This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Fine-Grained GRPO for Precise Preference Alignment in Flow Models

Yujie Zhou[1,4*] **Pengyang Ling**[2,4*] **Jiazi Bu**[1,4*] **Yibin Wang**[3,5]
Yuhang Zang[4] **Jiaqi Wang**[4,5†] **Li Niu**[1,6†] **Guangtao Zhai**[1]
[1]Shanghai Jiao Tong University [2]University of Science and Technology of China
[3]Fudan University [4]Shanghai AI Laboratory [5]Shanghai Innovation Institute [6]Miguo.ai
https://bujiazi.github.io/g2rpo.github.io/

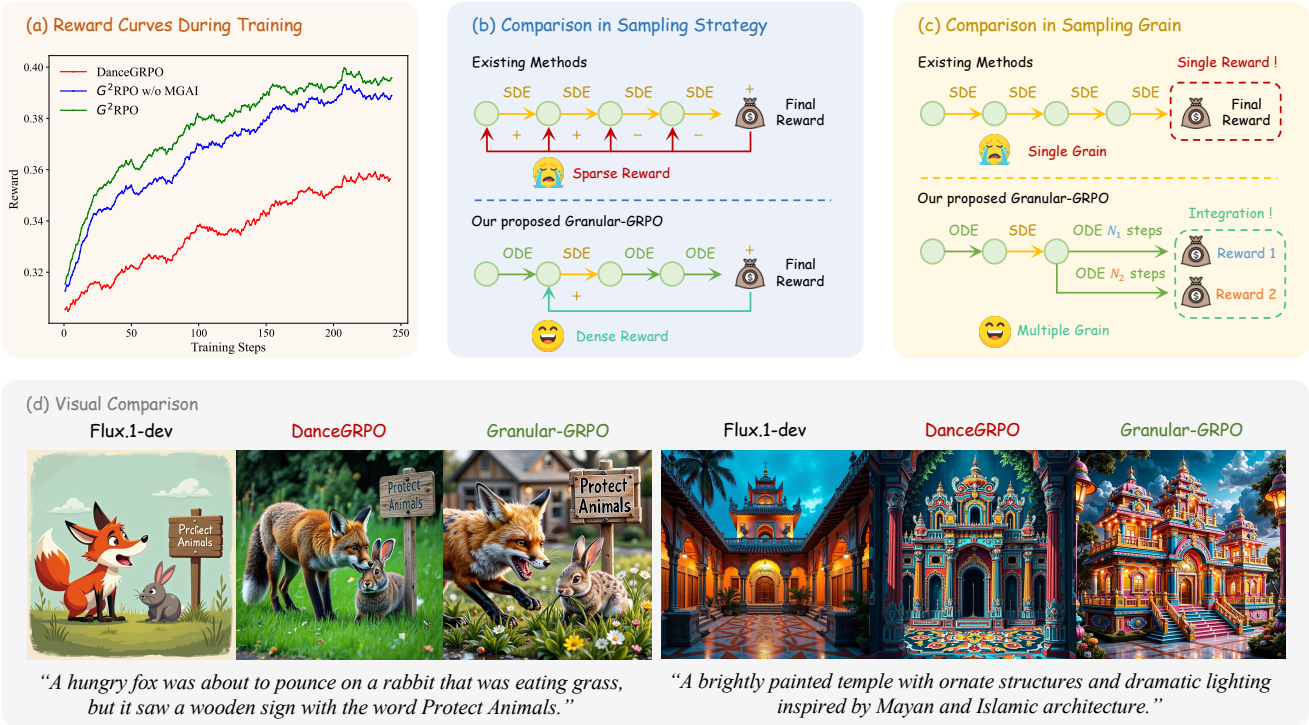

Figure 1. (a) G^{2}RPO outperforms DanceGRPO in reward scores (HPS-v2.1 in this figure). (b) Sampling strategy comparison. G^{2}RPO acquire a dense reward by confining stochasticity to individual sampling steps. (c) Sampling grain comparison. G^{2}RPO achieves a comprehensive evaluation of each sampling direction by integrating advantages from multi-granularity ODE denoising. (d) Visual Comparison. Compared to the baseline method, the images generated by G^{2}RPO are more aligned with human preferences.

Abstract

*The incorporation of online reinforcement learning (RL) into diffusion and flow-based generative models has recently gained attention as a powerful paradigm for aligning model behavior with human preferences. By leveraging stochastic sampling via Stochastic Differential Equations (SDEs) during the denoising phase, these models can explore a variety of denoising trajectories, enhancing the exploratory capacity of RL. However, despite their ability to discover potentially high-reward samples, current approaches often struggle to effectively align with preferences due to the sparsity and narrowness of reward feedback. To overcome this limitation, we introduce a novel framework called Granular-GRPO (**G^{2}RPO**), which enables fine-grained and comprehensive evaluation of sampling directions in the RL training of flow models. Specifically, we propose a **Singular Stochastic Sampling** mechanism that supports step-wise stochastic exploration while ensuring strong*

*Equal contribution. † Corresponding authors.

correlation between injected noise and reward signals, enabling more accurate credit assignment to each SDE perturbation. Additionally, to mitigate the bias introduced by fixed-granularity denoising, we design a ***Multi-Granularity Advantage Integration*** *module that aggregates advantages computed across multiple diffusion scales, resulting in a more robust and holistic assessment of sampling trajectories. Extensive experiments on various reward models, including both in-domain and out-of-domain settings, demonstrate that G^{2}RPO outperforms existing baselines, highlighting its effectiveness and generalization capability.*

1. Introduction

Recent advances in generative models, particularly diffusion models [15, 35, 36] and flow models [23, 25, 27], have revolutionized visual content creation, offering unprecedented capabilities in generating high-quality images [6, 7, 10, 19, 28, 31] and videos [4, 5, 8, 13, 18, 22, 39, 46]. However, a key challenge remains in aligning model outputs with the diverse and complex human preferences. To tackle this challenge, reinforcement learning from human feedback [3, 11, 42] has emerged as a promising solution, including paradigms such as Proximal Policy Optimization (PPO) [32], Direct Policy Optimization (DPO) [30], and Group Relative Policy Optimization (GRPO) [34]. Among these, GRPO stands out as an innovative online reinforcement learning approach. Leveraging group comparisons to optimize policies, GRPO eliminates the need for a separate value model, achieving greater flexibility and scalability.

To integrate GRPO into flow-based generative models, Flow-GRPO [24] and DanceGRPO [45] substitute the deterministic ODE sampler with an SDE formulation, wherein the injected stochasticity deliberately perturbs the denoising direction at each step. Although the resulting samples enable exhaustive per-step exploration for reinforcement learning, they simultaneously underscore the difficulty of attributing the final reward to any specific random perturbation, thereby constraining model trainability. Specifically, most existing flow-based GRPO methods encounter two core issues in evaluating group denoising directions: 1) **Sparse reward**: As shown in Fig. 1 (b), the final reward signal is uniformly assigned to each SDE sampling step, which cannot be precisely aligned with the sampling direction at each step, leading to inaccuracies for the optimization at individual steps. 2) **Incomplete evaluation**: While the same single-step SDE sampling direction can yield different rewards under varying denoising intervals, an ideal high-quality direction (Sample 2 and 3 shown in Fig. 3) intuitively is robust to the granularity of the remaining sampling trajectory. This means that although the absolute quality of the final results may vary with the number of steps, an excellent denoising direction still maintains a superior relative position within the group. Nevertheless, as shown in Fig. 1(c), existing methods suffer from the drawback that each denoising direction is rigidly associated with a fixed number of denoising steps. Consequently, it leads to a single granularity of the denoised images, making it impossible to conduct a thorough comparison across the group.

To address these limitations, we propose **G**ranular-**GRPO** (G^{2}RPO), a novel online reinforcement learning framework specifically designed for precise and comprehensive reward signals. First, mirroring the sparse-reward problem [14, 21] that plagues RLHF, the reward signal in the SDE sampling process is delivered only after an entire sequence of decisions. This long delay undermines the credit-assignment chain, preventing the linking of the terminal reward with any specific earlier action and thereby inducing sluggish, unstable learning. Therefore, we propose a simple yet effective sampling strategy, termed **Singular Stochastic Sampling**. As illustrated in Fig 1 (b), this strategy applies the SDE formulation at a single time step to generate a group of denoising directions, while employing deterministic ODE sampling for all other steps. By concentrating stochasticity at one specific step, the proposed method establishes a strong correlation between the reward signal and the injected noise, enabling stable model optimization. Secondly, we propose a **Multi-Granularity Advantage Integration** (MGAI) module. As depicted in Fig. 1(c), instead of binding each denoising direction to a fixed subsequent denoising granularity, the denoising directions in the same group are assigned to a spectrum of denoising steps, producing images with different granularities. The corresponding reward signals of these images are then fused into a unified advantage estimate, yielding a comprehensive evaluation of the current state's value. With the support of the Singular Stochastic Sampling strategy and the Multi-Granularity Advantage Integration module, G^{2}RPO can provide a more precise and comprehensive reward signal, thereby enhancing the upper limit of the GRPO model training. As shown in Fig 1 (a), our reward curves exhibit stable improvements over the baseline during training. Additionally, Fig 1 (d) illustrates the images generated by G^{2}RPO, highlighting its advantages in text prompt adherence and detail fidelity. Our contributions are summarized:

- **Granular-GRPO**: A novel flow-based GRPO framework designed to provide a precise and comprehensive evaluation of the denoising directions sampled by the SDE, thereby improving the precision of model optimization.
- **Singular Stochastic Sampling**: A sampling strategy confines stochasticity to individual sampling steps, addressing the sparse reward issue associated with long-range stochasticity injection.
- **Multi-Granularity Advantage Integration**: A module aggregates multi-scale advantages to enable a comprehensive evaluation of each sampling direction.

2. Related Work

2.1. Alignment for Large Language Models

Recent years have witnessed a paradigm shift from supervised fine-tuning [9, 37] to online reinforcement learning [1, 16, 33] when aligning Large Language Models (LLMs) with human intent, which is known as Reinforcement Learning from Human Feedback (RLHF) [2, 26]. Early RLHF pipelines typically involve training a reward model from pairwise comparisons to predict human preferences and guide a policy model through reinforcement learning algorithms like Proximal Policy Optimization (PPO) [32]. Despite their effectiveness, PPO introduces intensive computational overhead due to its iterative optimization process and the need for frequent interactions with the environment. To address these computational challenges, value-free alternatives have been proposed, such as Group Relative Policy Optimization (GRPO) [34]. Adopted by leading LLM DeepSeek-R1 [12], GRPO aims to optimize policies based on relative preferences within a group of samples, providing a robust signal for policy improvement, particularly when absolute rewards are difficult to define or noisy. These advancements in LLM alignment provide a strong foundation for exploring similar human-centric optimization strategies in visual generation domains.

2.2. Alignment for Flow Models

Diffusion and Flow models [15, 27, 31, 35, 36], which offer flexible visual creation through an iterative denoising process, have revolutionized the field of visual synthesis and become a pivotal part of generative models. Building on the success of aligning LLMs with human preferences, similar techniques have recently been transplanted to diffusion and flow models [10, 19, 28]. Pioneer works like DDPO [3] and ReFL [44] apply PPO to finetune diffusion models for improved aesthetic performance and human feedback alignment. These methods face challenges inherent to RL, including high variance, low efficiency, and sparse reward. Diffusion-DPO [38] adapts the Direct Preference Optimization (DPO) [30] framework to directly optimize diffusion models from paired preference data, bypassing the need for an explicit reward model but suffering from distribution shift since no new samples are collected during training. Recent efforts such as DanceGRPO [45] and Flow-GRPO [24] enable GRPO-style policy updates by converting the ODE sampling into an equivalent SDE to each timestep, thereby acquiring a group of denoising directions for statistical sampling and RL exploration for flow models. More recently, MixGRPO [20] has improved training efficiency through a hybrid ODE-SDE sampling approach while maintaining comparable performance. However, these methods are generally constrained by sparse rewards due to long-range stochasticity injection and the binding of each sampling direction to a fixed denoising granularity. These paradigms restrict the ability to conduct a comprehensive evaluation of each sampling direction, limiting the optimization ceiling of GRPO training.

3. Granular-GRPO

As an online RL algorithm, flow-based GRPO methods utilize ODE-to-SDE conversion with the original model's marginal distribution to sample a group of denoising directions for optimization. A core issue underlying this paradigm is to obtain precise and comprehensive assessments of each sampling direction. To this end, we introduce the Granular-GRPO framework to (i) confine stochasticity to individual steps (Singular Stochastic Sampling) for more precise reward signals, and (ii) integrate the advantages derived from multi-granularity denoising results (Multi-Granularity Advantage Integration) to acquire a more comprehensive evaluation, as shown in Fig. 2.

3.1. Preliminary

For the flow-based GRPO methods [20, 24, 40, 45], the denoising process is first modeled as a multi-step Markov decision process (MDP). Given a prompt $\boldsymbol{c}$, the agent with a flow model p_θ produce a reverse-time trajectories defined as $\Gamma = (\mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T-1}, \mathbf{a}_{T-1}, \ldots, \mathbf{s}_0, \mathbf{a}_0)$, where $\mathbf{s}_t = (\boldsymbol{c}, t, \boldsymbol{x}_t)$ is the state at timestep t and $\boldsymbol{x}_t$ is the corresponding noisy sample. Specifically, $\boldsymbol{x}_T \sim N(0, I)$ and $\boldsymbol{x}_0$ is the denoised image. The action $\mathbf{a}_t$ represents the single step denoising process with the policy π_θ, indicating a sampling direction $\frac{d\boldsymbol{x}}{dt}$ from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$, i.e. $\boldsymbol{x}_{t-1} \sim \pi_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c})$.

SDE Sampling. As an online RL algorithm, GRPO needs grouped outputs and relative advantages to optimize policies. However, the flow matching model utilizes a deterministic ODE to sample the denoising direction:

$$d\boldsymbol{x}_t = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t)dt, \tag{1}$$

where, $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ is the model output, given the noisy sample $\boldsymbol{x}_t$ and timestep t.

To match GRPO's stochastic sampling requirements, Flow-GRPO [24] converts the ODE into an equivalent SDE sampling with the same marginal distribution:

$$d\boldsymbol{x}_t = (\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) + \frac{\sigma_t^2}{2t}(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)))dt + \sigma_t d\boldsymbol{w}_t, \tag{2}$$

where $d\boldsymbol{w}_t$ denotes Wiener process increments, and σ_t controls the stochasticity injected into the sampling direction. Furthermore, it can be discretized via the Euler–Maruyama scheme:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \big(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) + \tfrac{\sigma_t^2}{2t}(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t))\big)\Delta t + \sigma_t\sqrt{\Delta t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

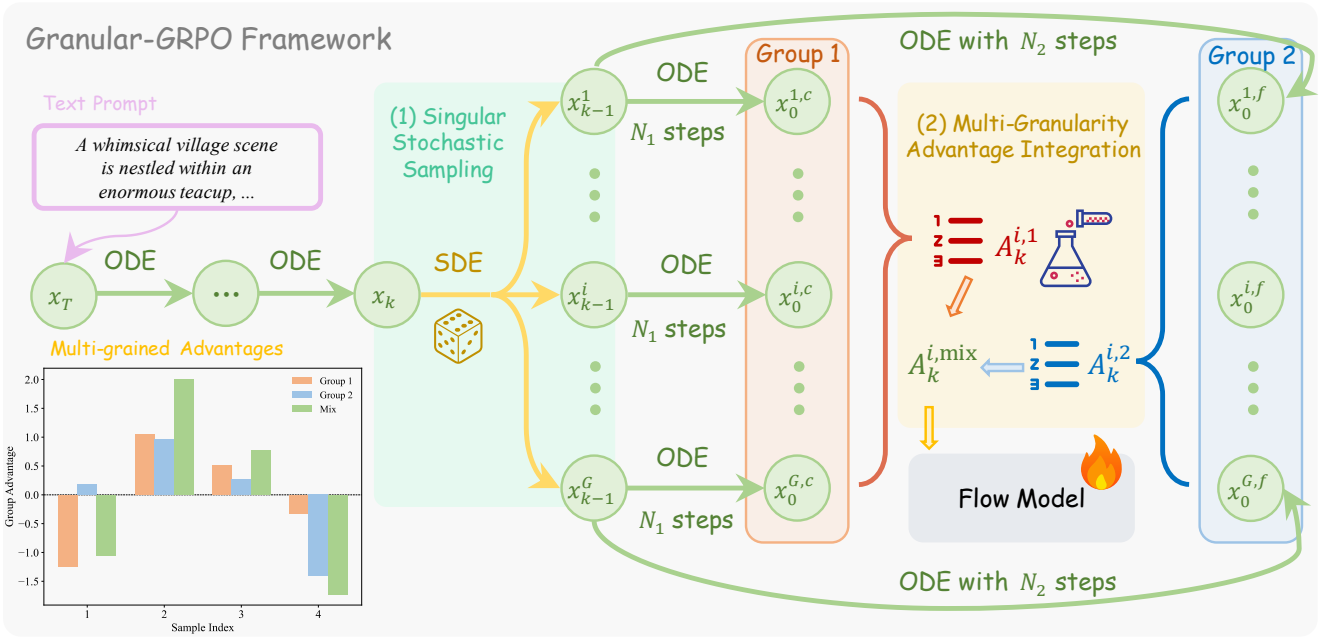

Figure 2. **Overview of G^{2}RPO**. Given a text prompt and an initial noise, our Singular Stochastic Sampling strategy employs SDE sampling solely at a single step and samples a group of distinct denoising directions. Then, the Multi-Granularity Advantage Integration module executes multi-granularity ODE denoising for each direction and integrates the advantages to produce a comprehensive evaluation for each sampling direction. For simplicity, the figure shows one coarse-grained path (denoted as c) and one fine-grained path (denoted as f).

As defined in Flow-GRPO, $\sigma_t = \eta\sqrt{\frac{t}{1-t}}$, where the noise level is controlled by hyperparameter η.

GRPO Training. With SDE sampling, flow-based GRPO methods introduce stochasticity at each timestep to generate a group of G images $\{\boldsymbol{x}_0^i\}_{i=1}^G$. Then the reward model assigns a score $R(\boldsymbol{x}_0^i, \boldsymbol{c})$ to $\boldsymbol{x}_0^i$, and the advantage is computed as:

$$A_t^i = \frac{R(\boldsymbol{x}_0^i, \boldsymbol{c}) - \text{mean}(\{R(\boldsymbol{x}_0^j, \boldsymbol{c})\}_{j=1}^G)}{\text{std}(\{R(\boldsymbol{x}_0^j, \boldsymbol{c})\}_{j=1}^G)}. \tag{4}$$

Note that the advantages A_0^i obtained from the final step image are uniformly broadcast to each step A_t^i to evaluate the SDE sampling directions. Finally, the policy model is optimized by maximizing the following objective:

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}_{\boldsymbol{c}\sim\mathcal{C},\{\boldsymbol{x}^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{c})} f(r, A, \theta, \varepsilon, \beta), \tag{5}$$

where

$$f(r, A, \theta, \varepsilon, \beta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}(\min(r_t^i(\theta)A_t^i, \text{clip}(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon)A_t^i) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{6}$$

$$r_t^i(\theta) = \frac{p_\theta\left(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, \boldsymbol{c}\right)}{p_{\theta_{\text{old}}}\left(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, \boldsymbol{c}\right)}. \tag{7}$$

Notably, β is the hyperparameter that controls the proportion of KL loss. Following the practices of DanceGRPO and MixGRPO, $\beta = 0$ to achieve a stable training process.

3.2. Singular Stochastic Sampling

Traditional flow-based GRPO methods introduce ODE-to-SDE conversion sampling at every step to inject stochasticity and uniformly assign the final image reward to each step's sampling direction. According to Eq. 4, the advantage A_t^i of each sampling direction at step t is equally assigned with A_0^i. Notably, the reward signal available only after multiple decision steps impedes the model's capability to link the final reward to each decision, thereby resulting in imprecise and sparse rewards.

In accordance with the design of DanceGRPO, we define M as the first half of the denoising timesteps, i.e., $M = \{T, T-1, \ldots, \lfloor T/2 \rfloor\}$. This selection prioritizes the early denoising steps, where the exploration space of the SDE is larger and significantly determines the overall direction of the entire denoising chain. In contrast, the improvements from GRPO in the later stages are relatively minor. Consequently, training is conducted exclusively on these timesteps after each round of group image sampling, thereby optimizing computational efficiency. Moreover, to acquire a dense reward for each SDE sampling direction, a simple yet effective strategy is to confine the stochasticity to the single step selected for optimization. As shown in Fig. 2, given a prompt $\boldsymbol{c}$ and initial noise $\boldsymbol{x}_T$, for each timestep $k \in M$, a common starting point $\boldsymbol{x}_k$ for the group

is acquired using ODE sampling from Eq. 1. Then, our Singular Stochastic Sampling strategy employs SDE sampling only at $\boldsymbol{x}_k$ and samples G distinct denoising directions to get the next noisy state $\{\boldsymbol{x}_{k-1}^i\}_{i=1}^G$. Each $\boldsymbol{x}_{k-1}^i$ undergoes $k-1$ steps of ODE sampling to generate a deterministic denoised image $\boldsymbol{x}_{0\leftarrow k}^i$. Based on our sampling strategy, the variance of the group reward $\{R(\boldsymbol{x}_{0\leftarrow k}^i, \boldsymbol{c})\}_{i=1}^G$ is entirely determined by the distinct denoising directions introduced by the SDE sampling at step k. A step-aware, precise advantage can be acquired:

$$A_k^i = \frac{R(\boldsymbol{x}_{0\leftarrow k}^i, \boldsymbol{c}) - \text{mean}(\{R(\boldsymbol{x}_{0\leftarrow k}^i, \boldsymbol{c})\}_{i=1}^G)}{\text{std}(\{R(\boldsymbol{x}_{0\leftarrow k}^i, \boldsymbol{c})\}_{i=1}^G)}. \quad (8)$$

The $f(r, A, \theta, \varepsilon, \beta)$ in Eq 5 can be formulated as:

$$f(r, A, \theta, \varepsilon, \beta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{K}\sum_{k\in M}\Bigg(\min\Big(r_k^i(\theta)A_k^i, \text{clip}\big(r_k^i(\theta), 1-\varepsilon, 1+\varepsilon\big)A_k^i\Big)\Bigg). \quad (9)$$

After the sampling phase is completed, the training of GRPO requires the computation of $p_\theta\left(\boldsymbol{x}_{k-1}^i \mid \boldsymbol{x}_k^i, \boldsymbol{c}\right)$ to obtain $r_k^i(\theta)$ refer to eq 7. In practice, each distinct sampling starting point $\boldsymbol{x}_k^i$ needs to be fed into the flow model to compute the corresponding ODE denoising direction $\boldsymbol{v}_k^i$. However, our sampling strategy shares a common starting point $\boldsymbol{x}_k$, allowing a group of G samples to reuse the same $\boldsymbol{v}_k$, which in turn improves training efficiency.

3.3. Multi-Granularity Advantage Integration

The proposed Singular Stochastic Sampling accurately constrains stochasticity into a single SDE step, where terminal signals are traced back to specific intermediate actions ensuring a causal density between the reward and the injected noise. Nevertheless, how to acquire a comprehensive reward signal for the denoising direction of the current step still requires further investigation.

As shown in Fig. 3, we observe that under identical $\boldsymbol{x}_k$ and prompt $\boldsymbol{c}$ conditions, the denoising trajectory generated by singular stochastic sampling is not robust when assessing the corresponding SDE denoising direction. Specifically, while the overall content of the images generated from different denoising trajectories remains similar, the details vary significantly due to the differing denoising intervals. This variation in denoising granularity leads to discrepancies in the scores assigned by the reward model, which in turn influences the numerical values of the advantage within the group and even the optimization direction.

To conduct a comprehensive assessment of the single-step SDE sampling direction, we propose a Multi-Granularity Advantage Integration module, which performs multi-granularity denoising on the sampled denoising directions within a group. The advantages of the images denoised at different granularities are then integrated to form the final evaluation. Specifically, as shown in Fig. 2, step k is the SDE sampling step, and G distinct denoising directions are sampled to acquire next noisy state $\{\boldsymbol{x}_{k-1}^i\}_{i=1}^G$. Under the conventional granularity condition, each $\boldsymbol{x}_{k-1}^i$ undergoes $k-1$ steps to obtain the final denoised image. The sequence of denoising timesteps can be represented as: $\mathcal{S} = \{1, 2, \ldots, k-1\}$. For our Multi-Granularity denoising module, a set of integer scaling factors $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_j\}, |\Lambda| = J$ is defined to represent different denoising granularities. Each λ_j implements interval sampling for different denoising granularity, that means sample every λ_j-th step from the total $k-1$ steps. The denoising timestep sequence $\mathcal{S}_j$ can be formally represented as:

$$\mathcal{S}_j = \{1, 1+\lambda_j, 1+2\lambda_j, \ldots, \left\lceil\frac{k-1}{\lambda_j}\right\rceil\lambda_j\}. \quad (10)$$

Our interval sampling approach ensures that the denoising process is performed at regular intervals defined by λ_j. As λ_j increases, the granularity becomes coarser, allowing for a more flexible and adaptive denoising process. For ease of illustration, $J = 2$ in the Fig. 2.

After N_j subsequent steps denoising for $\{\boldsymbol{x}_{k-1}^i\}_{i=1}^G$, a group of noise-free images $\{x_0^{i,j}\}_{i=1}^G$ are generated. Subsequently, different groups images gets the reward $\{R(\boldsymbol{x}_{0\leftarrow k}^{i,j}, \boldsymbol{c})\}_{i=1}^G$ from a reward model and then compute the intra-group advantages $\{A_t^{i,j}\}_{i=1}^G$ with Eq. 4. Similar to the joint training with multiple reward models (e.g., HPS-v2.1 and CLIP Score) in DanceGRPO, where the advantages from different reward models are directly summed to provide a multi-dimensional comprehensive evaluation of a single denoising direction, we combines the advantages from different granularities to get $\{A_t^{i,\text{mix}}\}_{i=1}^G$:

$$A_t^{i,\text{mix}} = \sum_{j}^{J} A_t^{i,j}. \quad (11)$$

Eq. 9 can be updated to the following formula:

$$f(r, A, \theta, \varepsilon, \beta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{K}\sum_{k\in M}(\min(r_k^i(\theta)A_k^{i,\text{mix}}, \text{clip}\left(r_k^i(\theta), 1-\varepsilon, 1+\varepsilon\right)A_k^{i,\text{mix}})). \quad (12)$$

Finally, we use Eq. 5 to optimize the policy π_θ across all timesteps in the set M, and a detailed algorithm is illustrated in Algorithm 1.

In summary, Granular-GRPO constrains the noise injection range through the Singular Stochastic Sampling to produce causally dense reward-stochasticity correspondence, while the MGAI module further enables comprehensive and robust assessment along each SDE sample, ultimately driving the policy to jointly enhance detail fidelity and global layout coherence.

Figure 3. **Visual Comparison of Images Denoised at Different Granularities.** Images denoised at different granularities exhibit variations in fine details and textures, leading to inconsistent scoring by the Reward Model (HPS-v2.1). This observation reveals the insufficiency of a single-granularity evaluation of group advantage.

Algorithm 1 Granular-GRPO Training Process

Require: Prompt dataset $\mathcal{C}$, policy model π_θ, reward model R, total sampling steps T
Require: SDE sampling timestep set M, Denoising granularities set Λ ($|\Lambda| = J$)
for training iteration $e = 1$ **to** E **do**
  Update old policy model: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Sample batch prompts $\mathcal{C}_b \sim \mathcal{C}$
  for prompt $\mathbf{c} \in \mathcal{C}_b$ **do**
    Init same noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$
    for $k \in M$ **do**
      for $t = T$ to 0 **do**
        if $t > k$ **then**
          ODE Sampling: $\boldsymbol{x}_{t-1}$
        else if $t == k$ **then**
          SDE Sampling a group samples: $\boldsymbol{x}^i_{k-1}$
        else if $t < k$ **then**
          for $\lambda_j \in \Lambda$ **do**
            ODE Sampling with granularity λ_j: $\boldsymbol{x}^{i,j}_{t-1}$
          end for
        end if
      end for
      Get a group of reward: $R(\boldsymbol{x}^{i,j}_{0\leftarrow k})$
      $A^j_k \leftarrow \frac{R(\boldsymbol{x}^{i,j}_{0\leftarrow k}, \mathbf{c}) - \mu^j}{\sigma^j}$
      $A^{\text{mix}}_k \leftarrow \sum_{j=1}^{J} A^j_k$
    end for
    Compute GRPO loss $J(\theta)$
  end for
  Update policy: gradient ascent on $J(\theta)$
end for

4. Experiments

4.1. Implementation Details

Datasets and Backbone. Following the same setting in DanceGRPO [45] and MixGRPO [20], the prompt dataset HPSv2 [43] is utilized. It contains 103,700 text prompts for training and 400 diverse prompts for testing. The text-to-image model employed for reinforcement learning is Flux.1-dev [19], a leading flow model in the community.

Evaluation Metrics. To evaluate the effectiveness and robustness of our G^{2}RPO, multiple reward models are applied as evaluation metrics. These reward models assess the alignment between generated images and human preferences from multiple dimensions. Specifically, HPS-v2.1 **(HPS)** [43], CLIP Score **(CLIP)** [29], and Pick Score **(PS)** [17] collectively assess semantic alignment and visual coherence. Image Reward **(IR)** [44] focuses on visual quality and aesthetic appeal, while Unified Reward **(UR)** [41] is the SOTA reward model that leverages the powerful visual understanding capabilities of LLM comprehensively evaluates both alignment with the caption and overall quality.

Evaluation Setting. Similar to DanceGRPO and MixGRPO, two experimental settings are employed. Firstly, a single HPS-v2.1 reward model is utilized for training to verify the upper limit of improvement for *in-domain performance*, i.e., both training and testing are conducted using the same reward model. However, as demonstrated by DanceGRPO, HPS-v2.1 is prone to model hacking due to biases in the training set, leading to degradation in other evaluation metrics. Similarly, our primary experiment involves joint training with both HPS-v2.1 and CLIP Score as reward models to acquire stable and robust results. Sub-

"In the dust of the abandoned attic, there is a silver locket box in an open music box. The swan embossed on it is exactly the same as the carved pattern on the lid of the music box. Film texture."

"On the chessboard, the black king chess piece was smiling proudly, and beside it was a sad white king chess piece. The game was decided."

"In the ruined city, a little boy looked up in surprise at the huge, still robot in front of him, with a cinematic lens and a realistic style."

"The texture of the movie. An elderly historian wearing white cotton gloves carefully examined a yellowed sheepskin scroll map with a magnifying glass, with a solemn expression."

"In the Wolong Nature Reserve, there are two giant pandas. One adult giant panda is leisurely eating bamboo, and the other cub is lying on its back curiously."

"A public welfare poster has a clear dividing line in the middle of the picture. On the left is the dry and cracked land and withered trees, and on the right is the vibrant oasis and clear lake water."

Figure 4. **Qualitative Results.** Comparison with existing flow-based GRPO methods, in which our G^{2}RPO demonstrates superior performance in human preference alignment.

sequently, we evaluate the model using *out-of-domain metrics*, i.e., reward models that were not seen during training.

Sampling Phase. Following DanceGRPO, a shared initialization noise is used to generate a group of 12 images from the same text prompt. The total sampling step $T = 16$ to enhance computational efficiency and the advantage clip $\varepsilon = 5$ in Eq. 6. The parameter $\eta = 0.7$ in Eq. 3 directly determines the noise level, which in turn defines the size of the stochastic exploration space in the SDE. All comparison methods share the aforementioned parameter configuration. For the Multi-Granularity Advantage Integration strategy, the set of distinct granularities $\Lambda = \{1, 2, 3\}$.

Training Phase. All experiments are conducted using 16× NVIDIA H200 GPUs, with a batch size of 1. The AdamW optimizer is used, configured with a learning rate of 2×10^{-6} and a weight decay of 1×10^{-4}. Mixed precision training is implemented using bfloat16 (bf16) format.

Table 1. **Quantitative Results.** Comparison of results on various reward metrics.

Reward Model	Method	HPS	CLIP	PS	IR	UR
/	Flux.1-dev	0.305	0.388	0.226	1.040	3.621
HPS	DanceGRPO	0.353	**0.375**	0.228	1.233	**3.548**
	MixGRPO	0.378	0.358	0.225	1.266	3.421
	G^{2}RPO w/o MGAI	0.376	0.351	0.228	1.286	3.469
	G^{2}RPO	**0.385**	0.355	**0.229**	**1.313**	3.487
HPS & CLIP	DanceGRPO	0.331	0.389	0.227	1.128	3.569
	MixGRPO	0.363	0.399	0.230	1.436	3.661
	G^{2}RPO w/o MGAI	0.372	0.395	0.234	1.421	3.688
	G^{2}RPO	**0.376**	**0.406**	**0.235**	**1.483**	**3.783**

4.2. Main Results

Quantitative Evaluation. As shown in Tab. 1, DanceGRPO and MixGRPO serve as the baselines for comparison with our G^{2}RPO. Flow-GRPO is not included since

Table 2. **Ablation Study.** Comparison for different denoising granularities.

Metrics	Λ	HPS	CLIP	PS	IR	UR
HPS & CLIP	$\{1\}$	0.372	0.395	0.234	1.421	3.688
	$\{1,2\}$	0.375	0.404	0.234	1.468	3.759
	$\{1,3\}$	**0.378**	0.404	0.234	1.465	3.760
	$\{1,2,3\}$	0.376	**0.406**	**0.235**	**1.483**	**3.783**

it employs a, similar ODE-to-SDE conversion across all timesteps as contemporary DanceGRPO, but fit a specific reward target. In addition, the Singular Stochastic Sampling strategy without multi-granularity denoising (G^2RPO w/o MGAI) is also proposed. It can be observed that when HPS-v2.1 is used solely as the training reward model, our Singular Stochastic Sampling achieves a relative improvement of 6.52% compared to DanceGRPO. This indicates that constraining the stochasticity to a single step yields a precise reward signal, which in turn provides a more faithful optimization signal and enables GRPO to enhance the optimization ceiling. However, as demonstrated by DanceGRPO, optimizing solely with HPS-v2.1 can induce model hacking, which in turn compromises other out-of-domain evaluation metrics. Secondly, for the setting of multi-reward (HPS-v2.1 and CLIP Score) optimization, the results indicate that G^2RPO outperforms other baselines across both in-domain and out-of-domain rewards. Under the multi-granularity denoising condition, groups of images at different granularities provide a more comprehensive evaluation of the SDE sampling direction. This robust assessment facilitates improvements in out-of-domain metrics, particularly on Unified Reward enhanced with LLM knowledge.

Qualitative Comparison. Fig. 4 presents the qualitative comparison among the original Flux.1-dev, DanceGRPO, MixGRPO, and our proposed G^2RPO. It can be observed that G^2RPO provides enhanced detail fidelity and improves the consistency with the text prompt, achieving superior alignment with human preferences. For instance, in the second column, our G^2RPO faithfully captures the specified expressions and even the nuances of the chess pieces as described in the prompt, delivering finer details and higher visual quality. Moreover, in the "poster" case depicted in the last column, G^2RPO not only adheres to the spatial requirement of a clear left-right demarcation but also renders the reflections of the trees with remarkable clarity. The overall style of the image generated by G^2RPO is more consistent with the aesthetic demands of poster design.

4.3. Ablation Study

As described in Section 3, the Multi-Granularity Advantage Integration module integrates multi-granularity ODE sampling results to evaluate the SDE sampling direction comprehensively. Different granularities represent diverse sampling intervals during denoising, controlled by the parameter set Λ. To validate the effectiveness of multi-granularity fusion, we perform ablation studies on various Λ set shown in Tab. 2. It can be found that as the number of selected granularities increases, the evaluation of each sampling direction becomes more comprehensive, facilitating a robust assessment through multi-granularity advantage fusion, which significantly improves performance across both in-domain and out-of-domain reward models. Additional ablation studies are presented in **supplementary material**.

Table 3. Comprehensive evaluation of total denoising steps.

Inference Steps	Method	HPS	CLIP	PS	IR	UR
10 Steps	Flux.1-dev	0.289	0.388	0.225	0.939	3.504
	DanceGRPO	0.325	0.390	0.227	1.129	3.576
	MixGRPO	0.358	0.401	0.230	1.431	3.641
	G^2RPO	**0.378**	**0.408**	**0.235**	**1.519**	**3.805**
20 Steps	Flux.1-dev	0.300	0.389	0.226	1.034	3.575
	DanceGRPO	0.329	0.388	0.228	1.136	3.586
	MixGRPO	0.363	0.401	0.230	1.430	3.651
	G^2RPO	**0.376**	**0.407**	**0.235**	**1.511**	**3.806**

4.4. Further exploration of varying inference steps

In the main experiment Tab. 1, we maintained the same inference settings as DanceGRPO and MixGRPO. Notably, our MGAI module, which evaluates denoising samples of different granularities in a mixed manner, provides a more comprehensive assessment. With the assistance of this module, G^2RPO exhibits stronger robustness to varying denoising step configurations. As shown in Tab. 3, all flow-based GRPO methods are jointly trained with HPS-v2.1 and CLIP. When the total inference timesteps are reduced to 20 or even 10 steps, G^2RPO still achieved significant performance improvements across various in-domain and out-of-domain evaluation reward models, thereby demonstrating the value of multi-granularity assessment.

5. Conclusion

This paper addresses the critical limitations of precisely evaluating the quality of denoising directions sampled by Flow-based GRPO for human preference alignment. We introduce Granular-GRPO (G^2RPO), a novel online RL framework that precisely localizes stochasticity to a single step within the denoising process and provides a comprehensive evaluation of SDE denoising directions by integrating the advantages derived from images at different denoising granularities. This innovative design enables the provision of dense, precise reward signals, thereby fundamentally improving optimization accuracy and leading to a more robust and higher-quality alignment. Our extensive experiments consistently demonstrate that G^2RPO achieves superior performance across diverse reward conditions, marking a significant advancement in aligning generative models with human preferences.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant No. 62471287). This project is funded in part by Shanghai Artificial Intelligence Laboratory, Shanghai Innovation Institute.

References

[1] Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*, 2023. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2, 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[5] Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Bytheway: Boost your text-to-video generation model to higher quality in a training-free way. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12999–13008, 2025. 2

[6] Jiazi Bu, Pengyang Ling, Yujie Zhou, Yibin Wang, Yuhang Zang, Tong Wu, Dahua Lin, and Jiaqi Wang. Dicache: Let diffusion model determine its own cache. *arXiv preprint arXiv:2508.17356*, 2025. 2

[7] Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025. 2

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2

[9] Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*, 2023. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3

[11] 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*, 36:79858–79885, 2023. 2

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[14] Joshua Hare. Dealing with sparse rewards in reinforcement learning. *arXiv preprint arXiv:1910.09281*, 2019. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[16] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 3

[17] 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. 6

[18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[19] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 6

[20] 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. *arXiv preprint arXiv:2507.21802*, 2025. 3, 6

[21] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2 (5):7, 2024. 2

[22] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 2

[23] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[24] 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. 2, 3

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[26] 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. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3

[28] 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. 2, 3

[29] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[30] 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. 2, 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 3

[33] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993, 2024. 3

[34] 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. 2, 3

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3

[37] Hao Sun. Supervised fine-tuning as inverse reinforcement learning. *arXiv preprint arXiv:2403.12017*, 2024. 3

[38] 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. 3

[39] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[40] 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*, 2025. 3

[41] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 6

[42] Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *arXiv preprint arXiv:2602.02380*, 2026. 2

[43] 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. 6

[44] 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 3, 6

[45] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 2, 3, 6

[46] Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13315–13325, 2025. 2