From Sparse to Dense: Multi-View GRPO for Flow Models via Augmented Condition Space

Jiazi Bu[1,2,6*], Pengyang Ling[3*], Yujie Zhou[1,6*], Yibin Wang[4,9], Yuhang Zang[6], Tianyi Wei[2], Xiaohang Zhan[7], Jiaqi Wang[9], Tong Wu[8†], Xingang Pan[2†], and Dahua Lin[5,6,10]

[1]Shanghai Jiao Tong University, China [2]S-Lab, Nanyang Technological University, Singapore [3]University of Science and Technology of China, China [4]Fudan University, China [5]The Chinese University of Hong Kong, Hong Kong, China [6]Shanghai AI Laboratory, China [7]Adobe Research, USA [8]Stanford University, USA [9]Shanghai Innovation Institute, China [10]CPII under InnoHK, Hong Kong, China

Fig. 1: Gallery of MV-GRPO. Our MV-GRPO substantially elevates the generation quality of flow models (Flux.1-dev in this figure), particularly in terms of fine-grained details and photorealism. *Prompts are listed in Sec.E of the Supplementary Material.*

Abstract. Group Relative Policy Optimization (GRPO) has emerged as a powerful framework for preference alignment in text-to-image (T2I) flow models. However, we have observed that the standard paradigm

* Equal contribution. †Corresponding author.

that evaluates a group of generated samples against a single condition suffers from insufficient exploration of inter-sample relationships, constraining both alignment efficacy and performance ceilings. To address this sparse single-view evaluation scheme, we propose Multi-View GRPO (**MV-GRPO**), a novel algorithm that enhances relationship exploration by augmenting the condition space to create a dense multi-view reward mapping. Specifically, for a group of samples generated from one prompt, MV-GRPO leverages a flexible Condition Enhancer to generate semantically adjacent yet diverse captions. These captions enable multi-view advantage re-estimation, capturing diverse semantic attributes and providing richer optimization signals. By deriving the probability distribution of the original samples conditioned on these new captions, they can be incorporated into the training process without costly sample regeneration. Extensive experiments demonstrate that MV-GRPO achieves superior alignment performance over state-of-the-art methods.

Keywords: Reinforcement Learning · GRPO · Diffusion/Flow Model

1 Introduction

Over the past few years, diffusion/flow models [13, 21, 24, 33] have emerged as the dominant paradigm in the generative scheme, demonstrating unprecedented capability in synthesizing high-fidelity visual content [9,15,26,29]. While pre-training on massive datasets [5, 23, 30] endows these models with impressive generative versatility, ensuring their outputs align with human preferences and task-specific downstream constraints still poses a critical ongoing challenge [6]. Recent advances in Reinforcement Learning (RL)-based post-training approaches [2,10,28,31] have demonstrated considerable efficacy in bridging this gap. Through optimization anchored in reward models [22,37,39,42] that faithfully reflect human preferences, these methods effectively align model outputs with desired behaviors and task constraints.

Among these advancements, Group Relative Policy Optimization (GRPO) [32] has stood out for its efficiency and stability. Initially grounded in Large Language Models (LLMs), GRPO estimates the advantage of each sample relative to a group average under a given condition (e.g., a textual prompt), thereby eliminating the need for a complex value network and fostering a scalable, flexible framework for preference alignment. A line of research [7,12,20,38,44,47] has adapted GRPO for visual generation by substituting the standard ODE solvers with SDEs to introduce stochasticity during the flow sampling process.

As reward estimation relies on noise-free samples generated via computationally expensive iterative denoising, it is essential to fully exploit the relationships among these samples for preference alignment. However, existing methods typically operate under a "Single-View" paradigm: they evaluate the generated group solely against the single initial condition. This reward evaluation protocol can be reinterpreted as *a sparse, one-to-many mapping from the condition space $\mathcal{C}$ to the data space $\mathcal{X}$*, as shown in Fig. 2 (a). Fundamentally, this paradigm models intra-group relationships by ranking samples based on their alignment with

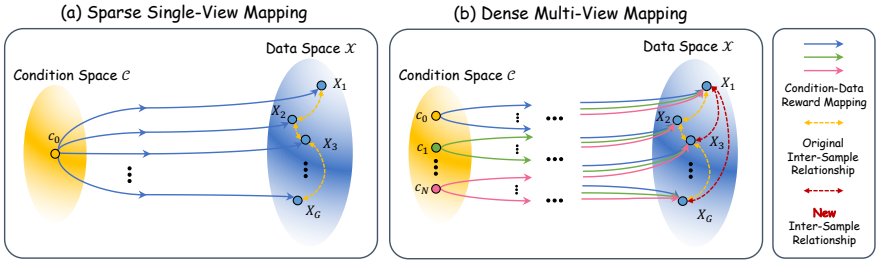


Fig. 2: Reward Evaluation in GRPO Training. (a) Standard flow-based GRPO methods evaluate generated samples under the single original condition, resulting in sparse reward mapping and insufficient inter-sample relationship exploration. (b) Our MV-GRPO leverages an augmented set of conditions to facilitate a dense multi-view mapping, fostering a comprehensive exploration of relationship among samples.

a singular condition, ignoring the multifaceted nature of visual semantics. For instance, as illustrated in Fig. 3, given a SDE sample describing a cat and a dog within a teacup, it may rank poorly under one condition ("A cat and a dog in a teacup.") but highly under another similar condition specifying visual attributes like lighting, motion or composition. Consequently, relying solely on the ranking derived from a single prompt is insufficient to gauge the nuanced relationships among samples, resulting in an inherently sparse reward mapping. In contrast, by incorporating the diverse rankings induced by novel prompts, we can effectively densify the condition-data reward signal. This strategy serves two purposes: (i) enabling a more comprehensive exploration of intra-group relationships from multiple perspectives, and (ii) establishing intrinsic contrasts by identifying ranking shifts across different conditions, thereby facilitating preference-aligned generation under various conditions.

In light of the above analysis, we propose **M**ulti-**V**iew **GRPO** (**MV-GRPO**), a novel reinforcement learning framework that provides a dense supervision paradigm via an Augmented Condition Space. Specifically, MV-GRPO introduces a flexible **Condition Enhancer** module to sample a set of semantically adjacent conditions around the original condition anchor. As depicted in Fig. 2 (b), these augmented conditions, together with the original prompt, form a multi-view condition set that is used to jointly evaluate the relative advantage relationships among the generated samples. This design offers two key benefits: (i) the multi-view evaluation paradigm reinforces the thoroughness of intra-group sample assessment and inherently facilitates the model's capacity to learn ranking variations under diverse perspectives, promoting heightened awareness of conditional perturbations for enhanced preference alignment, and (ii) by augmenting the condition space $\mathcal{C}$ rather than the computationally expensive data space $\mathcal{X}$, we incur only modest overhead by reusing the previously generated noise-free samples. Extensive experiments demonstrate that MV-GRPO significantly outperforms standard single-view baselines, achieving superior visual quality (as shown in Fig. 1) and generalization capabilities. Our contributions are twofold:

1. **Multi-View GRPO**: We identify the sparsity of single-view reward evaluation in flow-based GRPO and propose MV-GRPO, a novel algorithm that leverages a Condition Enhancer to construct an augmented condition set. By re-evaluating the probabilities of original samples under the augmented conditions, it enables multi-view optimization without costly regeneration.
2. **Empirical Validation**: MV-GRPO achieves superior performance over existing baselines, excelling in both in-domain and out-of-domain evaluation.

2 Related Work

2.1 Diffusion and Flow Matching

Diffusion models [8, 13, 33, 34] have achieved exceptional performance in generative modeling by learning to reverse a gradual noising process, enabling high-fidelity visual synthesis across various modalities [3, 4, 11, 46]. The introduction of Latent Diffusion Models (LDMs) [29] further reduces the computational cost by performing the diffusion process in a compressed latent space. Instead of simulating a stochastic diffusion path, flow models [9, 19, 21] directly learn a continuous-time velocity field that moves along straight lines between the noise and data distributions, offering better stability and scalability, and giving rise to numerous state-of-the-art generative models like Flux series [15, 16], Qwen-Image [41], HunyuanVideo series [14, 40] and WAN series [36].

2.2 Alignment for Diffusion and Flow Models

Aligning Diffusion and Flow models with human preferences has evolved from early PPO-style policy gradients [2, 31, 43] and DPO variants [25, 28, 35] toward more efficient online reinforcement learning frameworks like Group Relative Policy Optimization (GRPO) [32]. To enable GRPO to Flow Matching, foundational works such as Flow-GRPO [20] and DanceGRPO [44] reformulate deterministic Ordinary Differential Equation (ODE) sampling into equivalent Stochastic Differential Equation (SDE) trajectories, facilitating the stochastic exploration necessary for policy optimization while preserving marginal probability distributions. Building upon this, several variants have emerged to refine the alignment process: TempFlow-GRPO [12] and Granular-GRPO [49] introduce dense credit assignment for precise T2I alignment. Then, efficiency is further addressed by MixGRPO [17] through a hybrid ODE-SDE sampling mechanism and by Branch-GRPO [18] via structured branching rollouts. DiffusionNFT [48] optimizes the forward process directly via flow matching, defining an implicit policy direction by contrasting positive and negative generations. Despite these advancements, existing frameworks typically follow a sparse, one-to-many reward evaluation paradigm, leading to insufficient and suboptimal exploration. In this work, we enable a dense condition-data reward mapping through efficiently augmenting the condition space, achieving more comprehensive advantage estimation and improved alignment performance.

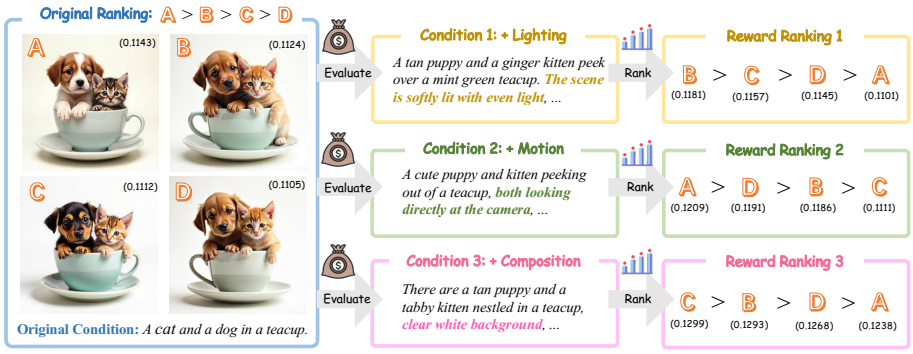


Fig. 3: Reward Ranking Varies with Conditions. Reward rankings of SDE samples across multiple semantically similar yet different conditions exhibit large variations, indicating that relying on a single condition for advantage estimation is inadequate.

3 Method

As illustrated in Fig. 3, given a prompt condition, a set of images generated via SDE-based sampling usually shares the same main subject content while varying in attributes or local details that are not explicitly specified in the prompt. Consequently, evaluating the whole group only under the original prompt provides an incomplete view of their relative merits. Notably, when the prompt is perturbed (Condition 1, 2 and 3 in Fig. 3), the reward ranking of these samples also change accordingly, suggesting that single-view evaluation is insufficient for capturing their nuanced relationships. Motivated by this observation, we extend flow-based GRPO from single-view reward evaluation to dense multi-view supervision. Specifically, given a group of samples generated from one prompt, we use a Condition Enhancer to construct a set of semantically related conditions, re-estimate the advantages of the same samples under these augmented conditions, and optimize the model without sample regeneration. The details of our method are presented in the following subsections.

3.1 Preliminary: Flow-based GRPO

Flow Matching as MDP. Flow-based GRPO [20,44] formulates the generation process as a multi-step Markov Decision Process (MDP). Let $\mathbf{c} \in \mathcal{C}$ be the condition. The policy π_θ, parameterized by θ, induces a reverse-time generation trajectory $\Gamma = (\mathbf{s}_T, \mathbf{a}_T, \ldots, \mathbf{s}_0, \mathbf{a}_0)$. Here, the state $\mathbf{s}_t = (\mathbf{c}, t, \boldsymbol{x}_t)$ encompasses the current noisy latent $\boldsymbol{x}_t$ at timestep t, initializing from $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$ and terminating at the clean sample $\boldsymbol{x}_0$. The action $\mathbf{a}_t$ corresponds to the single-step denoising update derived from the policy $\pi_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \mathbf{c})$.

Sampling with SDE. Standard flow matching models [9,15] typically utilize a deterministic Ordinary Differential Equation (ODE) for sampling:

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

where $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c})$ is the predicted flow velocity. To satisfy the stochastic exploration requirements of GRPO, prior works [20, 44] substitute the ODE with a Stochastic Differential Equation (SDE) that preserves the marginal distribution:

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

where $d\mathbf{w}_t$ represents the Wiener process increments. The term $\sigma_t = \eta\sqrt{\frac{t}{1-t}}$ modulates the magnitude of injected noise, governed by the hyperparameter η. For practical implementation, this is discretized via the Euler-Maruyama scheme:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \left(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c}) + \frac{\sigma_t^2}{2t}(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c})) \right) \Delta t + \sigma_t \sqrt{\Delta t}\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ denotes the Gaussian noise for stochastic exploration.

Training. Given a condition $\mathbf{c}$, a generation rollout produces a set of G outputs $\{\boldsymbol{x}_0^i\}_{i=1}^G$. The relative advantage A_t^i of $\boldsymbol{x}_0^i$ is then derived by comparing its reward value $R(\boldsymbol{x}_0^i, \mathbf{c})$ against the aggregate group statistics as follows:

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

Finally, the policy model is optimized by maximizing the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{C}, \{\boldsymbol{x}^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{c})} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_{\text{clip}}(r_t^i, A_t^i) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \tag{5}$$

where the coefficient β balances the KL regularization term that keeps the learned policy close to the reference policy during training, and:

$$\mathcal{L}_{\text{clip}}(r_t^i, A_t^i) = \min\left(r_t^i(\theta)A_t^i, \text{clip}(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon)A_t^i \right), \tag{6}$$

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

ε in Eq. 6 is the clipping parameter bounding policy updates.

3.2 Condition Enhancer

To facilitate a comprehensive evaluation of visual samples, we consider sampling auxiliary conditions from the local manifold surrounding the anchor condition $\mathbf{c}$ in the condition space $\mathcal{C}$ for a dense multi-view assessment. We formalize the Condition Enhancer operator as $\mathcal{E} : \mathcal{C} \times \mathcal{X} \to 2^{\mathcal{C}}$, which maps an anchor condition $\mathbf{c}$ and a sample group $\mathbf{X}_G = \{\boldsymbol{x}_0^i\}_{i=1}^G$ to an augmented condition set:

$$\mathcal{V}_K = \mathcal{E}(\mathbf{c}, \mathbf{X}_G) = \{\mathbf{c}' \in \mathcal{C} \mid \mathbf{c}' \sim p_{\mathcal{E}}(\cdot|\mathbf{c}, \mathbf{X}_G)\}, \tag{8}$$

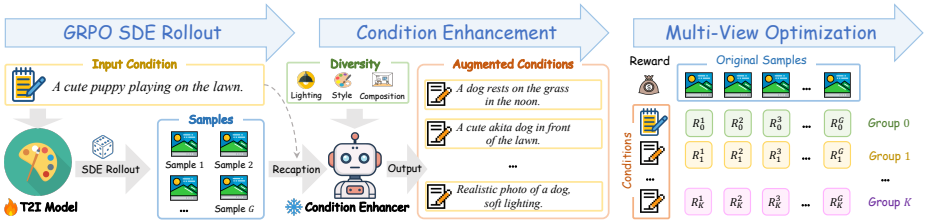


Fig. 4: Overview of MV-GRPO. MV-GRPO leverages a flexible Condition Enhancer module (a pretrained VLM or LLM) to generate diverse augmented conditions for dense multi-view reward signals, facilitating comprehensive advantage estimation.

in which $\mathcal{V}_K$ denotes the resulting augmented condition set containing K additional views, and $p_{\mathcal{E}}$ represents the sampling distribution of $\mathcal{E}$ given $\mathbf{c}$ and $\mathbf{X}_G$. In practice, we provide two implementations of $\mathcal{E}$:

Online VLM Enhancer. To dynamically capture the visual semantics of generated samples, a pretrained Vision-Language Model (VLM) is employed as an online Condition Enhancer $\mathcal{E}_{\text{VLM}}$. During the training loop, $\mathcal{E}_{\text{VLM}}$ projects each sample $\boldsymbol{x}_0^i \in \mathbf{X}_G$ back to the condition space to obtain posterior conditions:

$$\mathcal{V}_K^{\text{post}} = \left\{ \mathbf{c}_i^{\text{post}} \in \mathcal{C} \mid \mathbf{c}_i^{\text{post}} \sim p_{\mathcal{E}_{\text{VLM}}}(\cdot|\mathbf{c}, \boldsymbol{x}_0^i, \mathtt{P}_{\text{VLM}}), i = 1 \dots K \right\}, \tag{9}$$

where the prompt $\mathtt{P}_{\text{VLM}}$ instructs $\mathcal{E}_{\text{VLM}}$ to describe visual contents within $\boldsymbol{x}_0^i$. For each enhancement given by Eq. 9, $\mathtt{P}_{\text{VLM}}$ is randomly sampled from an instruction set $\mathcal{P}_{\text{VLM}}$ covering diverse descriptive perspectives (e.g., `lighting`, `composition`, `style`, etc.). The above design guarantees the diversity of augmented conditions from two aspects: (i) First, each $\mathbf{c}_i^{\text{post}}$ is derived from a unique SDE sample $\boldsymbol{x}_0^i \in \mathbf{X}_G$; (ii) Second, $\mathcal{E}_{\text{VLM}}$ is queried with varied instructions $\mathtt{P}_{\text{VLM}} \in \mathcal{P}_{\text{VLM}}$ focusing on different attributes. In the implementation, we set $K = G$ to fully leverage the generated samples within $\mathbf{X}_G$.

Offline LLM Enhancer. As a complementary strategy based purely on textual semantics, a pretrained Large Language Model (LLM) is utilized as an offline Condition Enhancer $\mathcal{E}_{\text{LLM}}$, which directly samples prior conditions given the anchor condition $\mathbf{c}$:

$$\mathcal{V}_K^{\text{prior}} = \left\{ \mathbf{c}_i^{\text{prior}} \in \mathcal{C} \mid \mathbf{c}_i^{\text{prior}} \sim p_{\mathcal{E}_{\text{LLM}}}(\cdot|\mathbf{c}, \mathtt{Mem}, \mathtt{P}_{\text{LLM}}), i = 1 \dots K \right\}, \tag{10}$$

where the prompt $\mathtt{P}_{\text{LLM}}$ instructs $\mathcal{E}_{\text{LLM}}$ to rewrite the condition. Mirroring the online mode to ensure diversity, (i) `Mem` represents a historical output buffer introduced to prevent duplicate responses, and (ii) $\mathtt{P}_{\text{LLM}}$ is randomly chosen from an editing prompt set $\mathcal{P}_{\text{LLM}}$, which includes three operations: `addition`, `deletion`, and `rewriting`. Crucially, since $\mathcal{E}_{\text{LLM}}$ operates independently of image generation, it can be executed entirely offline before training. *The full details of all VLM/LLM prompts are provided in Sec.B of the Supplementary Material.*

Algorithm 1 Multi-View GRPO Training Process

Require: Prompt dataset $\mathcal{P}$, policy model π_θ, reward model R, total sampling steps T, SDE sampling timestep set M, group size G, total training iterations E
Require: Condition Enhancer $\mathcal{E}$, number of augmented conditions K
for training iteration $e = 1$ **to** E **do**
  Update old policy model: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Sample batch prompts $\mathcal{P}_b \sim \mathcal{P}$
  for anchor condition $\mathbf{c} \in \mathcal{P}_b$ **do**
    # 1. SDE Sampling for G Samples
    Initialize noise: $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$
    for $t = T$ **down to** 1 **do**
      if $t \in M$ **then**
        SDE Sampling: generate $\boldsymbol{x}_{t-1}^i$ for $i = 1 \dots G$
      else
        ODE Sampling: generate $\boldsymbol{x}_{t-1}$
      end if
    end for
    Obtain a group of generated samples $\mathbf{X}_G = \{\boldsymbol{x}_0^i\}_{i=1}^G$
    # 2. Condition Enhancement
    Generate augmented condition set $\mathcal{V}_K = \{\mathbf{c}_k\}_{k=1}^K \sim p_\mathcal{E}(\cdot|\mathbf{c}, \mathbf{X}_G)$
    # 3. Multi-View Advantage Estimation
    Compute rewards $R(\boldsymbol{x}_0^i, \mathbf{c})$ and advantages $A_t^{i,\mathbf{c}}$ for the anchor condition $\mathbf{c}$
    for augmented condition $\mathbf{c}_k \in \mathcal{V}_K$ **do**
      Compute rewards $R(\boldsymbol{x}_0^i, \mathbf{c}_k)$ and advantages $A_t^{i,\mathbf{c}_k}$
    end for
    # 4. MV-GRPO Objective Computation
    Compute the Multi-View GRPO objective $\mathcal{J}_{\text{MV-GRPO}}(\theta)$ using Eq. 11
  end for
  Update policy: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{\text{MV-GRPO}}(\theta)$
end for

3.3 Multi-View GRPO

Building upon the expanded prompts generated through condition enhancement and their associated condition-data mappings, we develop MV-GRPO, a multi-view flow-based GRPO framework that densely couples generated samples with diverse conditions. The overview of MV-GRPO is illustrated in Fig. 4.

Training Objective. The model is fine-tuned on a mixed set of both the original condition and the augmented conditions. The final MV-GRPO objective is constructed by aggregating the policy gradient losses across the anchor view $\mathbf{c}$ and the K augmented conditions in $\mathcal{V}_K$, with the KL term omitted for brevity:

$$\mathcal{J}_{\text{MV-GRPO}}(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{C},\{\boldsymbol{x}_0^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{c}),\{\mathbf{c}_k\}_{k=1}^K\sim p_\mathcal{E}(\cdot|\mathbf{c},\mathbf{X}_G)}$$
$$\Bigg[\underbrace{\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\min\left(r_t^i(\theta)A_t^{i,\mathbf{c}}, \text{clip}(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon)A_t^{i,\mathbf{c}}\right)}_{\text{Objective for the Original Condition}} + \tag{11}$$
$$\underbrace{\sum_{k=1}^{K}\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\min\left(r_t'^i(\theta, \mathbf{c}_k)A_t^{i,\mathbf{c}_k}, \text{clip}(r_t'^i(\theta, \mathbf{c}_k), 1-\varepsilon, 1+\varepsilon)A_t^{i,\mathbf{c}_k}\right)}_{\text{Objective for Augmented Conditions}}\Bigg],$$

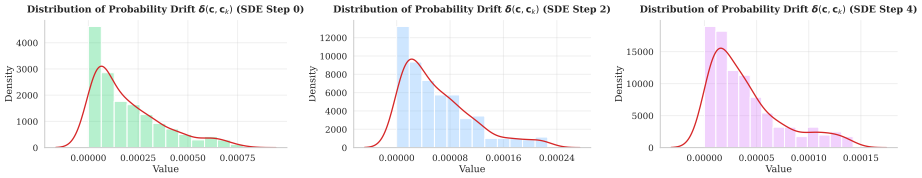


Fig. 5: Distribution of Probability Drift at Different SDE Steps. Most condition pairs exhibit a drift near zero, demonstrating that the SDE transition probability is effectively preserved when substituting the original with augmented conditions.

where $A_t^{i,\mathbf{c}_k}$ is the advantage for the sample $\boldsymbol{x}_0^i$ under an augmented condition $\mathbf{c}_k$ (derived from Eq. 4 by substituting $\mathbf{c}$ with $\mathbf{c}_k$), with $r_t^i(\theta)$ and $r_t^{'i}(\theta, \mathbf{c}_k)$ denoting the importance sampling ratios conditioned on $\mathbf{c}$ and $\mathbf{c}_k$, respectively:

$$r_t^i(\theta) = \frac{p_\theta(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t^i, \mathbf{c})}{p_{\theta_{\text{old}}}(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t^i, \mathbf{c})}, r_t^{'i}(\theta, \mathbf{c}_k) = \frac{p_\theta(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t^i, \mathbf{c}_k)}{p_{\theta_{\text{old}}}(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t^i, \mathbf{c}_k)}. \tag{12}$$

The training pipeline of MV-GRPO is detailed in Algorithm 1.

Theoretical Perspective. To justify optimizing the policy conditioned on an augmented view $\mathbf{c}_k$ using trajectories generated under the anchor $\mathbf{c}$, we examine the transition probability dynamics. Recall from Eq. 3 that the single-step transition from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$ (denoted as step size Δt) follows a Gaussian distribution. The transition mean $\boldsymbol{\mu}_\theta$ and covariance $\boldsymbol{\Sigma}_t$ derived from the SDE solver are given by:

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \mathbf{c}) = \boldsymbol{x}_t + \left(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c}) + \frac{\sigma_t^2}{2t}(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c}))\right)\Delta t, \tag{13}$$

$$\boldsymbol{\Sigma}_t = \sigma_t^2 \Delta t \mathbf{I}. \tag{14}$$

Consequently, the policy $\pi_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \mathbf{c})$ can be modeled as $\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \mathbf{c}), \boldsymbol{\Sigma}_t)$, where the probability density is formulated as:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \mathbf{c}) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_t|}} \exp\left(-\frac{1}{2}\left\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \mathbf{c})\right\|^2_{\boldsymbol{\Sigma}_t^{-1}}\right). \tag{15}$$

When evaluating this transition under a new augmented condition $\mathbf{c}_k \in \mathcal{V}_K$, the sampled point $\boldsymbol{x}_{t-1}$ (which was generated via $\mathbf{c}$) is fixed. The probability density of observing this specific transition under the new view $\mathbf{c}_k$ is given by:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \mathbf{c}_k) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_t|}} \exp\left(-\frac{1}{2}\left\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \mathbf{c}_k)\right\|^2_{\boldsymbol{\Sigma}_t^{-1}}\right). \tag{16}$$

The probability drift induced by the condition perturbation is defined as the absolute difference in log-probability densities:

$$\boldsymbol{\delta}(\mathbf{c}, \mathbf{c}_k) = \left|\log p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \mathbf{c}) - \log p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \mathbf{c}_k)\right|. \tag{17}$$

We sampled 500 pairs of $(\mathbf{c}, \mathbf{c}_k)$ through the VLM enhancer and calculate their corresponding probability drift, with the resulting distribution plotted in Fig. 5. Specifically, it can be observed that the drift is minimal for the vast majority of cases across different SDE steps, which is ensured by our Condition Enhancer through sampling semantically adjacent augmented conditions. Given the negligible difference in transition probabilities, $\mathcal{V}_K = \{\mathbf{c}_k\}_{k=1}^{K}$ offers a meaningful signal for dense supervision and can be seamlessly incorporated into GRPO training. More discussion is provided in Sec.G of the Supplementary Material.

4 Experiments

4.1 Implementation Details

Datasets and Models. Following previous works [17, 44, 49], the HPD [42] dataset is employed as the prompt dataset. It comprises over 100K prompts for training and a separate set of 400 prompts for evaluation. We adopt Flux.1-dev [15] as the training backbone, an advanced open-source T2I flow model recognized for its superior visual quality. For the Condition Enhancer, we utilize two leading models from the Qwen series: Qwen3-VL-8B [1] is deployed as the online VLM enhancer, while Qwen3-8B [45] serves as the offline LLM enhancer. Further details are provided in Sec.B of the Supplementary Material.
Baselines. The compared methods encompass the vanilla Flux model [15], Flow-GRPO [20], DanceGRPO [44], TempFlow-GRPO [12] and DiffusionNFT [48].
Evaluation Metrics. To comprehensively assess the effectiveness of MV-GRPO, a diverse set of metrics is employed for evaluation: (i) *Leading VLM-based Reward Models*: HPS-v3 [22] and UnifiedReward-v1/v2 (UR-v1/v2) [39]; (ii) *CLIP/BLIP-based Reward Models*: HPS-v2 [42], CLIP [27] and ImageReward (IR) [43].
Sampling Details. Each SDE rollout is conducted with a group size of $G = 12$. The total number of sampling steps is set as $T = 16$ for efficiency. The noise level throughout the sampling process is governed by the hyperparameter η, which is fixed at 0.7 in the expression $\sigma_t = \eta\sqrt{\frac{t}{1-t}}$. To ensure a fair comparison, all baseline methods adopt the identical configuration described above.
Training Details. We build MV-GRPO upon Flow-GRPO-Fast [20], an efficient variant of Flow-GRPO [20]. The training steps are configured as $\{0, 2, 4, 6\}$. Following prior studies [44, 49], we train MV-GRPO under two experimental settings: (i) *Single-Reward*, where the model is fine-tuned using a single state-of-the-art reward model, specifically either HPS-v3 or UnifiedReward-v2; (ii) *Multi-Reward*, in which HPS-v3 and CLIP are jointly utilized as reward signals to improve training robustness and prevent potential reward-hacking. Additional experimental results on other GRPO frameworks and T2I models can be found in Sec.H of the Supplementary Material.
Optimization Details. If not specified, all experiments in this section are conducted on 16×NVIDIA H200 GPUs with the batch size setting to 1. We employ the AdamW optimizer, specifying a learning rate of 2×10^{-6} and a weight decay of 1×10^{-4}. `bfloat16` (bf16) mixed-precision training is adopted for efficiency.

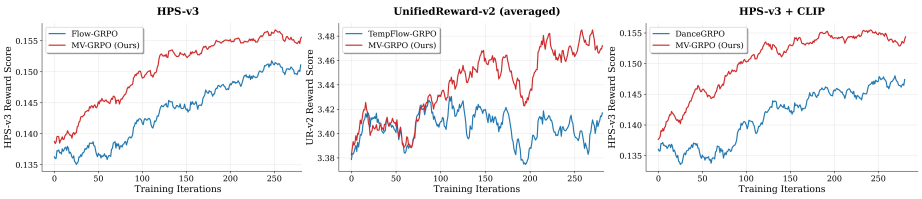


Fig. 6: Reward Curves during Training. Our MV-GRPO outperforms baselines in both convergence speed and performance ceiling under various training settings. "average" in the middle panel refers to the mean of UR-v2's three dimensions.

Table 1: Quantitative comparison of different methods. The best results are in **bold**, while the second-best result is underlined. UR-v2-A, UR-v2-C, and UR-v2-S denote the Alignment, Coherence, and Style dimensions of UnifiedReward-v2, respectively.

Reward Model	Method	HPS-v3	UR-v2-A	UR-v2-C	UR-v2-S	HPS-v2	CLIP	IR	UR-v1
/	Flux.1-dev [15]	0.133	3.250	3.634	3.223	0.305	0.389	1.042	3.604
HPS-v3	Flow-GRPO [20]	0.147	3.235	3.670	3.236	0.326	0.380	1.149	3.548
	DanceGRPO [44]	0.145	3.238	3.663	3.224	0.322	**0.381**	1.165	**3.585**
	TempFlow-GRPO [12]	0.150	3.240	3.683	3.299	0.332	0.376	1.184	3.559
	DiffusionNFT [48]	0.149	**3.244**	3.676	3.259	0.330	0.364	1.161	3.551
	MV-GRPO (LLM)	0.153	3.232	3.692	3.313	0.336	0.373	1.176	3.557
	MV-GRPO (VLM)	**0.155**	3.229	**3.701**	**3.320**	**0.340**	0.370	**1.193**	3.569
UnifiedReward-v2	Flow-GRPO [20]	0.129	3.176	3.668	3.268	0.297	0.373	1.003	3.429
	DanceGRPO [44]	0.132	3.168	3.671	3.280	0.303	0.370	1.006	3.412
	TempFlow-GRPO [12]	0.140	3.243	3.687	3.331	0.318	**0.374**	1.140	3.433
	DiffusionNFT [48]	0.139	3.225	3.680	3.338	0.294	0.360	1.076	3.306
	MV-GRPO (LLM)	0.142	3.239	**3.771**	3.429	0.330	0.362	1.118	3.476
	MV-GRPO (VLM)	**0.143**	**3.245**	3.734	**3.454**	**0.332**	0.367	**1.187**	**3.507**
HPS-v3 + CLIP	Flow-GRPO [20]	0.141	3.258	3.637	3.233	0.322	0.391	1.165	3.627
	DanceGRPO [44]	0.143	3.250	3.655	3.221	0.320	0.386	1.133	3.574
	TempFlow-GRPO [12]	0.147	3.274	3.583	3.213	0.336	**0.397**	1.227	3.656
	DiffusionNFT [48]	0.147	3.255	3.632	3.264	0.337	0.389	1.231	3.630
	MV-GRPO (LLM)	0.149	**3.276**	3.695	3.347	0.337	0.389	1.243	3.650
	MV-GRPO (VLM)	**0.152**	3.268	**3.720**	**3.396**	**0.342**	0.393	**1.268**	**3.671**

4.2 Main Results

Quantitative Evaluation. As presented in Tab. 1, MV-GRPO demonstrates consistent superiority under both single reward (HPS-v3 or UnifiedReward-v2) and multi-reward (HPS-v3 + CLIP) settings. Specifically, with online VLM condition enhancer, MV-GRPO achieves the best performance across most metrics, particularly excelling in HPS metrics, ImageReward, coherence (UR-v2-C), and style (UR-v2-S), while offline LLM enhancer yields the second-best results. This can be attributed to the VLM enhancer's ability to generate tailored sample-specific posterior captions, which more precisely describe the generated images and offer more discriminative reward signals than the LLM enhancer's prior conditions. Furthermore, combining HPS-v3 and CLIP yields notable improvements for both metrics, proving that integrating complementary signals (HPS-v3 for semantic quality, CLIP for text alignment) boosts overall generation. These results validate our dense multi-view mapping paradigm enables more compre-

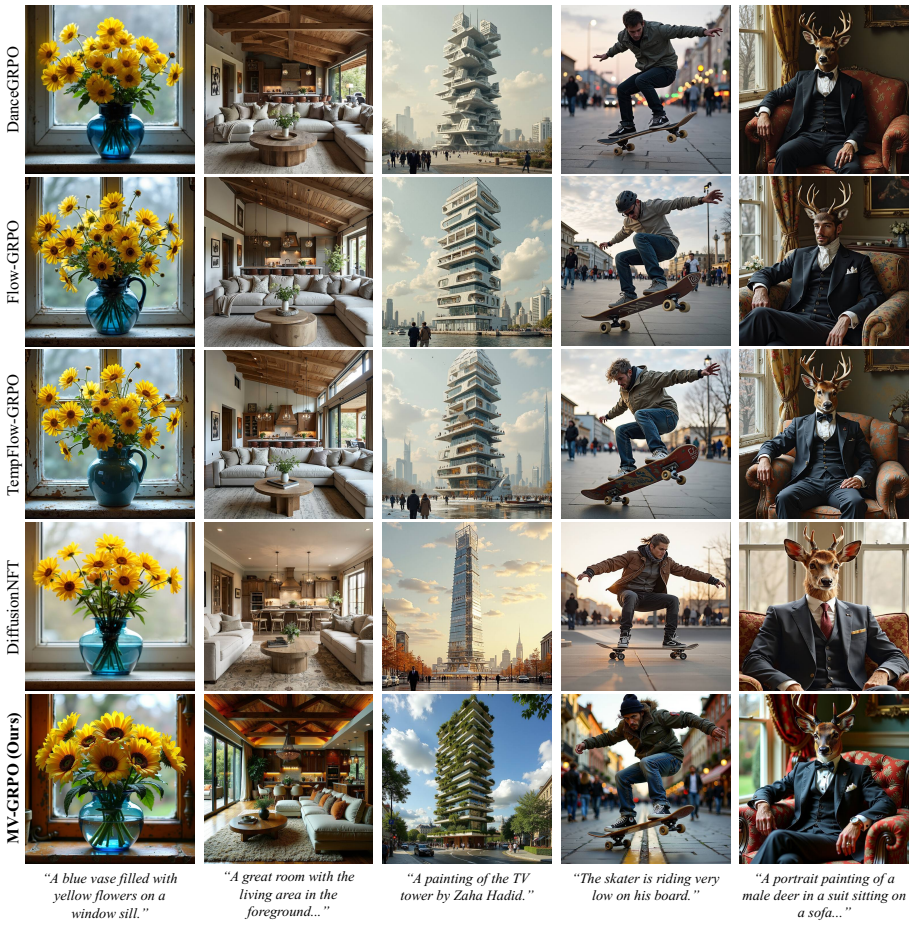


Fig. 7: Qualitative Comparisons with Baselines on HPS-v3.

hensive optimization and achieves superior performance. The reward curves for the VLM enhancer during training are illustrated in Fig. 6. A detailed analysis of the training overhead is provided in Sec.F of the Supplementary Material.

Qualitative Comparison. As depicted in Fig. 7 and Fig. 8, MV-GRPO consistently outperforms its competitors in semantic alignment, visual fidelity, and structural coherence. In the *"room"* and *"tower"* cases (Fig. 7), it renders fine indoor and architectural details with superior clarity. For the *"skater"* case, MV-GRPO enhances the scene's tension by vividly synthesizing facial expressions and clothing wrinkles. Similarly, in the *"daffodil"* and *"cave"* examples (Fig. 8), MV-GRPO enriches the compositions with intricate background elements such as furniture, moons, starry skies, and floral details, significantly elevating the cinematic atmosphere and aesthetic appeal of the generated images. Finally, in the *"ski"* case, MV-GRPO not only generates detailed figures but also opti-

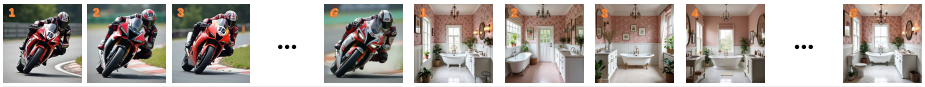
Fig. 8: Qualitative Comparisons with Baselines on UnifiedReward-v2.

mizes the lighting and composition to create a more immersive and expansive snow-covered environment. More qualitative results are presented in Sec.D of the Supplementary Material.

4.3 Ablation Study

We conducted ablation studies on the online VLM enhancer setting trained with reward model HPS-v3. The results are presented in Tab. 2.

Effects of Condition Number. The number of augmented conditions K determines the density of the condition-data reward mapping. Empirically, a larger K facilitates a more thorough exploration of intra-group relationships, leading to better optimization and superior performance. As can be observed in Tab. 2



Original Condition: *Motorcycle racer leaning low coming out of a turn.*

Augmented Condition 1: *Motorcycle racer in red and black gear leans into a turn on a racetrack, surrounded by trees and a guardrail.*

Augmented Condition 2: *Racer in red and black suit leans into a turn on a red sportbike, navigating a racetrack with precision, under overcast lighting.*

Augmented Condition 3: *A red racing motorcycle, ridden by a professional racer in full gear, leaning into a turn on a racetrack.*

Original Condition: *A pink bathroom with a toilet, shower, and other bathroom accessories.*

Augmented Condition 1: *Cozy, vintage-style bathroom with a clawfoot bathtub, white vanity, and potted plants, soft indoor light, creating a charming, inviting atmosphere.*

Augmented Condition 2: *Elegant vintage bathroom with white clawfoot tub, double sinks, ornate mirrors, and terracotta tiles.*

Augmented Condition 3: *Pink floral wallpaper bathroom with white wainscoting, a freestanding bathtub, and a vanity with two sinks. The room features a chandelier, framed artwork, and potted plants.*

Fig. 9: Augmented Conditions. MV-GRPO generates diverse augmented conditions by leveraging variations among SDE samples and multi-view descriptive prompts.

Table 2: Ablation experiments on MV-GRPO components and hyperparameters.

Components	Values & Choices	HPS-v3	UR-v2-A	UR-v2-C	UR-v2-S	HPS-v2	CLIP	IR	UR-v1
(a) Condition Number K	w/o $\mathcal{V}_K$	0.149	**3.238**	3.668	3.269	0.332	**0.378**	1.152	3.553
	$K = 3$	0.152	3.235	3.684	3.293	0.335	0.377	1.170	3.567
	$K = 6$	0.154	3.233	3.696	**3.326**	0.338	0.372	1.178	**3.573**
	K = 12	**0.155**	3.229	**3.701**	3.320	**0.340**	0.370	**1.193**	3.569
(b) Condition Diversity	w/o SDE query	0.153	3.233	3.692	3.304	0.336	0.374	1.183	3.560
	w/o $\mathcal{P}_{\text{VLM}}$	0.152	**3.237**	3.687	3.295	0.334	**0.376**	1.174	3.564
	MV-GRPO	**0.155**	3.229	**3.701**	**3.320**	**0.340**	0.370	**1.193**	**3.569**
(c) Enhancer Scale	Qwen3-VL-2B	0.153	**3.238**	3.694	3.317	**0.344**	**0.378**	1.175	**3.575**
	Qwen3-VL-8B	**0.155**	3.229	**3.701**	**3.320**	0.340	0.370	**1.193**	3.569

(a), performance improves with K but tends to saturate at higher values. Notably, even with $K = G/2 = 6$, the model achieves competitive performance with $K = G = 12$. However, given that the overhead of condition augmentation is small, $K = G = 12$ is chosen to maximize the density of the reward mapping.

Effects of Condition Diversity. The diversity of augmented conditions stems from two aspects: (i) each augmented condition is derived from a distinct SDE sample, and (ii) a diverse set of VLM prompts $\mathcal{P}_{\text{VLM}}$ covering various descriptive perspectives is employed to query the VLM. We ablate (i) by generating all conditions from the same ODE sample, and (ii) by removing the multi-perspective prompt set $\mathcal{P}_{\text{VLM}}$. As shown in Tab. 2 (b), removing either component leads to a notable decline in performance. This confirms that both sample-level stochasticity and prompt-level semantic variety are crucial for constructing a robust and diverse augmented condition space. Examples of the augmented conditions used during training are illustrated in Fig. 9.

Effects of Enhancer Scale. We investigate the impact of the Condition Enhancer's parameter scale by comparing the originally adopted Qwen3-VL-8B with its lightweight variant, Qwen3-VL-2B. As depicted in Tab. 2 (c), Qwen3-VL-8B yields better overall performance on primary metrics such as HPS-v3, IR, and UR-v2-C/S. However, it is noteworthy that the 2B model delivers remarkably competitive results, even marginally surpassing its 8B variant on metrics like UR-v1 and HPS-v2. This indicates that the core advantage of MV-GRPO

stems fundamentally from the dense multi-view evaluation mechanism itself, rather than merely relying on the capacity of the Condition Enhancer.

5 Limitation and Discussion

Method Applicability. Despite the advancements of MV-GRPO in preference alignment for flow models, it faces certain constraints. First, its effectiveness may be limited in tasks with rigid or predefined conditioning signals (e.g., class-conditional generation on a specific dataset), where meaningful condition enhancements are difficult to formulate. Furthermore, the quality of augmented conditions is bounded by the visual understanding and reasoning capabilities of current VLMs or LLMs. However, we anticipate that this limitation will be naturally mitigated as the performance of these models continues to advance.
Evaluation Scope. Although we evaluate MV-GRPO against various baselines under multiple reward settings, our main experiments are centered on Flux.1-dev and Qwen3 models, with limited exploration of the broader model and hyperparameter space. Nevertheless, since Flux and the Qwen-series models are representative strong open-source models widely adopted in the community, we believe the current results provide meaningful evidence for the effectiveness of MV-GRPO in practical flow model alignment settings.

6 Conclusion

In this work, we identify that standard flow-based GRPO relies on a sparse, single-view reward evaluation scheme that causes insufficient exploration of intra-group relationships and suboptimal performance. To this end, we introduce MV-GRPO, a reinforcement learning algorithm that shifts alignment from single-view to dense, multi-view supervision. MV-GRPO leverages a flexible Condition Enhancer module to augment the condition space with semantically related conditions, enabling dense multi-view reward mapping and more comprehensive advantage estimation without sample regeneration. Across multiple reward settings and evaluation metrics, our experiments show that MV-GRPO consistently improves preference alignment over strong baselines. We hope this work can inspire future research on extending multi-view alignment to reinforcement learning for other generative modalities, such as text-to-video and 3D generation.

7 Acknowledgement

This project is funded in part by Shanghai Artificial Intelligence Laboratory, Shanghai Innovation Institute, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. Dahua Lin is a PI of CPII under the InnoHK. This project is also supported by NTU SUG-NAP.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
5. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13320–13331 (2024)
6. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400 (2023)
7. Deng, H., Yan, K., Mao, C., Wang, X., Liu, Y., Gao, C., Sang, N.: Densegrpo: From sparse to dense reward for flow matching model alignment. arXiv preprint arXiv:2601.20218 (2026)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
10. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)
11. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
12. He, X., Fu, S., Zhao, Y., Li, W., Yang, J., Yin, D., Rao, F., Zhang, B.: Tempflow-grpo: When timing matters for grpo in flow models. arXiv preprint arXiv:2508.04324 (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
15. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024), accessed: 2026-06-25
16. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025), accessed: 2026-06-25

17. Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., Zhong, Z.: Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802 (2025)
18. Li, Y., Wang, Y., Zhu, Y., Zhao, Z., Lu, M., She, Q., Zhang, S.: Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. arXiv preprint arXiv:2509.06040 (2025)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
21. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
22. Ma, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15086–15095 (2025)
23. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
24. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
25. Peng, L., Wu, B., Cheng, H., Zhao, Y., He, X.: Sudo: Enhancing text-to-image diffusion models with self-supervised direct preference optimization. arXiv preprint arXiv:2504.14534 (2025)
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
28. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
31. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
32. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
34. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

35. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wang, Y., Li, Z., Zang, Y., Wang, C., Lu, Q., Jin, C., Wang, J.: Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. arXiv preprint arXiv:2505.03318 (2025)
38. Wang, Y., Li, Z., Zang, Y., Zhou, Y., Bu, J., Wang, C., Lu, Q., Jin, C., Wang, J.: Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv preprint arXiv:2508.20751 (2025)
39. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
40. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025)
41. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
42. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
43. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
44. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
45. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
46. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
47. Yue, Z., Ni, Z., Ye, F., Zhang, J., Shen, S., Mi, Z.: Know your step: Faster and better alignment for flow matching models via step-aware advantages. arXiv preprint arXiv:2602.01591 (2026)
48. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117 (2025)
49. Zhou, Y., Ling, P., Bu, J., Wang, Y., Zang, Y., Wang, J., Niu, L., Zhai, G.: G2rpo: Granular grpo for precise reward in flow models. arXiv preprint arXiv:2510.01982 **3** (2025)