiffusionNFT and RAM. These gains also transfer to held-out prompts and rewards, with better performance in most comparisons and competitive results otherwise.

Outline Section 2 analyzes the instabilities of online reward tilting and of signed regression, and shows that existing flow RL methods are gradient-equivalent to regularized signed regression. Section 3 formulates SIFR as a constrained optimization problem, characterizes its saddle point, and gives the practical algorithm. Section 4 evaluates the optimization of RL objectives and the transfer to held-out rewards and prompts, and ablates the design choices.

2 FROM POSITIVE WEIGHTS TO SIGNED REGRESSION

Rectified Flow Let π_0 be a source distribution on $\mathbb{R}^d$, such as $\mathcal{N}(0, I)$, and let π_1 be the target data distribution of the same dimension. Given noise-data pairs (X_0, X_1) with $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$, rectified flow (RF) learns a velocity field $v_\theta(x,t)$ by minimizing the mean squared error

$$\min_{\theta} \mathbb{E}\big[\|v_\theta(X_t,t) - (X_1 - X_0)\|^2\big], \qquad X_t = (1-t)X_0 + tX_1, \tag{1}$$

where X_t interpolates between the endpoints, and $\mathbb{E}[\cdot]$ averages over (X_0, X_1) and an independent time $t \in [0,1]$. The minimizer over unrestricted velocity fields is the conditional expectation

$$v^\star(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x].$$

One of the key properties of RF is marginal preservation: under standard regularity conditions, the process Z_t generated by the ODE $\dot{Z}_t = v^\star(Z_t,t)$, initialized with $Z_0 \sim \pi_0$, has the same distribution as X_t at every $t \in [0,1]$. In particular, $Z_1 \sim \pi_1$. See Liu et al. (2023); Liu (2022) for further details.

Reward Tilting of Rectified Flow Assume we are given a reward function $r(x) \geq 0$ without loss of generality. The reward-tilted distribution is $\pi_1^r(x) = r(x)\pi_1(x)/Z_r$, where $Z_r = \mathbb{E}[r(X_1)]$. It is easy to see that the RF velocity that generates π_1^r can be learned by minimizing the weighted flow regression loss:

$$\min_\theta \ \mathbb{E}\big[r(X_1)\|v_\theta(X_t,t) - (X_1 - X_0)\|^2\big] \implies v^r(x,t) = \frac{\mathbb{E}[r(X_1)(X_1 - X_0) \mid X_t = x]}{\mathbb{E}[r(X_1) \mid X_t = x]}. \quad (2)$$

Intuitively, reweighting the regression loss is equivalent to reweighting the data samples, so that high-reward samples are generated more frequently.

Online Reward-Weighted Finetuning The reward tilting above reweights samples from a fixed data distribution π_1. A more aggressive approach is to alternate generation and finetuning, using samples from the updated model at each round, as in online rejection sampling finetuning for language models (Touvron et al., 2023; Shao et al., 2024). For RF, the corresponding update with rewards $r(x) \geq 0$ can be expressed as

$$\theta^{(k+1)} = \arg\min_\theta \ \mathbb{E}_{\theta^{(k)}}\big[r(X_1)\|v_\theta(X_t,t) - (X_1 - X_0)\|^2\big]. \quad (3)$$

At round k, the current velocity field $v_{\theta^{(k)}}$ generates rollouts X_1 and is updated with optimizer steps on their reward-weighted loss. The updated velocity field generates the next round's rollouts, so training proceeds online with data continually refreshed by the current velocity field.

Unfortunately, iterative tilting as in eq. (3) fails to work in practice, for two reasons:

- *Repeated tilting progressively concentrates probability mass near reward maximizers and leads to distributional collapse (appendix C.1).*
- *Nonnegative weights encourage the model to fit high-reward samples but provide no explicit signal to unfit low-reward samples, limiting the effectiveness of the updates.*

Signed Tilting and Signed RF A natural next step is to replace $r(X_1)$ with a signed advantage $A(X_1)$, giving

$$\theta^{(k+1)} = \arg\min_\theta \ \mathbb{E}_{\theta^{(k)}}\big[A(X_1)\|v_\theta(X_t,t) - (X_1 - X_0)\|^2\big], \qquad A(x) = \frac{r(x) - \mu_r}{\sigma_r + \epsilon}, \quad (4)$$

where μ_r and σ_r are the reward mean and standard deviation (Shao et al., 2024). Optimizing this signed MSE *minimizes* the loss for above-average samples $r(X_1) > \mu_r$ and *maximizes* it for below-average samples, which yields an unfitting signal.

However, this unfitting signal introduces a new source of instability: an MSE loss with negative weights can be unbounded below. To understand its stability, it is useful to expand the quadratic terms of the loss as

$$\mathbb{E}_{\theta^{(k)}}\big[m(X_t,t)\|v_\theta(X_t,t)\|^2 - 2n(X_t,t)^\top v_\theta(X_t,t)\big],$$

where the quadratic and linear coefficients (whose dependence on the round k we suppress) are

$$m(x,t) = \mathbb{E}_{\theta^{(k)}}[A(X_1) \mid X_t = x], \qquad n(x,t) = \mathbb{E}_{\theta^{(k)}}[A(X_1)(X_1 - X_0) \mid X_t = x].$$

Hence, the optimization in (4) is stable when $m(x,t) > 0$. In practice, however, it is hard to know if $m(x,t) > 0$ holds unless we estimate $m(x,t)$. When $m(x,t) > 0$, the solution is

$$v^A(x,t) = \frac{n(x,t)}{m(x,t)} = \frac{\mathbb{E}_{\theta^{(k)}}[A(X_1)(X_1 - X_0) \mid X_t = x]}{\mathbb{E}_{\theta^{(k)}}[A(X_1) \mid X_t = x]}. \quad (5)$$

When $m(x,t) < 0$, the ratio $v^A(x,t)$ is a conditional maximum, and the conditional loss is unbounded below. If both sign regions of $m(\cdot,t)$ have positive training measure, then v^A can only be a saddle of the signed loss (Proposition 1). Moreover, the velocity field v^A becomes singular near the zero boundary $\{x : m(x,t) = 0\}$.

Despite these singularities, the velocity field v^A can still define valid sampling dynamics through Signed Rectified Flow (Signed RF) (Liao et al., 2026). Assume $Z_A = \mathbb{E}[A(X_1)] > 0$. With an independent Gaussian source and under the regularity assumptions of Signed RF, one can show that ODE trajectories driven by v^A and initialized from π_0 remain in the positive region $\{x : m(x,t) >$

$0\}$ for $t \in [0, 1)$, and the generated density is a truncated density of $\pi_1(x)A(x)$ within the positive region:

$$\rho_1(x) = A(x)\pi_1(x)\mathbf{1}\{x \in \Omega_{\text{reach}}\}/Z_A, \qquad \Omega_{\text{reach}} \subseteq \{x : A(x) > 0\},$$

where the reachable region Ω_{reach} is determined by the flow dynamics and satisfies the normalization condition $\mathbb{E}[A(X_1)\mathbf{1}\{X_1 \in \Omega_{\text{reach}}\}] = Z_A$. Thus, Signed RF preserves the advantage-weighted density on a reachable subset of the positive region while excluding the negative region.

However, learning this field v^A remains difficult: minimizing the signed loss does not recover its saddle, and singularities near the zero boundary of v^A make direct neural approximation and numerical integration sensitive to errors. These difficulties motivate regularized signed regression.

Regularized Signed Regression Regularization toward reference models offers a natural way to control both distributional collapse from repeated tilting and the optimization instability of signed MSE. Interestingly, although existing flow RL methods such as DiffusionNFT (Zheng et al., 2026) and RAM (Bergmeister et al., 2026) are formulated as minimizing nonnegative MSE losses, we show that their stop-gradient operations and parameterizations yield update gradients equivalent to those of signed MSE with reference regularization, even though this structure is not explicit in their original objectives. We derive and summarize these gradient-equivalent objectives in the following unified form:

$$\mathcal{L}_{\text{reg}}^{(k)}(\theta) = \mathbb{E}_{\theta^{(k)}}\Big[w(X_1)\|v_\theta(X_t, t) - (X_1 - X_0)\|^2 + c_{\text{ref}}^{(k)} D(v_\theta, v_{\text{ref}}^{(k)}) + c_{\text{ref}}^{\text{fix}} D(v_\theta, v_{\text{ref}}^{\text{fix}})\Big], \tag{6}$$

where $D(v, v_{\text{ref}}) = \|v(X_t, t) - v_{\text{ref}}(X_t, t)\|^2$. Here v_{ref} denotes the reference model, for which we consider two choices. The fixed reference $v_{\text{ref}}^{\text{fix}}$, such as a pretrained teacher model, remains unchanged throughout training. The dynamic reference $v_{\text{ref}}^{(k)}$ is parameterized by either $\theta_{\text{ema}}^{(k)}$, an exponential moving average (EMA) of model parameters across RL rounds, or the current snapshot $\theta^{(k)}$. Both reference predictions are detached during model updates. Table 1 lists, for each method, the weight $w(X_1)$ and the coefficients $c_{\text{ref}}^{(k)}$ and $c_{\text{ref}}^{\text{fix}}$ that reproduce its update gradient.

Method	$v_{\text{ref}}^{(k)}$	$w(X_1)$	$c_{\text{ref}}^{(k)}$	$c_{\text{ref}}^{\text{fix}}$
ORW-CFM-W2 (Fan et al., 2025)	—	$e^{\tau r(X_1)}$	0	α
AWM (Xue et al., 2026)	EMA	$(1-t)A(X_1)$	$\beta(1-t)$	0
DiffusionNFT (Zheng et al., 2026)	No EMA	$\beta A(X_1)$	0	0
	EMA		$\beta(\beta - A(X_1))$	
RAM (Bergmeister et al., 2026)	No EMA	$\eta A(X_1)$	0	1
	EMA		$-\eta A(X_1)$	
AdvantageFlow (Kveton et al., 2026)	EMA	$(1-t)^2 A(X_1)$	$(1-t)^2\gamma(X_1)$	$(1-t)^2\lambda$
FlowAWR (Fu et al., 2026)	EMA	$A(X_1)$	$1 - A(X_1)$	0

Table 1: Existing flow RL methods admit gradient-equivalent objectives of the form in eq. (6); derivations are given in appendix C.5. The $v_{\text{ref}}^{(k)}$ column gives the dynamic reference, either an EMA of model parameters or the detached current snapshot $\theta^{(k)}$ (No EMA), and a dash means none; DiffusionNFT and RAM have both variants. ORW-CFM-W2 uses positive exponential weights, while the other methods use weights proportional to signed advantages. The reported hyperparameters are $\gamma(X_1) \in \{1.1, 1 - A(X_1)\}$ for AdvantageFlow, $\eta \in \{100, 1000\}$ for RAM, $\beta \in \{0.1, 1\}$ for DiffusionNFT, and a task-dependent τ (e.g., 1 or 10) for ORW-CFM-W2.

3 Stable Signed Flow Reinforcement

Section 2 shows that existing flow RL methods implicitly use both a signed MSE loss and a regularization toward a reference model, and that the regularization can mitigate these instabilities. However, in these methods the strength is a coefficient fixed before training (c_{ref} in Table 1), and the choice is delicate: too little leaves the loss unbounded and training unstable, whereas too much anchors the model to its reference, which slows training and lowers the performance ceiling. In this section, we formalize signed flow RL as a constrained optimization problem, so that the reward improves quickly and to a high ceiling while training remains stable.

Stable Signed Flow Reinforcement Formally, given $\theta^{(k)}$, we obtain $\theta^{(k+1)}$ by solving

$$\min_{\theta} \mathbb{E}_{\theta^{(k)}}\big[A(X_1)\|v_\theta(X_t,t)-(X_1-X_0)\|^2\big] \quad \text{s.t. } \mathbb{E}_{\theta^{(k)}}\big[D(v_\theta, v_{\text{ref}}^{(k)})\big] \le \delta, \tag{7}$$

where $D(v_\theta, v_{\text{ref}}^{(k)}) = \|v_\theta(X_t,t) - v_{\text{ref}}^{(k)}(X_t,t)\|^2$ is the squared velocity drift as in eq. (6), $v_{\text{ref}}^{(k)}$ is the reference model in round k, and δ is the drift budget. Multiple references, such as $v_{\text{ref}}^{\text{fix}}$ and $v_{\text{ref}}^{(k)}$, can be included with a separate constraint each; we use one here for simplicity.

The Lagrange dual of eq. (7) yields a minimax problem $\max_{\lambda\ge 0}\min_\theta \mathcal{L}^{(k)}(\theta,\lambda)$ with

$$\mathcal{L}^{(k)}(\theta,\lambda) = \mathbb{E}_{\theta^{(k)}}\big[A(X_1)\|v_\theta(X_t,t)-(X_1-X_0)\|^2 + \lambda\big(D(v_\theta, v_{\text{ref}}^{(k)}) - \delta\big)\big] \tag{8}$$

One can search for a saddle point of the Lagrangian by alternating updates of θ and λ, such as

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \epsilon\nabla_\theta \mathcal{L}^{(k)}(\theta^{(k)},\lambda^{(k)}), \quad \lambda^{(k+1)} \leftarrow \max(\lambda^{(k)} + \epsilon\nabla_\lambda \mathcal{L}^{(k)}(\theta^{(k+1)},\lambda^{(k)}),\ 0).$$

where the θ update minimizes a regularized loss with coefficient λ, while the λ update seeks to enforce the constraint by increasing λ when the drift exceeds δ and decreasing it when it falls below.

Saddle Point For the fixed round-k rollouts, a finite saddle point $(v^\star,\lambda^\star)$ of $\mathcal{L}^{(k)}$ satisfies

$$v^\star(x,t) = \frac{n(x,t) + \lambda^\star v_{\text{ref}}^{(k)}(x,t)}{m(x,t)+\lambda^\star}, \qquad m(x,t)+\lambda^\star > 0, \quad \text{almost everywhere,}$$

together with the standard feasibility and complementary slackness conditions. Here $v^\star$ coincides with the signed RF velocity with weight $A+\lambda^\star$ when $v_{\text{ref}}^{(k)} = v_{\theta^{(k)}}$, with $\lambda^\star$ adaptively adjusted to satisfy both the stability condition $m(x,t)+\lambda^\star \ge 0$ and the drift constraint. The stability condition is necessary because the multiplier must eliminate negative conditional curvature almost everywhere at a finite saddle point. Intuitively, otherwise, minimizing the loss with respect to v would drive its magnitude to infinity and violate the drift constraint, pushing λ upward to restore $m(x,t)+\lambda \ge 0$.

3.1 Practical Algorithm

Reference choice By default, we use a step constraint with $v_{\text{ref}}^{(k)} = v_{\theta_{\text{ema}}^{(k)}}$, where $\theta_{\text{ema}}^{(k)}$ is an exponential moving average (EMA) of the parameters over past rounds with decay $\beta = 0.5$; the current snapshot $\theta^{(k)}$ is the special case $\beta = 0$. Optionally, a frozen pretrained reference $v_{\text{ref}}^{\text{fix}}$ controls departure from the pretrained model. The two constraints can also be combined with separate budgets and multipliers; the Lagrangian then adds both drift penalties, each weighted by its own multiplier.

Multiplier update At the start of RL training, we initialize the budget to κ times the RF loss of the pretrained model on its own rollouts, $\delta = \kappa\,\mathbb{E}_{\theta^{(0)}}\|v_{\theta^{(0)}}(X_t,t)-(X_1-X_0)\|^2$. The budget remains fixed throughout training. After each round of model updates, we measure the drift $D^{(k)} = \mathbb{E}_{\theta_{\text{ema}}^{(k)}}\big[D(v_{\theta^{(k+1)}}, v_{\text{ref}}^{(k)})\big]$ and update its moving average $D_{\text{ema}}^{(k)}$. The multiplier is then updated by

$$\log\lambda^{(k+1)} = \text{clip}\Big(\log\lambda^{(k)} + \tau\,\text{clip}\Big(\frac{D_{\text{ema}}^{(k)}}{\delta} - 1,\ -1,\ 1\Big),\ \log\lambda_{\min},\ \log\lambda_{\max}\Big). \tag{9}$$

Here τ sets the step size, and $\lambda_{\min}$ and $\lambda_{\max}$ bound λ. Algorithm 1 summarizes the procedure.

Algorithm 1 SIFR with an EMA step constraint

Require: Reward model r; trainable velocity field $v_{\theta^{(0)}}$.
Initialize $\theta_{\text{ema}}^{(0)} \leftarrow \theta^{(0)}$, budget δ, and $\lambda^{(0)} > 0$.
for $k = 0, 1, 2, \ldots$ **do**
  Roll out $X_1 \leftarrow \text{ODESolve}(v_{\text{ref}}^{(k)})$; evaluate $r(X_1)$ and compute $A(X_1)$.
  Update $v_{\theta^{(k+1)}} \leftarrow v_{\theta^{(k)}}$ with S optimizer steps using
    $\nabla_\theta\, \mathbb{E}_{\theta_{\text{ema}}^{(k)}}\Big[A(X_1)\|v_\theta(X_t,t)-(X_1-X_0)\|^2 + \lambda^{(k)} D(v_\theta, v_{\text{ref}}^{(k)})\Big]$.
  Measure drift $D^{(k)} \leftarrow \mathbb{E}_{\theta_{\text{ema}}^{(k)}}\big[D(v_{\theta^{(k+1)}}, v_{\text{ref}}^{(k)})\big]$; update $D_{\text{ema}}^{(k)}$.
  Update multiplier $\lambda^{(k+1)} \leftarrow \lambda^{(k)}$ using eq. (9).
  Update EMA $\theta_{\text{ema}}^{(k+1)} \leftarrow \beta\,\theta_{\text{ema}}^{(k)} + (1-\beta)\theta^{(k+1)}$.
end for

4 EXPERIMENTS

We evaluate SIFR on SD3.5 Medium (2.5B) (Esser et al., 2024; Stability AI, 2024), FLUX.2 Klein (9B) (Black Forest Labs, 2026), and Cosmos 3 Super Text2Image (64B) (NVIDIA, 2026). We first compare post-training performance with existing flow RL methods, including DiffusionNFT (Zheng et al., 2026) and RAM (Bergmeister et al., 2026), then broaden the evaluation to longer training horizons and different prompt sets. Finally, we conduct ablation studies of key design choices.

4.1 EXPERIMENTAL SETUP

Rewards and prompt sets We optimize four objectives in separate runs. For preference alignment, we consider two settings: *single-reward training* with PickScore (Kirstain et al., 2023), and *multi-reward training* with the sum of raw PickScore, HPSv2.1 (Wu et al., 2023), and ClipScore (Radford et al., 2021). Both settings use the Pick-a-Pic training prompts (Kirstain et al., 2023). For the OCR task and compositional generation with GenEval (Ghosh et al., 2023), we follow the reward and prompt pipelines of Flow-GRPO (Liu et al., 2025). To assess performance beyond the training objectives, we report ImageReward (Xu et al., 2023), HPSv3 (Ma et al., 2025), UnifiedReward-2.0 (UniRwd2) (Wang et al., 2025b; CodeGoat24, 2026), and Aesthetic Score (Schuhmann, 2022) as held-out rewards. We further evaluate on held-out prompts from each task, the 600 UniGenBench prompts (Wang et al., 2025a) augmented with 570 physical prompts from Cosmos 3 (NVIDIA, 2026), and 200 DrawBench prompts (Saharia et al., 2022) with 5 images per prompt.

Training configuration We finetune all models with LoRA (Hu et al., 2022) (rank 32, $\alpha = 64$), using AdamW (Loshchilov & Hutter, 2019) at a learning rate of 3×10^{-4}. At each RL round, we sample 48 prompt groups with a group size of 24. Rollout images are generated at 512×512 for SD3.5 Medium and FLUX.2 Klein, and 1024×1024 for Cosmos 3 Super Text2Image. For experiments on SD3.5 Medium and FLUX.2 Klein, SIFR and RAM use an Euler sampler with 20 steps, while DiffusionNFT uses its original DPM-Solver++ sampler with its per-task step count. For experiments on Cosmos 3 Super Text2Image, all three methods use an Euler sampler with 35 steps.

4.2 OPTIMIZATION PERFORMANCE ON RL OBJECTIVES

Observation 1. SIFR more effectively optimizes the rewards used for RL training, achieving higher rewards across model scales and continuing to improve with longer training.

Model-based rewards We first compare single-reward and multi-reward training on SD3.5 Medium, FLUX.2 Klein, and Cosmos 3 Super Text2Image. Figure 2 tracks the optimized rewards on held-out prompts during training. SIFR consistently leads across all three models and both reward settings, improving rewards more quickly and reaching higher values within 500 rounds. Continuing single-reward training on Cosmos 3 Super Text2Image to 2000 rounds confirms that this advantage persists: DiffusionNFT and RAM largely plateau, while SIFR continues to improve (Figure 1a). SIFR reaches their plateau score of about 24.4 in roughly one quarter of the rounds required by the baselines. SIFR also continues to improve under multi-reward training. Tables 2 and 5 summarize the 500-round results. In particular, when PickScore, HPSv2.1, and ClipScore are jointly optimized, SIFR outperforms both baselines on each of the three scores across all three models on held-out prompts.

Rule-based rewards We next test SIFR on text rendering (OCR) and compositional generation (GenEval), which use rule-based rewards. All three methods approach saturation, with SIFR achieving the highest score in five of the six model–task comparisons (Table 3). We therefore also report the number of training rounds to first reach a score of 0.97. SIFR reaches 0.97 no later than either baseline in all six comparisons; on SD3.5 Medium OCR and both Cosmos 3 Super Text2Image tasks it needs about half the rounds of the faster baseline.

Long-caption training Cosmos 3 Super Text2Image additionally supports the long, detailed prompts produced by prompt upsampling (NVIDIA, 2026). To further test the robustness of these gains, we replace the Pick-a-Pic training prompts with 57,918 such long captions, with a median length of 185 words. On 2048 held-out long captions, SIFR reaches PickScore 25.35 at round 500, above the baseline peaks within 500 rounds, 24.56 for RAM and 23.85 for DiffusionNFT.

Method	Training rewards			Held-out rewards			
	PickScore	HPSv2.1	ClipScore	ImgRwd	HPSv3	UniRwd2	Aesthetic
SD3.5 Medium (2.5B)	21.78	0.295	0.289	1.02	9.18	3.194	5.96
+ DiffusionNFT	23.06	0.348	0.318	1.56	10.52	3.249	6.12
+ RAM	23.37	0.361	0.328	1.61	11.92	**3.294**	6.08
+ SIFR	**23.76**	**0.374**	**0.339**	**1.67**	**12.12**	3.290	**6.15**
FLUX.2 Klein (9B)	21.82	0.299	0.290	1.18	9.47	3.249	5.86
+ DiffusionNFT	23.69	0.364	0.328	1.63	12.20	3.290	6.21
+ RAM†	23.26	0.351	0.315	1.57	11.64	**3.304**	**6.23**
+ SIFR	**24.00**	**0.385**	**0.336**	**1.67**	**12.86**	3.290	**6.23**
Cosmos 3 Super T2I (64B)	22.27	0.312	0.301	1.29	9.12	3.284	5.95
+ DiffusionNFT	23.46	0.351	0.307	1.60	10.76	3.319	6.42
+ RAM	23.70	0.365	0.311	1.62	11.40	3.329	6.46
+ SIFR	**24.07**	**0.378**	**0.317**	**1.67**	**11.95**	**3.355**	**6.62**

Table 2: **Multi-reward training across model scales.** Reward scores on held-out Pick-a-Pic prompts after 500 training rounds. † RAM on FLUX.2 Klein uses the round-300 checkpoint; training subsequently collapses.

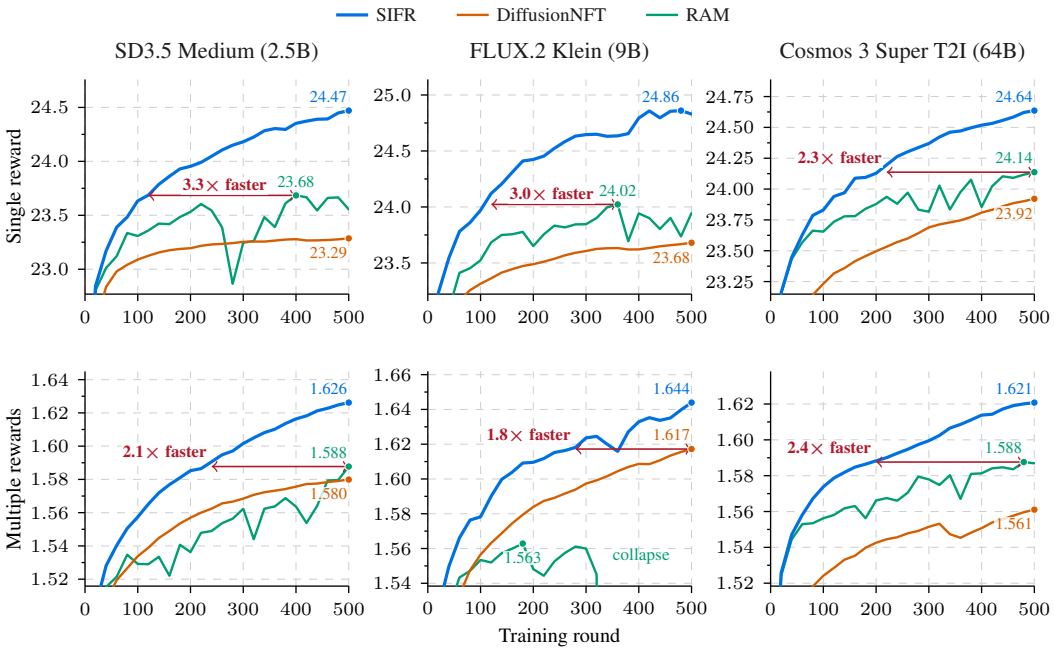
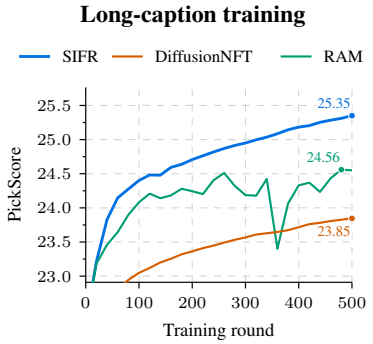


Figure 2: Training reward on held-out prompts over 500 rounds.

Method	SD3.5 Medium		FLUX.2 Klein		Cosmos 3 Super T2I	
	OCR	GenEval	OCR	GenEval	OCR	GenEval
Pretrained	0.631	0.652	0.783	0.809	0.728	0.846
DiffusionNFT	**0.985** (80)	0.972 (500)	0.988 (40)	0.971 (160)	0.976 (100)	0.975 (260)
RAM	0.974 (160)	0.973 (**270**)	0.986 (30)	0.989 (80)	0.982 (100)	0.970 (290)
SIFR	0.984 (**40**)	**0.978** (**270**)	**0.989** (**25**)	**0.994** (**60**)	**0.988** (**50**)	**0.987** (**140**)

Table 3: **Rule-based reward optimization.** Peak OCR and GenEval Overall scores within 500 rounds, with the first evaluation round reaching 0.97 in parentheses (fewer is better).



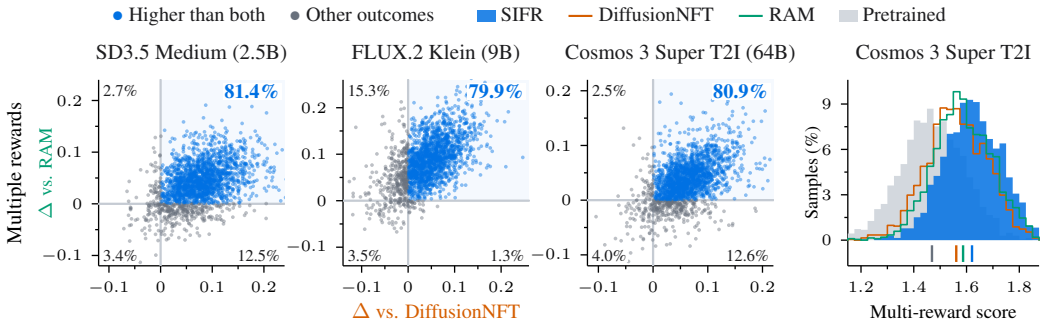


Figure 3: Per-prompt reward gains on held-out Pick-a-Pic prompts after multi-reward training; Figure 10 shows single-reward training. Each point is one prompt and noise seed: SIFR's gain over DiffusionNFT (horizontal) and RAM (vertical); blue points improve over both. Right: per-sample reward distribution on Cosmos 3 Super T2I; ticks below the axis mark the means.

4.3 Transferability across Held-Out Rewards and Prompts

Observation 2. Across SD3.5 Medium, FLUX.2 Klein, and Cosmos 3 Super Text2Image, SIFR's gains are shared by most samples and extend to held-out rewards and new prompt sets.

Per-prompt reward gains Higher mean rewards can be driven by a small number of high-scoring samples. For each backbone, we compare outputs generated from the same held-out prompts and initial noise seeds. Across all three backbones and both training objectives, SIFR achieves a higher reward than both DiffusionNFT and RAM on 79.9%–92.4% of the matched pairs (Figures 3 and 10), showing that the gains are shared by most samples.

Held-out rewards The gains also extend to reward models outside the training objective. After multi-reward training, SIFR achieves the highest ImageReward and HPSv3 scores across all three backbones (Table 2). On Cosmos 3 Super Text2Image, it leads all four held-out evaluators.

Held-out prompt sets To test transfer across prompt distributions, we evaluate the same checkpoints on UniGenBench augmented with physical prompts, DrawBench, and 2048 held-out long captions. On Cosmos 3 Super Text2Image, SIFR achieves the highest scores on all seven reported rewards across all four prompt sets (Table 4). Across all three backbones and both objectives, it also achieves the highest PickScore and Aesthetic Score on UniGenBench and DrawBench (Tables 6 and 7).

Method	Training rewards			Held-out rewards			
	PickScore	HPSv2.1	ClipScore	ImgRwd	HPSv3	UniRwd2	Aesthetic
Pick-a-Pic (held-out)							
DiffusionNFT	23.46	0.351	0.307	1.60	10.76	3.319	6.42
RAM	23.70	0.365	0.311	1.62	11.40	3.329	6.46
SIFR	**24.07**	**0.378**	**0.317**	**1.67**	**11.95**	**3.355**	**6.62**
UniGenBench + physical prompts							
DiffusionNFT	23.12	0.313	0.294	1.08	12.03	3.376	6.01
RAM	23.50	0.330	0.296	1.14	12.66	3.401	6.08
SIFR	**23.93**	**0.346**	**0.297**	**1.18**	**13.11**	**3.416**	**6.30**
DrawBench							
DiffusionNFT	23.82	0.331	0.294	1.42	12.46	3.339	5.87
RAM	24.03	0.347	0.297	1.50	13.17	3.330	5.92
SIFR	**24.36**	**0.359**	**0.299**	**1.55**	**13.64**	**3.356**	**6.24**
Long captions							
DiffusionNFT	22.79	0.297	0.274	0.84	11.95	3.554	5.79
RAM	23.19	0.312	0.277	0.89	12.54	3.586	5.90
SIFR	**23.61**	**0.326**	**0.278**	**0.98**	**13.22**	**3.600**	**6.07**

Table 4: **Transfer across prompt sets.** Reward scores for Cosmos 3 Super Text2Image on four prompt sets after 500 rounds of multi-reward training on Pick-a-Pic.

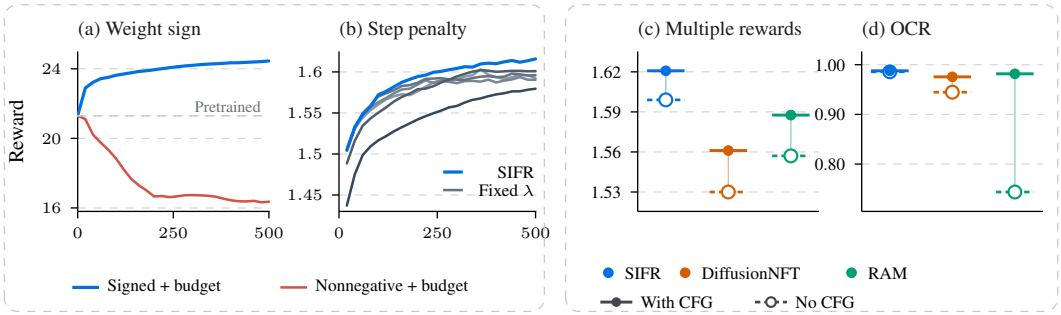


Figure 4: Ablations and evaluation guidance. Panels (a,b) plot reward against training round. (a) On SD3.5 Medium PickScore training, removing negative advantages causes collapse. (b) On FLUX.2 Klein multi-reward training, the updated step multiplier reaches a higher final reward than every fixed coefficient. (c,d) Cosmos 3 Super Text2Image checkpoints evaluated with and without CFG.

4.4 Ablation studies

Observation 3. Signed advantages and the drift constraint each contribute to stable training, and together they make RL faster and raise its reward ceiling. We further test SIFR's robustness and find that its gains persist with and without CFG.

Negative advantages We first replace $A(X_1)$ with $\max\{A(X_1), 0\}$, keeping the remaining training settings unchanged. On SD3.5 Medium, both nonnegative variants collapse, with or without the drift penalty (Figure 4a). Retaining negative advantages avoids this failure, showing that regularization alone cannot replace the unfitting signal.

Multiplier update We next compare the multiplier updated by eq. (9) with five fixed coefficients between 0.01 and 0.5 on FLUX.2 Klein multi-reward training (Figure 4b). At round 500, the updated multiplier yields higher reward than every fixed choice. Figure 5 shows an example of how the multiplier changes under feedback control during training.

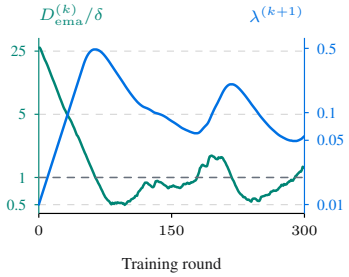


Figure 5: Feedback control example.

Evaluation guidance Finally, we evaluate each Cosmos 3 Super Text2Image checkpoint with and without CFG, using the same prompts and noise seeds (Figure 4c,d). SIFR leads both objectives in both settings. Even without CFG, it exceeds the guided baselines.

5 Conclusion

We presented SIFR, which combines signed advantages with a constraint on velocity drift whose multiplier adapts to a prescribed budget. The signed advantages supply the unfitting signal that nonnegative reward weighting lacks, and the adaptive constraint keeps signed regression stable; together they make RL faster and raise its reward ceiling. Across all 12 backbone–objective pairs on SD3.5 Medium, FLUX.2 Klein, and Cosmos 3 Super Text2Image, SIFR matches or exceeds DiffusionNFT and RAM; on preference rewards it reaches their peak scores two to four times faster and keeps improving where they plateau, and its gains transfer to held-out rewards and prompt sets.

AI use statement

We used large language models to assist with writing and revising the manuscript, including improving grammar, clarity, and organization. The authors reviewed the AI-assisted revisions for technical accuracy and consistency with the intended meaning, and take responsibility for the final content of the paper.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. Blog post and open-weight model release, 2026. URL https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

CodeGoat24. UnifiedReward-2.0-qwen35-9b. Hugging Face model checkpoint, 2026. URL https://huggingface.co/CodeGoat24/UnifiedReward-2.0-qwen35-9b.

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 *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.08861.

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 *International Conference on Machine Learning (ICML)*, volume 235, pp. 12606–12633, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with Wasserstein regularization. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9690d4746230cfea3d067fca695ba648-Abstract-Conference.html. arXiv:2502.06061.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zheming Fu, Ruizhe He, Wei Shang, Xiaoxiao Ma, Lei Wang, Chang Liu, and Siming Fu. FlowAWR: Online adaptive flow reinforcement via advantage-weighted rectification. *arXiv preprint arXiv:2606.30376*, 2026.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 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, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2210.01504.

Hilbert J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, 2005.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Branislav Kveton, Anup Rao, Subhojyoti Mukherjee, Krishna Kumar Singh, and Viet Dac Lai. AdvantageFlow: Advantage-weighted least squares for RL in flow models. *arXiv preprint arXiv:2605.26013*, 2026.

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.

Runlong Liao, Baiyu Su, Lizhang Chen, and Qiang Liu. Signed Rectified Flow: Negativity-controlled generation. *arXiv preprint arXiv:2607.18516*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.05470.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

Qiang Liu. Let us flow together. Lecture notes (working draft), 2024. URL https://www.cs.utexas.edu/~lqiang/PDF/flow_book.pdf. Version of December 24, 2024.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_HPSv3_Towards_Wide-Spectrum_Human_Preference_Score_ICCV_2025_paper.html.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2401.06121.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 *International Conference on Machine Learning (ICML)*, 2007.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, 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 (NeurIPS)*, volume 35, pp. 36479–36494, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Christoph Schuhmann. CLIP+MLP Aesthetic Score Predictor. GitHub repository, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor. Checkpoint: sac+logos+ava1-l14-linearMSE.pth.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015.

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, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Stability AI. Stable Diffusion 3.5 Medium. Hugging Face model checkpoint, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *International Conference on Machine Learning (ICML)*, 2020.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2019.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2006.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. UniGenBench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025a. URL https://arxiv.org/abs/2510.18701.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b. URL https://arxiv.org/abs/2503.05236.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning RL with pretraining in diffusion models. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2509.25050.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.05868.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.16117.