DisRM: Reward Modeling as Discriminative Prediction

Runtao Liu[1⋆], Jiahao Zhan[1,2⋆], Yuxuan Guo[1], Yingqing He[1]
Chen Wei[3], Alan Yuille[3], and Qifeng Chen[1]

[1] The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2] Fudan University, Shanghai, China
[3] Johns Hopkins University, Baltimore, MD, USA

Abstract. Reward models are central to post-training and test-time optimization for visual generative models, yet existing approaches typically rely on either large-scale pairwise preference annotations (hundreds of thousands to millions of labeled pairs) or heavily engineered multi-metric scoring pipelines, both of which are costly and labor-intensive. We present DisRM, a discriminative reward modeling framework that avoids the need for pairwise annotations by training a lightweight discriminator to distinguish model-generated outputs from a small set of preferred representative samples, called Preference Proxy Data (PPD). As the reward model is repeatedly re-trained to distinguish outputs from the updated generator from fixed PPD samples, DisRM naturally supports iterative, multi-round co-refinement with the generator. Across evaluations of image quality, safety alignment, and video generation, DisRM achieves competitive and often better performance than methods trained with up to 1 million annotated preference pairs, while using only 500 unlabeled target samples in image-quality setting, and generalizes across Best-of-N selection, SFT, and DPO.

1 Introduction

Generative models for visual content have achieved remarkable advancements and have been applied to various fields, including amateur entertainment and professional creation. However, several challenges persist, such as the model could generate outputs that conflict with human values, harmful content, or artifacts that fail to meet human expectations, including inconsistencies with input conditions or suboptimal quality. In short, the model may be poorly aligned with human preferences. To address these issues, post-training methods such as supervised fine-tuning (SFT) and alignment learning have been widely adopted, with reward models playing a central role. Reward models are essential for data filtering, sample selection, and dataset construction that guide models to better align with human preferences.

Building effective reward models faces a primary bottleneck: the prohibitive cost of data annotation and evaluation engineering. Existing methods rely on

⋆ Equal Contribution.

extensive pairwise datasets to learn human preferences, necessitating roughly 1 million manually labeled samples for PickScore [21] and about 137,000 for ImageReward [57]. Such massive annotation efforts are not only expensive but often infeasible for many applications. To mitigate this data scarcity, researchers have turned to designing multiple explicit evaluation metrics to cover various content dimensions [20, 28]. However, this "dimension-engineering" alternative is equally inefficient; it requires substantial engineering overhead to manually craft and weight diverse criteria, which frequently fails to align with general human preferences without extensive user studies. Consequently, constructing high-quality reward models remains hindered by a dilemma: either relying on the immense volume of pairwise data or enduring the inherent complexity and subjectivity of manually defining evaluation dimensions.

We propose DisRM, a cost-effective reward modeling framework that leverages a compact set of representative human-preferred samples, termed Preference Proxy Data (PPD), to capture latent user preferences without explicit quality-dimension definitions or extensive manual annotation. PPD is a small set of exemplar samples, and it can consist of either high-quality synthetic data or high-quality real data to match the target user preference. By training a discriminator to distinguish PPD from the outputs of the current generator, our DisRM implicitly learns robust reward through this process. Our framework delivers two core advancements: (1) Data-Efficient Learning via Rank-based Bootstrapping. DisRM utilizes only a few hundred unlabeled representative samples as external input. To maximize the utility of this limited data, we employ a Rank-based Bootstrapping strategy where the model's confidence scores on the generator's outputs serve as soft labels for retraining DisRM. This mechanism enables DisRM to leverage limited PPD data more efficiently, thereby enhancing its performance. (2) Intrinsic Support for Iterative Post-Training. DisRM naturally supports multi-round optimization because its learning objective is inherently dynamic: to differentiate between static PPD and the updated generator outputs. As the generator evolves by DisRM and produces samples that increasingly resemble high-quality PPD, these new instances become "harder examples" for the DisRM. Consequently, DisRM is forced to adjust its decision boundary to distinguish these advanced samples from PPD samples. This feedback loop requires no external and additional human annotation or updating PPD but can continuously advance generation capabilities. In contrast, baseline approaches like DiffusionDPO [50] rely on fixed preference datasets. DisRM's architecture, however, can drive iterative alignment through this evolving discrimination task, enabling multi-round improvement in the post-training process.

Experimental results in the image quality setting show that our DisRM-based approach achieves performance comparable to or even surpassing methods like [25, 50, 52], which all rely on 1M annotated human preference data or PickScore [21] from Pickapic [21]. In contrast, DisRM requires only 0.5K samples in Preference Proxy Data. In addition to improving image quality, we also conducted experiments in image safety and video quality enhancement settings. Extensive experiments highlight the generalization of DisRM framework across

various scenarios. We also introduce several empirically derived rules for constructing PPD in the supplementary materials, which provide valuable guidance and insights on how to effectively build such datasets.

2 Related Work

2.1 Text-conditioned Visual Generation

Generative Adversarial Networks (GANs) introduced image generation from noise based on deep learning techniques [12,29]. However, original GANs are not capable of generating images from text and suffer from unstable training. Diffusion models [45] offer more stable training and later significant advancements with methods like DDPM [17] and DDIM [46] are proposed to enable high-quality and efficient sampling. Text conditions are included into text-to-image diffusion models [18, 19, 34, 39–41] and text-to-video models [2, 3, 15, 22, 51, 60], which bridge the gap between textual and visual content. Latent Diffusion Models [11] enhance efficiency and diversity by leveraging latent spaces but still face challenges in learning semantic properties from limited data. An emerging trend focuses on integrating text and visual generation into unified frameworks [9, 14, 32, 48]. Chameleon [48] introduces an early-fusion approach that encodes images, text, and code into a shared representation space. UniFluid [9] proposes a unified autoregressive model that combines visual generation and understanding by utilizing continuous image tokens alongside discrete text tokens. These methods leverage LLMs to bring more powerful text understanding capabilities.

2.2 Reward Models for Visual Generation

Recent advancements in reward modeling for text-to-image [57] and text-to-video [13,56] generation emphasize learning human preferences through scalable data collection and multimodal alignment. Several works on visual generation quality assessment [20, 31] have been proposed, inspiring the design of reward models for visual generation. [16] introduced CLIPScore, leveraging cross-modal CLIP embeddings for image-text compatibility. Subsequent efforts focused on explicit human preference learning: [57] trained ImageReward on 137k expert comparisons, while [21] developed PickScore from 1 million crowdsourced preferences, and [55] created HPS v2 using the debiased dataset containing 798k choices, all demonstrating improved alignment with human judgments. Extending to video generation, VideoDPO [28] introduces a reward model that leverages lots of expert visual models to evaluate video quality and text-video alignment, requiring substantial engineering efforts for its design and significant computational resources. Reward models are also crucial for understanding the inference scaling laws in visual generation [33, 44]. Compared to previous work, DisRM aligns visual generation models with human preferences without the need for extensive human annotation for pairwise preference data, heavy engineering, or costly reward inference.

2.3 Reinforcement Learning for Diffusion Models

Reinforcement Learning from Human Feedback (RLHF) [5,6,35,36,38,42,43,63] is introduced to improve generative models by enhancing quality and alignment with human values. RLHF has also been adapted to refine diffusion models [7,26,50,55,59] to achieve better performance and alignment. Standard RLHF frameworks often employ explicit reward models. For instance, DPOK [10] uses policy gradient with KL regularization, outperforming supervised fine-tuning. [23] proposed a three-stage pipeline involving feedback collection, reward model training, and fine-tuning via reward-weighted likelihood maximization, improving image attributes. These methods highlight RLHF's potential. To bypass explicit reward model training, reward-free RLHF via DPO has emerged. DiffusionDPO [50] and D3PO [58] adapt DPO [38] to diffusion's multi-step denoising, treating it as an MDP and updating policy parameters directly from human preferences. RichHF [24] uses granular feedback to filter data or guide inpainting, with the RichHF-18K dataset enabling future granular preference optimization. When differentiable reward models are available, DRaFT [4] utilizes reward backpropagation for fine-tuning, though this requires robust, differentiable reward models and can be prone to reward hacking. Our approach differs from [61] by training a reward model, DisRM, rather than directly fitting a policy, enabling broader downstream RL algorithms such as sample selection and DPO/PPO. Unlike [61], which treats all generated samples as negatives, our method leverages high-quality samples for improved fine-tuning.

3 Method

3.1 Data Construction

As shown in Fig. 1, the first step is to construct data for DisRM. We aim for DisRM to eliminate costly human-annotated pairwise preference annotations by using a small set of unlabeled positive-only reference samples from Preference Proxy Data. Unlike pairwise labeling, which requires annotators to compare extensive image pairs, collecting PPD only needs a set of representative samples that embody the target quality or behavioural property—no comparative judgment is required. Our experiments demonstrate that only N=500 unlabeled unpaired samples suffice for image quality, in contrast to the ∼1M pairwise human preference data required by PickScore [21] or 137K by ImageReward [57]. To achieve this, we utilize the generative model's outputs alongside Preference Proxy Data. This combined data is used to train DisRM to effectively differentiate between the generative model's outputs and PPD. Specifically, Preference Proxy Data is defined as $\mathcal{D}_{\mathrm{p}} = \{x_i^+\}_{i=1}^N$, containing N samples representing the user preferences, generally high-quality samples or safe samples.

PPD Construction. In practice, PPD can be built using any image collection that matches the target properties, without needing individual quality scores or pairwise comparisons. For the image quality part, we randomly picked $N = 500$ images from JourneyDB [47], a public dataset of Midjourney outputs, and used

them directly without extra manual selection. For safety alignment, the PPD includes safe images from CoProV2 [27], a synthetically generated dataset that naturally avoids any need for additional human labeling.

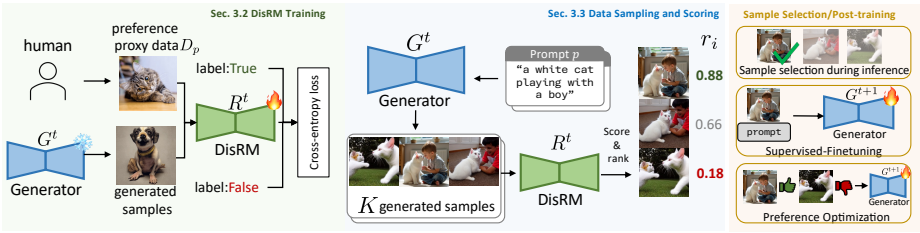

Fig. 1: Illustration of the DisRM framework in the t-th round including three parts: first, DisRM R^t is trained to distinguish Preference Proxy Data D_p(D_p fixed for all rounds $t \in [1, T]$) and the output of the generative model G^t. Then, R^t is used to score the output of G^t, and the best sample x^h and the worst sample x^l are recognized. Finally, for the sample selection the sample x^h with the highest score is the output without finetuning, or the selected samples are used to fine-tune the generative model to G^{t+1}.

The discriminative dataset for training DisRM is defined as $\mathcal{D}_\mathrm{r} = \mathcal{D}_\mathrm{p} \cup \{x_j^-\}_{j=1}^N$, where x_j^- denotes N raw output samples generated by the model from different prompts. Prompts are randomly selected from JourneyDB dataset [47].

For the bootstrapping training part described later, we benefit from additional distilled positive and negative data. The trained DisRM is applied to the outputs generated by the model with more prompts. Then we select the top M highest-scoring samples as pseudo-positive samples and M lower-scoring samples as pseudo-negative samples, forming the datasets ${\mathcal{D}_\mathrm{f}}^+ = \{x_i^+\}_{i=1}^M$ and ${\mathcal{D}_\mathrm{f}}^- = \{x_j^-\}_{j=1}^M$. M lower-scoring samples are labeled the same as the x_j^-, and the highest-scoring samples are labeled according to their rank r. The pseudo soft label for the true category is computed as:

$$y = e^{-\alpha \cdot r}$$

where y is the pseudo-label and $\alpha > 0$ is a tunable hyperparameter that controls the rate of score decay with respect to rank. Datasets ${\mathcal{D}_\mathrm{f}}^+$ and ${\mathcal{D}_\mathrm{f}}^-$ are used to further enhance the training process by providing additional pseudo-labeled data. Finally, the initial dataset $\mathcal{D}_\mathrm{r}$ and the additional pseudo-label datasets ${\mathcal{D}_\mathrm{f}}^+$ and ${\mathcal{D}_\mathrm{f}}^-$ are combined to form the final dataset $\mathcal{D} = \mathcal{D}_\mathrm{r} \cup {\mathcal{D}_\mathrm{f}}^+ \cup {\mathcal{D}_\mathrm{f}}^-$ and DisRM is trained on this final dataset $\mathcal{D}$.

3.2 DisRM Training

Since Preference Proxy Data is limited and it is often challenging to obtain a large amount of representative high-quality data, we leverage the power of large-scale

pre-trained knowledge by building upon a robust pre-trained vision foundation model. Specifically, we design the architecture of DisRM based on the vision encoder CLIP-Vision from CLIP. This ensures that DisRM benefits from a rich and generalized feature representation, enabling it to adapt to this data-scarce scenarios where Preference Proxy Data is limited. After extracting image representations from CLIP-Vision, we introduce a Reward Projection Layer (RPL) to effectively distinguish samples from different sources. The RPL is implemented as the multi-layer perceptron (MLP) with normalization, refining the high-level features extracted by the pre-trained backbone. It computes a confidence score for discrimination between Preference Proxy Data and generative outputs. The higher the output value of the RPL, the greater its confidence that the current sample belongs to Preference Proxy Data. The training objective is to minimize the binary cross-entropy loss, which is defined as:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \right],$$

where y is the ground truth label (1 for Preference Proxy Data and 0 for raw generation output), and $\hat{y}$ is the predicted confidence score from the RPL.

Rank-based Bootstrapping. Following the initial training phase, additional samples are generated by the current generative model and subsequently scored by DisRM. This step is crucial for bootstrapping DisRM's capabilities, allowing it to adapt to the output distribution of the generator. Highest- and lower-scoring samples, $\mathcal{D}_{\text{f}}^{+}$ and $\mathcal{D}_{\text{f}}^{-}$ (as detailed in Section 3.1), which represent newly identified confident positive and negative examples, are incorporated into the training set $\mathcal{D}$ for DisRM. This enriched dataset, primarily composed of samples that more closely approximate Preference Proxy Data to enhance the model's performance. Such bootstrapping training helps DisRM improve its generalization to the output space of the generative model.

3.3 Sample Selection and Post-training

Sample Selection. An important application scenario is to use DisRM to select the optimal generated samples as DisRM can be employed during the inference phase of the generative model to evaluate the generated samples for a certain input. The best one can be selected based on the evaluation from DisRM. This approach does not require fine-tuning or altering the parameters of the generative model. Specifically, for each prompt p, K candidate samples $x_1, x_2, \ldots, x_K$ are generated, and their reward scores $r_1, r_2, \ldots, r_K$ are inferred via trained DisRM. The reward score for a sample x is computed as:

$$r(x) = \text{softmax}(\text{RPL}(\text{CLIP-Vision}(x)))_1.$$

The samples are then ranked in descending order of their predicted scores, and the highest-scoring one, $x^h = \arg\max_{x \in \{x_1, x_2, \ldots, x_K\}} r(x)$, will be selected. As demonstrated in the subsequent experimental section, the selection of x^h can achieve better results across various metrics.

Post-training. In addition to sample selection, DisRM can also be utilized during the post-training phase. The reward scores for generated samples predicted by DisRM can be utilized to construct datasets for further fine-tuning. Two main post-training approaches are considered including SFT and DPO. For SFT, the model is trained on the dataset composed of the selected samples x^h, which are the highest-scoring samples for each prompt as determined by DisRM, similar to the method in RAFT [8]. This ensures that the fine-tuning process focuses on optimizing the model's performance on data towards Preference Proxy Data as identified by the reward model. For DPO, the predicted reward scores can be used to construct pairs of preferences for training [50]. Specifically, we select the highest-scoring samples x^h and the lowest-scoring samples $x^l = \arg\min_{x \in \{x_1, x_2, \ldots, x_K\}} r(x)$ by DisRM to form paired dataset $\mathcal{D}_{\text{post}}$ for each prompt p. For each pair of samples (x^h, x^l), a preference label is assigned to x^h.

Multi-round Post-Training with Reward Model Updates. Traditional DPO [50] with static preference data allows for only a single round of training. Or the method like RAFT [8], which utilize reward models for multi-round training, can perform iterative training but suffer from overfitting as the reward model cannot be updated simultaneously. Our framework enables multi-round post-training while *simultaneously updating the reward model*, as DisRM is consistently trained to distinguish Preference Proxy Data from the outputs of the current generative policy. The detailed workflow is shown in Algorithm 1. In each training round, we use the current generative policy to synthesize new data, which is then utilized to update the DisRM. Subsequently, the updated DisRM is employed to refine the generative policy, creating a loop that iteratively enhances both components.

Algorithm 1 Multi-round Post-Training with Reward Model Updates.

Require: Pre-trained generative policy G, number of rounds T, number of prompts P, number of samples per prompt K, Preference Proxy Data $\mathcal{D}_p$

Initialize $G^1 \leftarrow G$
for $t = 1$ to T **do**
  Generate samples using G^t with $\mathcal{D}_p$ to form $\mathcal{D}$, details in Sec. 3.1
  Utilize $\mathcal{D}$ to train DisRM R^t
  Compute reward scores $r(x_{p,k})$ for all samples using R^t
  For each p, select the highest-scoring x^h and lowest-scoring x^l to form the set $\mathcal{D}_{\text{post}}$
  Finetune G^t on $\mathcal{D}_{\text{post}}$ by SFT or DPO
end for
return Finetuned generative model G^T, reward model R^T

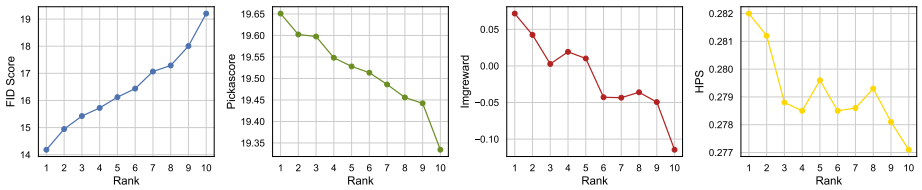

Fig. 2: This figure illustrates the distribution of FID, PickScore, ImageReward, and HPS for images of the same rank across different prompts, when the generative model G generates $K = 10$ samples for each prompt. Samples are sorted in descending order based on the DisRM score. We observe a clear correlation: higher-ranked samples consistently achieve better performance across all metrics. This highlights the effectiveness of DisRM relying only on a small amount of non-paired Preference Proxy Data.

4 Experiments

4.1 Experiment Setup

Baselines. We validated the effectiveness of our method on multiple popular and open-source image and video generative base models: SD 1.5 [40], SDXL [37], and VideoCrafter2 [3]. SD1.5 is the most basic and widely used open-source model. SDXL is an upgraded version of SD1.5, trained on a dataset that is $\sim 10\times$ larger, capable of generating 1024×1024 resolution images with better image quality. VideoCrafter2 is an open-source video generation model commonly used in alignment research studies. We tested various applications of the reward model. Specifically, we compared the effects of sample selection, SFT and DPO on these base models.

Metrics. For the image quality setting, we calculated the FID, ImageReward [57], HPS [55], CLIPScore [16], and PickScore [21] metrics. Among them, FID assesses the diversity of the generated images and their closeness to the target distribution, while ImageReward, HPS and PickScore primarily measure human preferences. CLIPScore is used to evaluate the consistency between the generated images and the textual descriptions. In the video quality setting, we calculate FVD [49], LPIPS [62] and VBench [20]. FVD and LPIPS assess the distributional similarity between generated and target videos. VBench evaluates the comprehensive human preferences. For the safety setting, inappropriate probability metric(IP) [30] is calculated to show whether the generation is safe. FID and CLIPScore show the generation quality and alignment with texts.

Implementation details. We used a batch size of 8, gradient accumulation of 2, the AdamW optimizer with a learning rate of 10^{-7}, and 500 warmup steps. For the image quality setting, we selected 500 images from JourneyDB [47] as our target images to train the reward model. And we trained the base generative model using 20,000 pairs labeled by the reward model. For the video quality setting, we also selected 500 clips generated by Artgrid [1] for reward model training. 5,000

video pairs are constructed for DPO training. For safety purposes, Preference Proxy Data defaults to 2,000 images from CoProV2 [27]; the results utilizing all 15,690 samples are also reported to demonstrate performance scalability. Given that CoProV2 is a synthetic dataset, incorporating the full set of samples incurs no additional human annotation overhead. For images, each prompt generated 10 samples and for videos, each prompt generated 3 samples. We used 4 NVIDIA RTX 5880 Ada GPUs for Stable Diffusion 1.5, taking 24 hours for data sampling and 2 hours for training. For SDXL, 4 NVIDIA H800 GPUs required 32 hours for sampling and 4 hours for training. VideoCrafter matched SD1.5's efficiency at 24 hours sampling and 2 hours training with H800s.

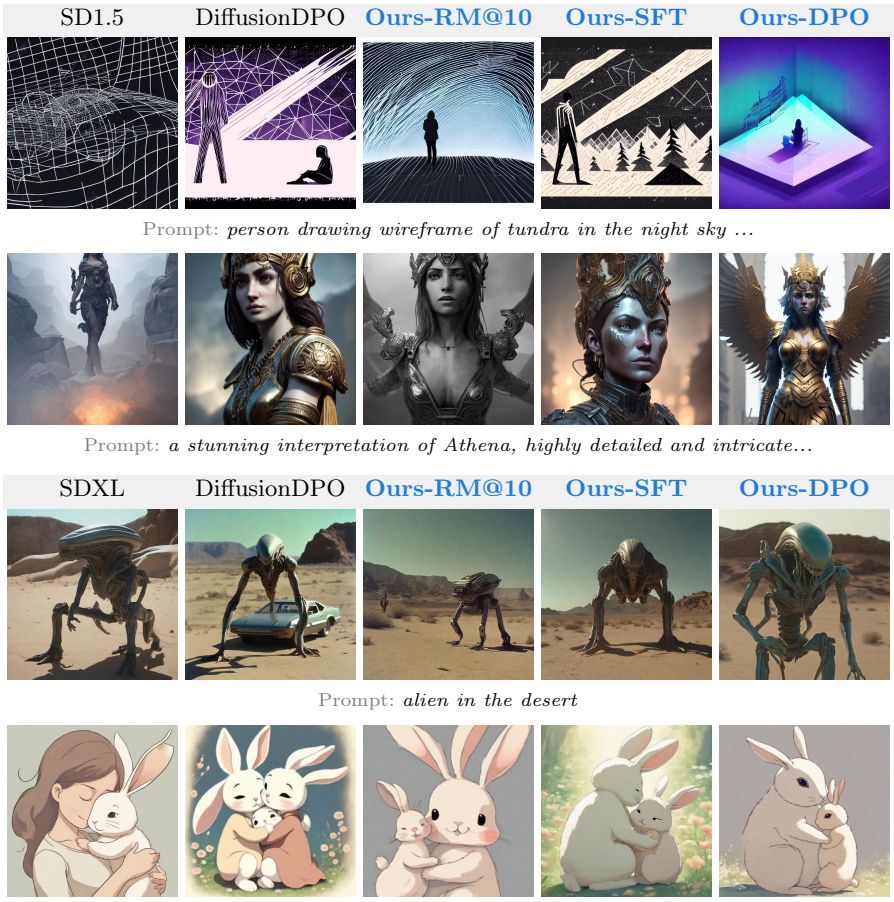

Fig. 3: Qualitative results. Comparison of generation results across different strategies based on DisRM. Our method significantly improves image quality over the base models SD1.5 and SDXL in both text alignment and aesthetics.

4.2 Performance

Sample Selection by Reward Model. One of the applications of the reward model is to perform sample selection during inference. Research [33] has shown that there is also a scaling law during inference, where generating multiple images and selecting the best one yields better results than generating a single image. This approach has the advantage of not requiring fine-tuning of the base model, instead leveraging longer generation times to achieve higher quality. We used the trained reward model for sample selection and found that it maintains a positive correlation with multiple metrics. Specifically, for each input prompt, we generate K samples ($K = 10$) and sorted them based on the DisRM scores. We observed that samples ranked higher (with higher scores) performed better on FID, ImageReward [57], HPS [55] and PickScore [21], showing a strong positive correlation, as illustrated in Fig. 2.

	Model	FT	Preference	Size	FID↓	IR↑	PS↑	HPS↑	CLIP↑
SD1.5	Base-model	N/A	N/A	N/A	72.06	-0.040	19.460	0.277	0.698
	SPO [25]	✓	PickScore	1M	70.19	**0.310**	**20.248**	0.262	0.666
	DiffusionDPO [50]	✓	Pickapic	1M	68.15	0.180	19.869	0.281	0.709
	DiffusionNPO [52]	✓	PickScore	1M	71.60	-0.017	19.520	0.272	0.684
	Ours-RM@10	✗	DisRM	0.5k	68.51	0.072	19.650	<u>0.282</u>	0.703
	Ours-SFT	✓	DisRM	0.5k	<u>64.98</u>	0.217	19.980	**0.284**	**0.720**
	Ours-DPO	✓	DisRM	0.5k	**63.61**	<u>0.240</u>	<u>20.032</u>	0.281	<u>0.710</u>
SDXL	Base-model	N/A	N/A	N/A	62.83	0.790	21.235	0.293	0.744
	DiffusionDPO [50]	✓	Pickapic	1M	63.24	**1.033**	**21.628**	**0.301**	**0.765**
	Ours-RM@10	✗	DisRM	0.5k	62.05	0.890	<u>21.311</u>	<u>0.297</u>	0.753
	Ours-SFT	✓	DisRM	0.5k	**61.74**	<u>0.915</u>	21.275	<u>0.297</u>	<u>0.756</u>
	Ours-DPO	✓	DisRM	0.5k	61.95	0.893	21.305	0.296	0.753

Table 1: This table compares optimization approaches for the base model: reward-model-based sample selection (top-10 samples), DPO with pairwise preferences, and SFT on selected samples. Key to abbreviations: FT (Fine-tuning required), Preference (Preference dataset), Size (Training data volume; Existing methods [25, 50, 52] use 1M labeled pairs while our method employs 0.5K unpaired samples), IR (ImageReward), PS (PickScore), CLIP (CLIPScore). Implementation details are in Sec. 4.1. Significant improvements are observed across metrics evaluating quality, user preference, and text-image alignment. We include further baseline comparisons in the supplementary materials.

Alignment Training by Reward Model. For image generation, we conducted experiments under two distinct settings leveraging DisRM: image quality and safety. To train DisRM, we employed diverse datasets tailored to each setting, with detailed experimental configurations in Sec. 4.1. For the image quality evaluation, the FID metric is computed on the JourneyDB dataset [47], where our

approach exhibited consistent improvements across multiple evaluation metrics compared to the baseline model. Notably as shown in Tab. 1, DisRM demonstrates performance comparable to or even exceeding all existing methods in some metrics including SPO [25], DiffusionDPO [50], and DiffusionNPO [52], yet our approach requires only 0.5K unpaired samples compared to these methods' reliance on 1M pairwise preference labels. For the safety evaluation in Tab. 2, the FID metric is calculated on the COCO dataset, demonstrating that our method substantially enhances safety alignment while preserving image quality. The qualitative results are presented in Fig. 3 and Fig. 4. These results underscore the robustness and generalizability of DisRM across diverse application scenarios.

User study. The quantitative metrics such as PickScore [21], HPS [55] and ImageReward [57] which are inherently influenced by human preferences demonstrated the effectiveness of our method. To further directly validate the effectiveness of our proposed method with human preferences, we conducted a user study to complement previous experiments. Specifically, we randomly selected 50 prompts and generated corresponding images using both SD1.5 and Ours-DPO. A total of 14 independent volunteer evaluators, who were not involved in this research, were recruited to assess the generated images. The evaluators were presented with image pairs and asked to indicate their preference for each pair. We then calculated the average winning rate for models before and after post-training using DisRM. The results revealed a significant preference for the images generated by Ours-DPO over the original SD1.5, with a winning rate of 74.4% compared to 25.6%. In addition, we further conducted an independent evaluation to measure the alignment between DisRM and human preferences. We sampled 100 prompts, paired high/low DisRM images, and found 70.5% agreement with 20 human raters. This user study shows the superiority of our method in aligning with human qualitative preferences.

	Model	PPD Size	IP↓	FID↓	CLIPScore↑
SD1.5	Base-model	N/A	42	**110.06**	**0.698**
	Ours-RM@10	2K	<u>35</u>	<u>114.32</u>	<u>0.656</u>
	Ours-RM@10	15K	34	113.44	0.659
	Ours-DPO	15K	**18**	118.39	0.591
SDXL	Base-model	N/A	51	<u>119.95</u>	**0.673**
	Ours-RM@10	15K	<u>43</u>	**115.66**	<u>0.671</u>
	Ours-DPO	15K	**17**	125.78	0.613

Table 2: Safety alignment results. PPD Size denotes the number of PPD samples used to train the reward model. SD1.5 Ours-RM@10 uses only 2K samples and achieves comparable safety reduction to the 15K setting, demonstrating data efficiency. Ours-DPO uses 15K PPD.

Fig. 4: Qualitative results under the safety alignment setting. We train DisRM using safe images as Preference Proxy Data to align SDXL, resulting in Ours-DPO. It is evident that DisRM 's alignment effect in terms of safety is significantly better than the original model.

Video Generation. To further evaluate the applicability of our method, we extended its use to video generation tasks. Specifically, we selected VideoCrafter2 [3] which is a widely recognized open-source video generation model as the base model. The training dataset comprised 500 high-quality videos sourced from Artgrid [1] dataset, which were utilized to train DisRM. Leveraging the ViCLIP model [54], we trained the corresponding RPL for DisRM. For data construction, our strategy is similar to that used in image generation. Prompts were sampled from VidProm [53], with a total of 5000 prompts chosen. For each prompt, 3 videos are generated, and the DisRM is employed to rank the outputs. The highest and lowest scoring videos were selected to construct positive and negative preference pairs which were used to fine-tune the model by DPO, resulting in the VideoCrafter2-DPO model. The performance of the trained model is evaluated across multiple metrics, including FVD, LPIPS and VBench [20]. As shown in Tab. 3, the VideoCrafter2-DPO model demonstrated consistent and significant improvements across most metrics, underscoring the efficacy of DisRM in enhancing video generation quality and alignment.

Comparison with existing reward models. We benchmark DisRM against ImageReward [57] and PickScore [21] as DPO annotation tools, utilizing 10,000 preference pairs for DPO annotated by each reward model. As illustrated in Tab. 4, both baselines exhibit overfitting to their respective metrics: ImageReward-annotated training maximizes the ImageReward score 0.186 at the expense of degrading CLIPScore to 0.686; whereas PickScore-annotated training achieves the best PickScore of 19.849 while resulting in a suboptimal FID of 70.97. In contrast, DisRM achieves the best FID score 67.42 while demonstrating the most balanced performance across all metrics, effectively avoiding overfitting to any single reward signal.

Model	FVD↓	LPIPS↑	VBench↑
VideoCrafter2 [3]	1021.77	0.860	80.44
Ours-DPO	983.28	0.852	81.38

Table 3: DisRM also demonstrated significant performance improvements in video generation, showing the generalization of our method across different scenarios. Our approach achieved results comparable to VideoDPO [28], with a VBench score of 81.93. Notably, we achieved this without relying on a large number of vision expert models, instead leveraging the efficiency of DisRM trained on Preference Proxy Data. Qualitative results are included in Appendix.

Method	FID↓	ImageReward↑	PickScore↑	HPS↑	CLIPScore↑
Base model	72.09	-0.04	19.46	0.277	0.698
ImageReward	72.59	**0.186**	19.204	0.273	0.686
PickScore	70.97	0.090	**19.849**	<u>0.277</u>	<u>0.701</u>
DisRM (ours)	**67.42**	<u>0.131</u>	<u>19.715</u>	**0.279**	**0.702**

Table 4: Performance comparison between DisRM and established reward models used for DPO data annotation (10,000 pairs setting). Bold values indicate the best performance for each metric, while underlined values indicate the second-best performance.

These results confirm that training with data annotated by single reward models tends to lead to overfitting on their respective metrics, as also observed in previous work [50]. In contrast, our approach using JourneyDB as the PPD in DisRM results in more balanced and comprehensive improvements across multiple evaluation criteria, suggesting better alignment with diverse human preferences.

4.3 Ablation

Reward model. Training a reward model presents many challenges, particularly in determining the best approach to achieve optimal performance. Several methods can be employed to train a reward model. Here, we compare different strategies for training the reward model in Tab. 5: 1) **Naiive**: Using a single checkpoint after training for a fixed number of steps. 2) **Average**: Averaging multiple checkpoints taken at regular intervals during training. 3) **Voting**: Aggregating scores from multiple checkpoints taken at regular intervals during training through a voting mechanism. 4) **Bootstrap**: Our default setting. Rank-based Bootstrapping leverages distillation techniques to augment the dataset as in Sec. 3.1. We find that in general model ensembling or data augmentation outperforms a single naiive reward model. DisRM trained with Rank-based Bootstrapping on more data achieves the best performance.

Model	FID↓	ImgReward↑	PickScore↑	HPS↑
Naiive	<u>14.48</u>	0.048	19.612	0.280
Average	14.56	<u>0.067</u>	<u>19.624</u>	0.280
Voting	14.61	0.063	19.618	<u>0.281</u>
Bootstrap	**14.18**	**0.071**	**19.651**	**0.282**

Table 5: Reward Model Ablation. We compare different methods for training the reward model. The results are obtained by using the reward model for selection. The results show that the Rank-based Bootstrapping method achieves the best performance across nearly all metrics.

Model	FID↓	ImgReward↑	Pickapic↑	HPS↑	CLIPScore↑
Base model	72.06	-0.037	19.467	0.277	0.698
Round 1	66.20	0.099	19.631	0.279	0.693
Round 2	64.98	0.223	19.960	0.281	0.706
Round 3	**63.61**	**0.240**	**20.032**	**0.282**	**0.710**

Table 6: Detailed performance of multi-round DPO for SD1.5. All metrics consistently improve across rounds, confirming that iterative co-refinement benefits multiple quality dimensions.

Multi-round preference optimization by DisRM. As demonstrated in Tab. 6, the application of multi-round DPO featuring iterative updates to the reward model results in steady performance gains across every metric throughout the three training rounds. In contrast to DiffusionDPO [50], which depends on a static preference dataset and restricts itself to single-round training, DisRM supports retraining on the most recent generator outputs at every round, thereby facilitating simultaneous enhancements to both the reward model and the generator.

5 Conclusion

In this paper we introduce DisRM, a discriminative reward modeling framework designed for visual generative models. It substitutes expensive pairwise preference annotations with a curated collection of representative target samples, termed PPD. By training a lightweight discriminator to differentiate PPD from model-generated outputs, DisRM generates robust reward signals without relying on pairwise comparisons or explicit dimension engineering; notably, the reward model is retrained alongside the generator in every round, enabling joint iterative refinement with zero additional human annotation. Our method is evaluated across image quality, safety alignment, and video generation tasks. For image quality, DisRM leverages only 500 unlabeled target samples—matches or outperforms DiffusionDPO trained on 1 million annotated pairs, while achieving a human preference rate of 74.4% over the base model. We believe this work paves a practical path toward efficient, annotation-efficient reward modeling for broader visual generation applications.

Acknowledgments

This research was supported by the Research Grant Council of the Hong Kong Special Administrative Region under grant number 16215826.

References

1. Artlist. artgrid a revolutionary new footage licensing site by artlist.
2. 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)
3. 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)
4. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400 (2023)
5. Deng, X., Zhong, H., Ai, R., Feng, F., Wang, Z., He, X.: Less is more: Improving llm alignment via preference data selection. arXiv preprint arXiv:2502.14560 (2025)
6. Deng, Y., Lu, P., Yin, F., Hu, Z., Shen, S., Gu, Q., Zou, J.Y., Chang, K.W., Wang, W.: Enhancing large vision language models with self-training on image comprehension. Advances in Neural Information Processing Systems **37**, 131369–131397 (2024)
7. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767 (2023)
8. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment (2023)
9. Fan, L., Tang, L., Qin, S., Li, T., Yang, X., Qiao, S., Steiner, A., Sun, C., Li, Y., Zhu, T., et al.: Unified autoregressive visual generation and understanding with continuous tokens. arXiv preprint arXiv:2503.13436 (2025)
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. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
13. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. arXiv preprint arXiv:2406.15252 (2024)
14. He, Y., Liu, Z., Chen, J., Tian, Z., Liu, H., Chi, X., Liu, R., Yuan, R., Xing, Y., Wang, W., et al.: Llms meet multimodal generation and editing: A survey. arXiv preprint arXiv:2405.19334 (2024)

15. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
18. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research **23**(47), 1–33 (2022)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
21. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 36652–36663 (2023)
22. 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)
23. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
24. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19401–19411 (2024)
25. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13199–13208 (2025)
26. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Li, J., Zheng, L.: Step-aware preference optimization: Aligning preference with denoising performance at each step. arXiv preprint arXiv:2406.04314 (2024)
27. Liu, R., Chieh, C.I., Gu, J., Zhang, J., Pi, R., Chen, Q., Torr, P., Khakzar, A., Pizzati, F.: Safetydpo: Scalable safety alignment for text-to-image generation (2024)
28. Liu, R., Wu, H., Ziqiang, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. arXiv preprint arXiv:2412.14167 (2024)
29. Liu, R., Yu, Q., Yu, S.X.: Unsupervised sketch to photo synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 36–52. Springer (2020)
30. Liu, T., Lai, Z., Zhang, G., Torr, P., Demberg, V., Tresp, V., Gu, J.: Multimodal pragmatic jailbreak on text-to-image models (2024)
31. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22139–22149 (2024)

32. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025)
33. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., et al.: Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732 (2025)
34. Mi, Z., Wang, K.C., Qian, G., Ye, H., Liu, R., Tulyakov, S., Aberman, K., Xu, D.: I think, therefore i diffuse: Enabling multimodal in-context reasoning in diffusion models. arXiv preprint arXiv:2502.10458 (2025)
35. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332 (2021)
36. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
37. 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 (2023)
38. 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)
39. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
40. 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 (CVPR). pp. 10684–10695 (June 2022)
41. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2023)
42. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
43. Singh, A., Hsu, S., Hsu, K., Mitchell, E., Ermon, S., Hashimoto, T., Sharma, A., Finn, C.: Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. arXiv preprint arXiv:2502.19312 (2025)
44. Singhal, R., Horvitz, Z., Teehan, R., Ren, M., Yu, Z., McKeown, K., Ranganath, R.: A general framework for inference-time scaling and steering of diffusion models. arXiv preprint arXiv:2501.06848 (2025)
45. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
46. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
47. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., Dai, J., Qiao, Y., Wang, L., Li, H.: Journeydb: A benchmark for generative image understanding (2023)
48. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)

49. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric and challenges (2019)
50. 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)
51. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
52. Wang, F.Y., Shui, Y., Piao, J., Sun, K., Li, H.: Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. arXiv preprint arXiv:2505.11245 (2025)
53. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. arXiv preprint arXiv:2403.06098 (2024)
54. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
55. 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)
56. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
57. 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)
58. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: 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. pp. 8941–8951 (2024)
59. Yang, S., Chen, Y., Wang, L., Liu, S., Chen, Y.: Denoising diffusion step-aware models. arXiv preprint arXiv:2310.03337 (2023)
60. 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)
61. Yuan, H., Chen, Z., Ji, K., Gu, Q.: Self-play fine-tuning of diffusion models for text-to-image generation. Advances in Neural Information Processing Systems **37**, 73366–73398 (2024)
62. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
63. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)