DRM: Diffusion-based Reward Model With Step-wise Guidance

Jaxon Zhang[1*], Binxin Yang[2], Hubery Yin[2], Chen Li[2], Jing Lyu[2]
[1]Peking University [2]WeChat Vision, Tencent Inc.

Abstract

Current mainstream methods of aligning diffusion models with human preferences typically employ VLM-based reward models. However, these reward models, pre-trained for semantic alignment, struggle to capture the essential perceptual qualities—such as aesthetics, composition, and visual harmony. In this work, we argue that a model capable of high-fidelity generation must possess a profound understanding of these visual attributes. Based on this insight, we introduce the Diffusion-based Reward Model (DRM), a novel paradigm that use the pre-trained diffusion model as a powerful evaluative backbone. A key advantage of the DRM is its unique ability to assess not only the final image but also the noisy intermediate latents at any stage of the generative process. We leverage this step-wise evaluative capacity in two ways. First, we propose Step-wise GRPO, a reinforcement learning algorithm that provides dense, per-step rewards to resolve the imprecise credit assignment problem in GRPO algorithm, leading to more stable and effective alignment. Second, we introduce Step-wise Sampling, a novel inference strategy that employs the DRM as a dynamic guide to evaluate multiple generation paths at each step, steering the process towards higher-quality outcomes. Extensive experiments confirm that our approach significantly enhances the final quality of generated images. Code: https://github.com/jjaxonx/DRM.

1. Introduction

Diffusion models [7, 11, 29, 32, 33] have demonstrated remarkable generative capabilities. However, their outputs often misalign with human preferences and intent, spurring the wave of research into human preference alignment. A direct approach to alignment involves fine-tuning the model on large-scale human feedback [25, 35]. While effective, this process is prohibitively expensive and labor-intensive for diffusion models. Consequently, an alternative paradigm has gained prominence: learning a reward model (RM) from limited preference data [15, 24, 38, 43, 46]. This learned reward function can then be used to generate synthetic preference data, significantly reducing the reliance on manual annotation for model alignment.

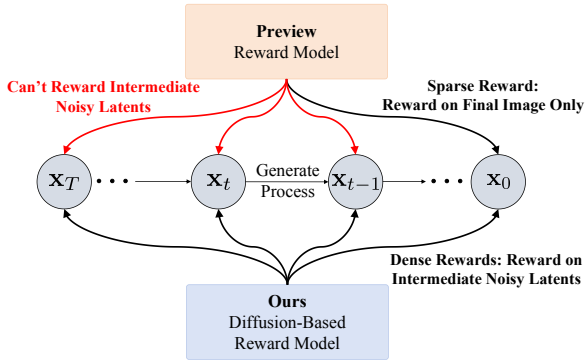

Figure 1. **Comparison between preview reward models and DRM.** Existing reward models treat the generation process as a black box, providing only a single, terminal reward based on the final output. Our DRM offers fine-grained reward for any noisy latent along the entire denoising trajectory.

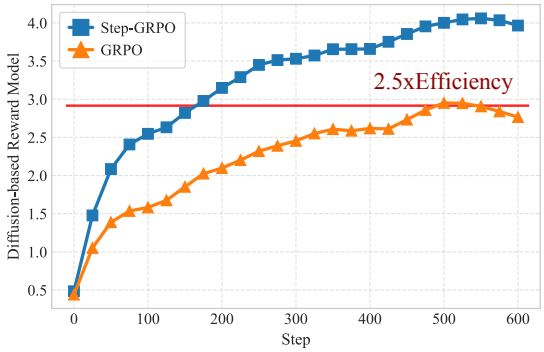

Figure 2. **Reward curves for various RL algorithms optimized using our DRM.** Our Step-GRPO, which leverages dense, per-step rewards, not only reaches a higher final reward but also converges **2.5x faster on step** than the standard GRPO baseline.

Early RMs [19, 42, 47] were typically fine-tuned from CLIP [30] backbones. With the advent of Vision-Language Models (VLMs) [2, 3, 21, 37], their superior visual understanding capabilities made them a more powerful choice for RM backbones, leading to their widespread adoption in visual quality assessment [6, 8, 17, 24, 38, 39, 43, 49].

*Work done during an internship at WeChat Vision, Tencent Inc.

However, these VLMs rely on a CLIP-style vision encoder, which is pre-trained to align images with text based on semantic similarity. This objective inherently prioritizes what an image contains over how it is presented. Consequently, the resulting feature representations are rich in semantics but impoverished in terms of crucial aesthetic and compositional attributes that are pivotal to human preference.

This limitation motivates the search for an alternative vision backbone, one inherently sensitive to the perceptual qualities that CLIP-style encoders neglect. We argue that pre-trained diffusion models are precisely such a backbone. This claim is built on an intuitive yet powerful insight: the ability to generate high-fidelity images necessitates a deep, implicit understanding of visual aesthetics, composition, and fine-grained details. Motivated by this insight, we pioneer the use of diffusion models as the backbone for reward modeling, systematically unlocking their powerful evaluative capabilities. We introduce our approach as the **Diffusion-based Reward Model (DRM)**.

The benefit of using a diffusion backbone is clear: a richer understanding of perceptual qualities like aesthetics and composition. Beyond this, the DRM possesses a more profound advantage: as shown in Figure 1, it does not merely judge the final image; it comprehends the entire generative trajectory, allowing it to assess noisy intermediate states at any given timestep. This unique, step-wise evaluative capacity provides a mechanism to address two key challenges in diffusion models.

(1) On the optimization front, prevailing reinforcement learning alignment methods, such as GRPO [23, 44], suffer from an imprecise credit assignment problem. They treat the multi-step generation process as a "black box," uniformly distributing the reward from the final image across all intermediate timesteps. This coarse approach fails to distinguish between beneficial and detrimental actions during generation. To resolve this, our Step-wise GRPO (Step-GRPO) algorithm leverages the DRM to provide immediate and precise rewards at each step. As visualized in Figure 2, this dense feedback signal enables far more effective and stable policy optimization. **(2)** For inference, we break the rigidity of deterministic samplers. Where conventional methods are locked into a single, uncorrectable path, our Step-wise Sampling strategy employs the DRM as a dynamic guide. At each step, it evaluates multiple potential futures and greedily chooses the one that best preserves quality, preventing the cascading failures common in fixed trajectories. In summary, our contributions are as follows:

❑ (1) We introduce the DRM, a novel paradigm for reward modeling. By using a pre-trained diffusion model as its backbone, the DRM inherits a rich understanding of perceptual qualities like aesthetics and composition, and crucially, it possesses the unique capability to assess noisy intermediate latents at any stage of the generative process.

❑ (2) We propose Step-GRPO, a reinforcement learning algorithm that resolves the credit assignment problem in diffusion model alignment. By using the DRM to provide dense, per-step rewards, Step-GRPO achieves significantly more stable and efficient policy optimization compared to methods that rely on a single, terminal reward.

❑ (3) We present Step-wise Sampling, a novel inference strategy that employs the DRM as a dynamic guide. This method evaluates multiple potential generation paths at each step, steering the process towards higher quality outcomes.

2. Related Work

2.1. Reward Model

Reward models are crucial for aligning diffusion generative models [1, 7, 11, 28, 29, 32, 33] with human preferences. Initially, methods relied on automated metrics like FID [10] and CLIP [30] to evaluate image quality and text-image consistency [12, 13, 26]. However, these metrics fall short of capturing human preferences due to training objectives and data. To bridge this gap, recent research focuses on fine-tuning CLIP models directly on human preference datasets, enabling them to better predict human judgments [19, 42, 47]. With the rise of powerful Vision Language Models (VLMs) [2, 3, 21, 37], they have become a natural choice for reward model backbones, leading to their widespread adoption in visual quality assessment [8, 17, 24, 38, 39, 43, 49]. A key limitation, however, is that these VLMs use a CLIP-style vision encoder that compresses an image into a semantic-heavy representation. This information bottleneck makes the VLM less sensitive to the image's structural integrity and other details. While LPO [48] have explored diffusion-based reward models, these efforts lack a systematic investigation. Motivated by the premise that "generation requires understanding," we systematically explore the diffusion model as a reward backbone, aiming to unlock its potential for more perceptive and accurate reward signals.

2.2. Alignment for Diffusion Models

Aligning diffusion models with human preferences is a significant area of investigation. Recent efforts have largely followed two paths. One line of work adapts direct preference optimization (DPO) [31] for diffusion models, as seen in D3PO [45] and Diffusion-DPO [35]. The other integrates online reinforcement learning, with Flow-GRPO [23] and DanceGRPO [44] being the first to apply it to flow-matching models, inspiring a surge of subsequent works [16, 18, 22, 36]. A central challenge in these approaches is the problem of credit assignment: the reward for the final image is uniformly applied to all steps in the generation process. TempFlow-GRPO [9] attempts to solve this

with a precise score allocation mechanism, but this introduces substantial sampling overhead during training. Our DRM is designed to address this challenge directly. By possessing the inherent capability to evaluate intermediate noisy latents, it offers a more direct and efficient approach to the credit assignment problem.

3. Method

3.1. Preliminary

Flow Matching. Let $x_0 \sim \mathcal{X}_0$ be a data sample from the real world image data distribution and $x_1 \sim \mathcal{X}_1$ a noise sample. Flow Matching[20] defines intermediate samples as

$$x_t = (1-t)x_0 + tx_1, \quad t \in [0,1], \tag{1}$$

and trains a velocity field $v_\theta(x_t, t)$ via the objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|v - v_\theta(x_t,t)\|_2^2\right], \quad v = x_1 - x_0. \tag{2}$$

At inference, the iterative denoising process can be naturally formalized as a Markov Decision Process [4]. At each step t, the state is $s_t = (c, t, x_t)$, where c denotes the prompt, and the action a_t corresponds to producing the denoised sample $x_{t-1} \sim \pi_\theta(x_{t-1}|x_t, c)$.

Flow-GRPO. RL aims to learn a policy that maximizes the expected cumulative reward. Given a prompt $\mathbf{c}$, the flow model p_θ samples a group of G individual images $\{\mathbf{x}_0^i\}_{i=1}^G$ and the corresponding reverse-time trajectories $\{(\mathbf{x}_T^i, \mathbf{x}_{T-1}^i, ..., \mathbf{x}_0^i)\}_{i=1}^G$. Then, the advantage of the i-th image is calculated by normalizing the group-level rewards as follows:

$$\hat{A}_t^i = \frac{R(\mathbf{x}_0^i, \mathbf{c}) - \text{mean}(\{R(\mathbf{x}_0^i, \mathbf{c})\}_{i=1}^G)}{\text{std}(\{R(\mathbf{x}_0^i, \mathbf{c})\}_{i=1}^G)}. \tag{3}$$

GRPO optimizes the policy model by maximizing the following objective:

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{C},\{\mathbf{x}^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{c})} f(r, \hat{A}, \theta, \epsilon, \beta), \tag{4}$$

where

$$f(r, \hat{A}, \theta, \epsilon, \beta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\min\big(r_t^i(\theta)\hat{A}_t^i, \text{clip}(r_t^i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t^i\big) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big), \tag{5}$$

with $r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i|x_t^i, c)}{p_{\theta_{\text{old}}}(x_{t-1}^i|x_t^i, c)}$. To satisfy GRPO's stochastic exploration requirements, [23] convert the deterministic ODE to an equivalent SDE:

$$x_{t+\Delta t} = x_t + \Big(v_\theta(x_t,t) + \frac{\sigma_t^2}{2t}\big(x_t + (1-t)v_\theta(x_t,t)\big)\Big)\Delta t + \sigma_t\sqrt{\Delta t}\,\epsilon, \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ injects stochasticity and $\sigma_t = a\sqrt{\frac{t}{1-t}}$.

3.2. Diffusion-based Reward Model

Architecture of Diffusion-based Reward Model (DRM). As illustrated in Figure 3, our DRM predicts human preferences by leveraging the intermediate features from the Diffusion Transformer (DiT) of a pre-trained diffusion model. Specifically, to ensure a fair comparison with VLM-based reward models in terms of parameter count, we adapt a pre-trained DiT by truncating its final transformer layers. For instance, we initialize our backbone with the pre-trained DiT from SD3.5-Medium (2.5B parameters). To align its scale with models like HPSv3-2B, we remove the last three transformer layers. Given a noisy latent representation x_t at a specific timestep t, it is fed into our modified DiT backbone. This process yields a sequence of visual features $f_v \in \mathbb{R}^{L\times d}$, where L is the sequence length and d is the feature dimension. The visual features f_v are then passed to a prediction head. First, a linear layer projects them to a lower-dimensional space, resulting in $f_p \in \mathbb{R}^{L\times d_p}$. Subsequently, f_p is reshaped into a spatial feature map $f_p \in \mathbb{R}^{h\times w\times \frac{d_p}{4}}$. Finally, this feature map is processed by a small convolutional network, followed by a pooling layer and a liner projection, to produce the final preference score s. The overall process of DRM can be formulated as follows:

$$f_p = \text{MLP}(f_v),\ f_v \in \mathbb{R}^{L\times d}, f_p \in \mathbb{R}^{L\times d_p} \tag{7}$$

$$s = \text{MLP}(\text{Pooling}(\text{Conv}(\text{ReShape}(f_p)))) \tag{8}$$

Training Loss. Our model is trained on a dataset composed of triplets (I^{win}, I^{lose}, p), where (I^{win}, I^{lose}) represents a pair of images with human preference labels (winner and loser), and p is their corresponding text prompt. The training process for a given pair is as follows, also illustrated in Figure 3. First, we encode the images into latent representations, x_0^{win} and x_0^{lose}, using a VAE encoder. Subsequently, for a randomly sampled timestep t, we simulate the forward diffusion process by adding Gaussian noise $\epsilon_t \in \mathcal{N}(0,1)$ to generate the noisy latents x_t^{win} and x_t^{lose}. These noisy latents are then fed into our DRM, conditioned on the timestep t, to obtain their respective preference scores:

$$s^{win} = \text{DRM}(x_t^{win}, t),\ s^{lose} = \text{DRM}(x_t^{lose}, t) \tag{9}$$

Following the Bradley-Terry (BT) model [5], we define the training loss as the negative log-likelihood of the probability that the winning image is preferred over the losing one:

$$\mathcal{L}_{DRM} = -\log(\sigma(s^{win} - s^{lose})), \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

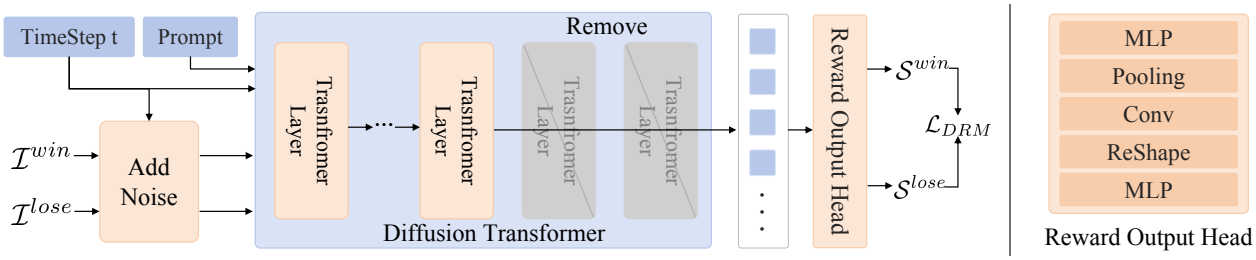

Figure 3. **Overview of the Diffusion-based Reward Model (DRM).** (Left) The training pipeline. During training, the DRM takes a pair of preferred and dispreferred images, both corrupted with noise at a specific timestep t, and predicts their respective reward scores. The model is then optimized via DR loss. (Right) The detailed architecture of our Reward Output Head.

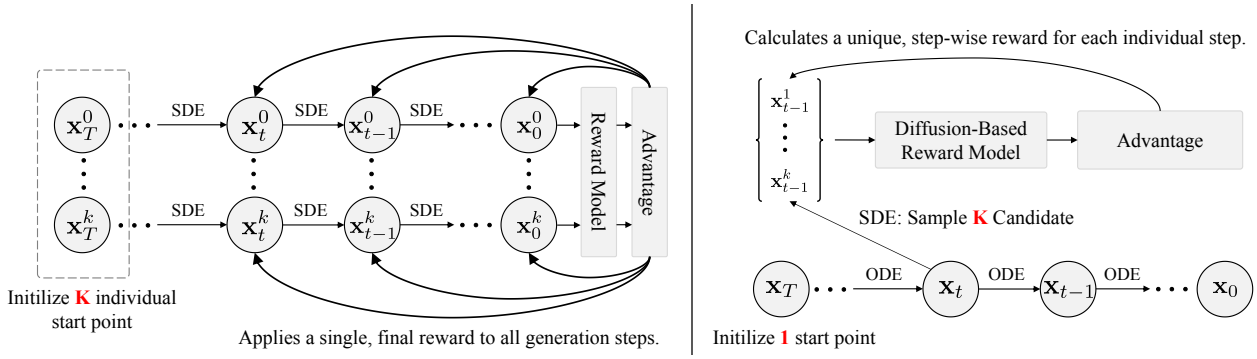

Figure 4. **GRPO vs. Step-wise GRPO.** (Left) Naive GRPO relies on a terminal reward. It samples multiple full trajectories, calculates a single reward at the final step (t=0), and applies this coarse reward uniformly to all preceding steps, leading to imprecise credit assignment. (Right) Step-wise GRPO introduces a dense, per-step reward signal. From a single initial point, it explores k candidate samples via SDE at each timestep, using the DRM to assign a precise, step-specific reward and advantage for more effective policy optimization.

3.3. Step-wise GRPO

Motivation. Prevailing reinforcement learning (RL) alignment algorithms, such as GRPO, largely treat the multi-step generation process as a "black box", performing time-agnostic policy optimization. This approach suffers from a fundamental limitation: it assigns the reward signal from the final generated image uniformly to every step in the generation trajectory. This coarse credit assignment mechanism overlooks the varying contributions of each intermediate step to the final image quality. To address this core issue, we leverage a unique capability of our Diffusion-based Reward Model (DRM). Because its backbone is initialized from pre-trained diffusion model weights, the DRM is inherently capable of evaluating noisy latents at any arbitrary timestep during the generation process. Capitalizing on this property, we introduce the Step-wise GRPO (Step-GRPO) algorithm. Instead of relying on a single, terminal reward, Step-GRPO provides a precise, step-specific reward for each intermediate state, enabling a more granular and effective policy optimization.

Step-wise GRPO (Step-GRPO). As illustrated in Figure 4, our method performs fine-grained policy optimization at each reverse diffusion timestep t. Specifically, starting from the current state $\mathbf{x}_{t+1}$, we sample a set of k candidate states for the next step, $\{\mathbf{x}_t^i\}_{i=1}^k$, via the SDE. These candidates are then fed into our DRM to obtain a corresponding set of immediate reward scores $\{R(\mathbf{x}_t^i, c)\}_{i=1}^k$ Unlike conventional advantage functions (Equal 3) that rely on a terminal reward, we define an immediate advantage for the decision at each step, formulated as:

$$\hat{A}_t^i = \frac{R(\mathbf{x}_t^i, \mathbf{c}) - \text{mean}(\{R(\mathbf{x}_t^i, \mathbf{c})\}_{i=1}^k)}{\text{std}(\{R(\mathbf{x}_t^i, \mathbf{c})\}_{i=1}^k)}. \quad (11)$$

This formulation shifts the focus of evaluation from the final, global outcome to the local decision at the current timestep—specifically, assessing the relative quality of transitioning from $\mathbf{x}_t$ to each candidate $\mathbf{x}_t^i$. This approach yields a more precise advantage estimate and provides a more direct and fine-grained supervisory signal for the policy gradient.

3.4. Step-wise Sampling

Beyond its role in providing step-wise rewards for RL-based fine-tuning, our Diffusion-based Reward Model (DRM) can also be leveraged to directly enhance generation quality at inference time. We introduce a novel sampling strategy, termed Step-wise Sampling, which offers

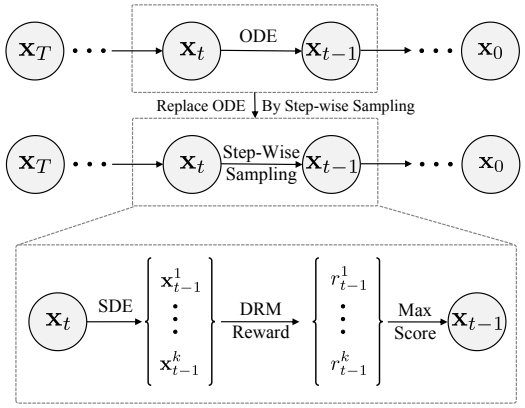

Figure 5. **Overview of Step-wise Sampling.** At each step t, we perform a branching into k candidates via SDE. The DRM scores these candidates, and the top-scoring latent is chosen to continue the trajectory.

a training-free, plug-and-play mechanism for improving model outputs. This approach provides a highly practical method for users to boost generation quality without any model fine-tuning. Conventional deterministic samplers follow a single, fixed trajectory, generating only one successor state $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ at each timestep. While efficient, this single-path approach is unforgiving; the quality of the final image is entirely dependent on the model's initial predictions, with no opportunity for corrective action during the generation process.

To overcome this limitation, Step-wise Sampling introduces an "explore-and-select" mechanism, as illustrated in Figure 5. At each timestep t, instead of following a single deterministic path, we first "explore" by leveraging SDE sample to generate k candidate states for the next step, forming a candidate set $\{\mathbf{x}_{t-1}^i\}_{i=1}^k$. This step effectively branches the generation process into multiple potential future trajectories. Next, in the "select" phase, we harness the unique power of our DRM to score each of these k candidates, obtaining a set of corresponding rewards $\{R(x_{t-1}^i, c)\}_{i=1}^k$. We greedily select the candidate with the highest score as the definitive state for the next step:

$$\mathbf{x}_{t-1} = \text{argmax}_{\mathbf{x}_{t-1}^i}(R(\mathbf{x}_{t-1}^i, c)). \tag{12}$$

By iteratively selecting the most promising path at each stage of generation, Step-wise Sampling can proactively steer the trajectory away from "bad" paths that might lead to low-quality results. This process robustly enhances the quality and alignment of the final image.

4. Experiment

4.1. Diffusion-based Reward Model

4.1.1. Implementation Details

Training Dataset. Following the methodology of HPSv3, we construct our training dataset by aggregating data from three sources: HPDv3, a subset of the Pick-A-Pic dataset, and a subset of the ImageReward dataset. The final dataset comprises a total of 1.4 million samples. Each sample is structured as a triplet (I^{win}, I^{lose}, p), consisting of a preferred image I^{win}, a dispreferred image I^{lose}, and their shared text prompt p.

Model. We initialize our backbone with the pretrained Diffusion Transformer from SD3.5-Medium. To ensure a fair comparison with VLM-based models in terms of parameter count, we truncate the final three transformer layers of the model. All remaining parameters are made trainable and are fine-tuned during training. The model is trained for one epoch on a cluster of 64 NVIDIA H20 GPUs, each with 96 GB of VRAM. We employ a constant learning rate of 1×10^{-5}, and a global batch size of 128, which corresponds to a per-GPU batch size of 2. All images are resized to a resolution of 512×512 pixels at training.

4.1.2. Preference Comparison

We evaluated our model against several leading reward models on standard benchmarks. As shown in Table 1, our approach achieves highly competitive performance, securing accuracies of 64.1% 73.4%, 82.2%, and 74.0% on the PickScore, HPDv2, and HPDv3 test sets, respectively. It is crucial to contextualize these results: unlike conventional RMs that are trained exclusively to judge final, clean images, our DRM is designed for the more challenging and general task of evaluating noisy latents at any step of the generation process. This broader training objective, which is fundamental to enabling our step-wise guidance methods, is not measured by standard benchmarks. This inherent design may introduce a slight trade-off in performance on clean-only evaluation tasks. In addition, the key advantage of our model lies in its remarkable parameter efficiency. Despite its modest size of only 2B parameters, our DRM significantly outperforms the similarly-sized, VLM-based HPSv3. This evidence strongly suggests that our diffusion-based architecture provides a more efficient and effective pathway to powerful reward modeling than simply scaling up conventional VLM backbones.

4.1.3. Ablation Study

Eeffect of Pretrained Weight. To validate our hypothesis that the model's strong performance stems from the generative prior embedded in the pre-trained diffusion weights, we conducted a critical ablation study. Specifically, we trained an identical model architecture from scratch, using random weight initialization instead of loading the pre-trained weights. The results after a single epoch of training are stark. As shown by comparing rows (a) and (e) in Table 1, the model initialized with pre-trained weights significantly outperforms the randomly initialized version across all test sets. To rule out the possibility that this discrepancy

#	Model	Weights	Epoch	Size	**ImageReward ↑**	**PickScore ↑**	**HPDv2 ↑**	**HPDv3 ↑**
-	CLIP ViT-H/14 [30]	-	-	-	57.1	60.8	65.1	48.6
-	Aesthetic Score Predictor [34]	-	-	-	57.4	56.8	76.8	59.9
-	ImageReward [42]	-	-	-	65.1	61.1	74.0	58.6
-	PickScore [14]	-	-	-	61.6	70.5	79.8	65.6
-	HPS [41]	-	-	-	61.2	66.7	77.6	63.8
-	HPSv2 [40]	-	-	-	65.7	63.8	83.3	65.3
-	MPS [47]	-	-	-	**67.5**	63.1	<u>83.5</u>	64.3
-	HPSv3 - 2B [27]	-	-	-	57.9	63.6	80.8	66.3
-	HPSv3 - 7B [27]	-	-	-	<u>66.8</u>	<u>72.8</u>	**85.4**	**76.9**
(a)	Ours	Random	1	256	52.4	57.5	65.0	59.3
(b)	Ours	Random	2	256	51.9	59.5	68.5	62.3
(c)	Ours	Random	3	256	53.7	59.0	70.1	63.0
(d)	Ours	Pre-trained	1	256	62.9	72.1	80.1	71.9
(e)	Ours	Pre-trained	1	512	64.1	**73.4**	82.2	<u>74.0</u>

Table 1. **Preference prediction accuracy (%) on the test sets of ImageReward, HPDv2 and HPDv3.** The best and second-best results are **bolded** and <u>underlined</u>. Our model achieves top-tier accuracy on PickScore. Its competitive scores on ImageReward, HPDv2 and HPDv3 reflect an expected trade-off, stemming from the DRM's core design. The DRM is trained to assess noisy latents throughout the generation process, not just the final clean outputs. This capability, fundamental to our approach, introduces a subtle domain shift when evaluated on benchmarks consisting solely of clean images, which accounts for the performance gap.

Timestep	**0**	**500**	**750**
DRM	74.0	73.0	65.11

Table 2. **DRM Accuracy vs. Timestep.** Performance on HPSv3 test set. Higher timestep correspond to higher noise level.

was merely due to the from-scratch model not having converged, we extended its training duration. The subsequent results, presented in rows (b) and (c), are conclusive: the pre-trained diffusion weights not only dramatically accelerate convergence but also enable the model to reach a higher performance ceiling. This confirms that the generative prior is indispensable for both training efficiency and the model's ultimate evaluative capabilities.

Eeffect of Training Image Size. To assess the impact of training image resolution on model performance, we conducted an ablation study comparing models trained on 256x256 and 512x512 images. The results are presented in Table 1, comparing rows (d) and (e). A clear trend emerges: increasing the training resolution leads to a consistent improvement in performance across all test sets. This indicates that our DRM has the capacity to leverage the fine-grained details present in higher-resolution data to make more accurate judgments, highlighting the importance of high-resolution training for achieving optimal performance.

4.1.4. Influence of Timestep

To validate the DRM's core capability of evaluating noisy latents, we tested its preference prediction accuracy on the HPSv3 test set across a spectrum of timestep. The results, detailed in Table 2, show a predictable decline in performance as the signal-to-noise ratio decreases. Nevertheless, the DRM's accuracy remains remarkably robust, confirming its efficacy as a reliable, step-wise reward signal throughout the entire denoising trajectory.

4.2. Step-Wise GRPO

4.2.1. Experimental Setting

Implementation Details. In our experiments, we benchmark the effectiveness of three distinct RMs: PickScore and HPSv3, and our proposed DRM. The base generative model for all experiments is SD3.5-Medium. To ensure a fair comparison, we fine-tune the generator using the state-of-the-art Flow-GRPO algorithm for all three RMs. For our DRM, we also conduct experiments with our novel Step-GRPO algorithm to showcase its unique step-wise guidance capabilities. For efficient fine-tuning, we optimize the generator using Low-Rank Adaptation (LoRA), with the rank set to 32 and the scaling factor α set to 64. We use a learning rate of 1e-4 with a policy clipping range of 1e-4. To ensure a fair comparison, we maintain the same workload across methods: standard GRPO aggregates 6 samples per GPU across 4 GPUs, resulting in a group size of 24 (6×4). In contrast, Step-GRPO (default $k = 6$) processes the same 24 samples total (6 per GPU) but computes updates using a local group size of 6 without cross-GPU aggregation. During inference, we employ the Flow Match Euler Discrete Scheduler with 50 sampling steps and a classifier-free guidance (CFG) scale of 4.5. To maintain a standardized benchmark, the evaluation is conducted on test datasets consistent with the Flow-GRPO.

Evaluation Metrics. For a comprehensive and objective evaluation of our method, we employ a suite of automated

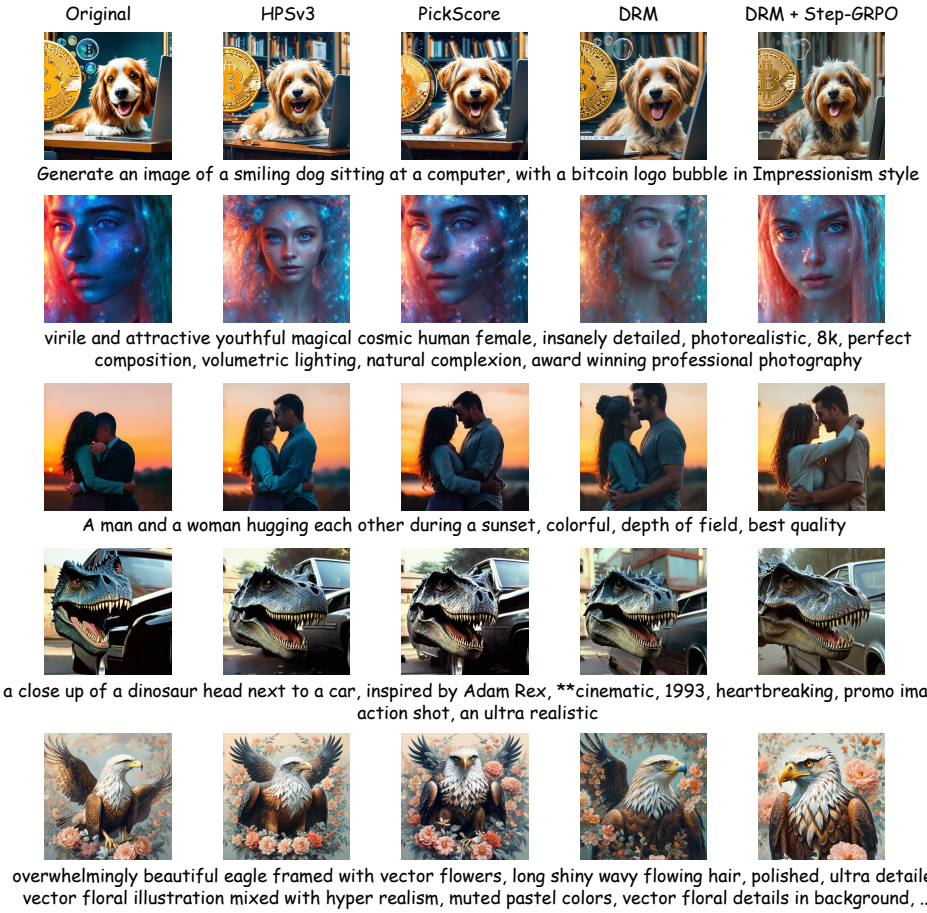

Figure 6. **Qualitative comparison of SD3.5-Medium optimized by various reward models.** Our approach clearly exhibits superior visual quality compared to the competing methods.

Model	ImageReward	PickScore	HPSv3
SD3.5-Medium	1.01	16.76	8.95
+ PickScore & GRPO	1.14	16.94	9.64
+ HPSv3 & GRPO	<u>1.15</u>	16.90	9.71
+ DRM & GRPO	1.14	<u>16.95</u>	<u>10.07</u>
+ DRM & Step-GRPO	**1.17**	**17.04**	**10.28**

Table 3. **Performance of SD3.5-Medium on the test set, optimized by different reward models.** Best and second-best results are in **bold** and <u>underlined</u>. Our full approach (DRM & Step-GRPO) outperforms all baselines, while DRM alone achieves the second-best performance on PickScore and HPSv3, validating the efficacy of both components.

metrics. Specifically, we utilize three models as evaluators: PickScore, HPSv3, and ImageReward. These models are established benchmarks for assessing critical aspects of generation quality, including text-image alignment, aesthetic appeal, and alignment with human preferences.

4.2.2. Quantitative Comparison

The quantitative results of RL fine-tuning experiments are summarized in Table 3. We evaluate the alignment of the fine-tuned models with human preferences using three established automated metrics: ImageReward, PickScore, and HPSv3. As shown, the baseline SD3.5-Medium model serves as our starting point. Applying the standard GRPO algorithm with any of the reward models—PickScore, HPSv3, or our DRM—yields consistent improvements across all evaluation metrics, validating the general effectiveness of RL-based fine-tuning. Notably, even when constrained to the standard GRPO framework, our DRM demonstrates highly competitive performance, particularly on the HPSv3 metric (10.07). However, the full potential of our approach is unlocked when our DRM is paired with the Step-GRPO algorithm. This combination decisively outperforms all other methods, establishing a new state-of-the-art across all three benchmarks. Specifically, our method achieves top scores of 1.17 on ImageReward, 17.04 on PickScore, and an impressive 10.28 on HPSv3. This consistent and superior performance provides strong empirical evidence for our central hypothesis: by leveraging a reward model capable of evaluating intermediate generation steps with an algorithm designed to utilize this granular feedback, we can achieve a more effective and robust

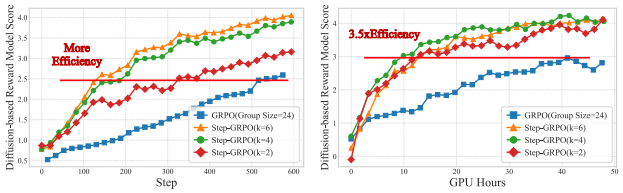

Figure 7. Reward curves with steps and GPU hours as the x-axis.

alignment with human preferences than methods that only provide a final reward.

4.2.3. Qualitative Comparison

To complement our quantitative findings, we conduct a qualitative analysis to visually assess the performance of our method. As illustrated in Figure 6, images generated by our DRM + Step-GRPO approach exhibit a clear superiority over those from the competing methods. Specifically, our method renders significantly more fine-grained details and shows a marked reduction in visual artifacts and generation errors. This advantage is particularly evident in its ability to preserve complex structures and generate realistic textures, areas where other methods often falter. These qualitative improvements provide compelling visual evidence that our approach excels at generating high-fidelity and aesthetically pleasing images, underscoring the benefits of leveraging step-wise guidance.

4.2.4. Training Efficiency and Convergence

As shown in Figure 2, Step-GRPO significantly outperforms standard GRPO in both convergence speed and final performance by leveraging step-wise feedback from a DRM. We conduct an ablation study on the group size, k, to further analyze its properties (Figure 7). The standard GRPO aggregates 6 samples per GPU across 4 GPUs, resulting in a group size of 24 (6×4), Step-GRPO (k=6) maintains the same workload (24 samples total, 6 per GPU), but computes updates using a local group size of 6 without cross-GPU aggregation. For smaller k values, we correspondingly set the per-GPU sample count to k (e.g., 2 samples per GPU for k=2). When measured by steps, our method exhibits superior convergence over GRPO even with k=2, and achieves faster reward growth as k increases (Figure 7 (left)). Regarding **GPU Hours**, our method converges $\sim3.5\times$ faster than GRPO (Figure 7 (right)). Notably, smaller k reduces per-iteration computational cost, resulting in similar GPU hour trajectories across $k\in\{2,4,6\}$.

4.3. Step-wise Sampling

In addition to Step-GRPO, we investigate the efficacy of Step-wise Sampling. Evaluated under the identical protocol described in Section 4.2, we further investigate the quality-efficiency trade-off by varying the candidate counts $k \in \{1,2,4,6\}$. As presented in Table 4, although generation time (512×512, 50 steps, bfloat16) inevitably scales with k, we observe consistent and notable improvements across all human preference metrics. Additionally, LPIPS evaluations confirm that this approach enhances diversity without inducing mode collapse. These quantitative gains are strongly corroborated by qualitative visual comparisons (Figure 8), which reveal superior visual quality, more coherent layouts, and enhanced aesthetic appeal. Together, these findings validate Step-wise Sampling as an effective inference technique that successfully balances computational cost with elevated generation quality and diversity.

Sampling	T(second)↓	ImageReward↑	PickScore↑	HPSv3↑	LPIPS↑
k=1	**2.88**	1.01	16.76	8.95	0.650
k=2	5.63	1.08	16.84	9.02	0.661
k=4	7.75	1.14	16.81	9.32	**0.663**
k=6	9.83	**1.15**	**16.93**	**9.49**	0.662

Table 4. **Performance of SD3.5-Medium on the test set with and without Step-wise Sampling.** It is evident that applying Step-wise Sampling leads to significant performance gains across all evaluation metrics.

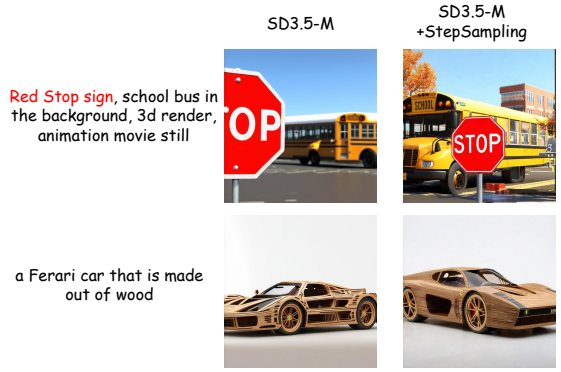

Figure 8. Step-wise Sampling enhances both the fidelity to the prompt and the aesthetic quality of the generated images.

5. Conclusion

In this paper, we introduced the Diffusion-based Reward Model (DRM), a novel paradigm that repurposes a pre-trained diffusion model's profound understanding of visual aesthetics as the evaluative backbone. The DRM's unique ability to assess noisy intermediate latents enabled two key innovations. For optimization, our Step-wise GRPO leverages dense, per-step rewards to resolve the credit assignment problem, achieving more stable and efficient alignment. For inference, our Step-wise Sampling strategy uses the DRM as a dynamic guide to proactively steer generation towards higher-quality results. Our extensive experiments confirm that DRM provides a more powerful solution for aligning diffusion models with human preference. We hope our work will inspire further exploration for reward modeling.

References

[1] Flux. https://github.com/black-forest-labs/flux/. 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[3] 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. 1, 2

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

[5] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 3

[6] Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025. 1

[7] 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. 1, 2

[8] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024. 1, 2

[9] Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[12] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 2

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[14] 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

[15] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1

[16] 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. 2

[17] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025. 1, 2

[18] Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025. 2

[19] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. 1, 2

[20] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[22] Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. *arXiv preprint arXiv:2512.21514*, 2025. 2

[23] 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

[24] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 1, 2

[25] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8009–8019, 2025. 1

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22139–22149, 2024. 2

[27] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025. 6

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[29] 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. 1, 2

[30] 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. 1, 2, 6

[31] 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

[32] 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. 1, 2

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 6

[35] 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. 1, 2

[36] Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit overoptimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 2025. 2

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2

[38] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024. 1, 2

[39] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 1, 2

[40] 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

[41] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023. 6

[42] 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. 1, 2, 6

[43] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 1, 2

[44] 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

[45] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 2

[46] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 1(2), 2025. 1

[47] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024. 1, 2, 6

[48] Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *arXiv preprint arXiv:2502.01051*, 2025. 2

[49] Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. Vq-insight: Teaching vlms for ai-generated

video quality understanding via progressive visual reinforcement learning. *arXiv preprint arXiv:2506.18564*, 2025. 1, 2