

FAKE IT TILL YOU MAKE IT: REWARD MODELING AS DISCRIMINATIVE PREDICTION

Anonymous authors

Paper under double-blind review

ABSTRACT

An effective reward model plays a pivotal role in reinforcement learning for post-training enhancement of visual generative models. However, current approaches of reward modeling suffer from implementation complexity due to their reliance on extensive human-annotated preference data or meticulously engineered quality dimensions that are often incomplete and engineering-intensive. Inspired by adversarial training in generative adversarial networks (GANs), this paper proposes GAN-RM, an efficient reward modeling framework that eliminates manual preference annotation and explicit quality dimension engineering. Our method trains the reward model through discrimination between a small set of representative, unpaired target samples (denoted as Preference Proxy Data) and model-generated ordinary outputs, requiring only a few hundred target samples. Comprehensive experiments demonstrate our GAN-RM’s effectiveness across multiple key applications including test-time scaling implemented as Best-of-N sample filtering, post-training approaches like Supervised Fine-Tuning (SFT) and Direct Preference Optimization (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 could be not well aligned with human preference.

Post-training, including supervised fine-tuning and alignment learning, have been proposed to address these issues, with reward models playing a pivotal role. Reward models are essential for data filtering, sample selection or constructing datasets that guide models to better align with human preferences. This paper proposes an efficient, low-cost, yet highly effective reward model and validates its effectiveness in the test-time scaling and post-training of visual generative models.

Building effective reward models presents significant challenges. First, constructing reward models often requires extensive datasets. Existing methods Kirstain et al. (2023); Xu et al. (2023) require hundreds of thousands to millions of manually labeled samples, which are expensive to collect. These datasets are typically annotated based on the output domain of a specific generative model, resulting in a domain gap when applying the trained reward model to generative models with different output domains. Additionally, to comprehensively evaluate the quality of generated content across multiple dimensions, existing methods often require the manual design of various evaluation metrics Huang et al. (2024); Liu et al. (2024b). This not only increases engineering costs but may also lead to suboptimal trade-offs between different dimensions. Moreover, it is difficult to ensure that the defined dimensions and their aggregation methods align well with general human preferences, often necessitating user studies to evaluate alignment Huang et al. (2024); Liu et al. (2024b). In summary, the challenges of constructing reward models include the difficulty of obtaining data, reliance on specific model output domains in terms of data, and the inherent subjectivity of human preferences, which are hard to define through designing dimensions.

Inspired by adversarial learning Goodfellow et al. (2020), we propose GAN-RM, an efficient and cost-effective reward modeling framework that leverages a small set of representative human-preferred samples—referred to as Preference Proxy Data. These samples encapsulate latent human preferences

without requiring manual annotation or explicit specification of quality dimensions. Our method offers several advantages: (1) GAN-RM eliminates the necessity for manual preference annotation. The only external data is a small set of unlabeled (a few hundred) representative samples, denoted as Preference Proxy Data. GAN-RM is trained to distinguish Preference Proxy Data from generative model outputs, thereby learning to assess generated samples. We employ a Rank-based Bootstrapping strategy, where the confidence scores from GAN-RM on these samples serve as soft labels. This approach leverages the additional data to retrain GAN-RM, enabling it to better capture latent human preferences. (2) GAN-RM supports multi-round post-training. In each round, samples identified as close to Preference Proxy Data are used to post-train the generator. In turn, the discriminator is retrained to differentiate these harder examples. Such iterative "fake it" process can progressively aligns generation quality with latent human preferences in Preference Proxy Data.

Experimental results show that our GAN-RM-based approach achieves performance comparable to or even surpassing methods like Wallace et al. (2024), which rely on 1M annotated human preference data from Pickapic Kirstain et al. (2023). In contrast, GAN-RM requires only 0.5K samples in Preference Proxy Data for the image quality experiment setting. In addition to improving image quality, we also conducted experiments in image safety and video quality enhancement settings. Extensive experiments highlight the generalization of GAN-RM framework across various scenarios.

2 RELATED WORK

2.1 TEXT-CONDITIONED VISUAL GENERATION

Generative Adversarial Networks (GANs) introduced image generation from noise based on deep learning techniques Goodfellow et al. (2020); Liu et al. (2020). However, original GANs are not capable of generating images from text and suffer from unstable training. Diffusion models Sohl-Dickstein et al. (2015) offer more stable training and later significant advancements with methods like DDPM Ho et al. (2020) and DDIM Song et al. (2020) are proposed to enable high-quality and efficient sampling. Text conditions are included into text-to-image diffusion models Rombach et al. (2022); Ramesh et al. (2022); Ho et al. (2022); Saharia et al. (2022); Ho & Salimans (2022); Mi et al. (2025) and text-to-video models Chen et al. (2024); Blattmann et al. (2023); Kong et al. (2024); Wang et al. (2025a); He et al. (2022); Yang et al. (2024b), which bridge the gap between textual and visual content. Latent Diffusion Models Gal et al. (2022) 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 Ma et al. (2025a); Fan et al. (2025); Team (2024); He et al. (2024b). Chameleon Team (2024) introduces an early-fusion approach that encodes images, text, and code into a shared representation space. UniFluid Fan et al. (2025) 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 Xu et al. (2023) and text-to-video He et al. (2024a); Xu et al. (2024) generation emphasize learning human preferences through scalable data collection and multimodal alignment. Several works on visual generation quality assessment Huang et al. (2024); Liu et al. (2024d) have been proposed, inspiring the design of reward models for visual generation. Hessel et al. (2021) introduced CLIPScore, leveraging cross-modal CLIP embeddings for image-text compatibility. Subsequent efforts focused on explicit human preference learning: Xu et al. (2023) trained ImageReward on 137k expert comparisons, while Kirstain et al. (2023) developed PickScore from 1 million crowdsourced preferences, and Wu et al. (2023) created HPS v2 using the debiased dataset containing 798k choices, all demonstrating improved alignment with human judgments. Extending to video generation, VideoDPO Liu et al. (2024b) 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 Ma et al. (2025b); Singhal et al. (2025). Compared to previous work, GAN-RM aligns visual generation models with human preferences without the need for extensive human annotation, heavy engineering, or costly reward inference.

2.3 REINFORCEMENT LEARNING FOR DIFFUSION MODELS

Reinforcement Learning from Human Feedback (RLHF) Schulman et al. (2017); Ouyang et al. (2022); Ziegler et al. (2019); Rafailov et al. (2023); Nakano et al. (2021); Deng et al. (2024); Singh et al. (2025); Deng et al. (2025) is introduced to improve generative models by enhancing quality and alignment with human values. RLHF has also been adapted to refine diffusion models Dong et al. (2023a); Wallace et al. (2024); Yang et al. (2023); Liang et al. (2024b); Wu et al. (2023) to achieve better performance and alignment. Standard RLHF frameworks often employ explicit reward models. For instance, DPOK Fan et al. (2023) uses policy gradient with KL regularization, outperforming supervised fine-tuning. Lee et al. (2023) 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 Wallace et al. (2024) and D3PO Yang et al. (2024a) adapt DPO Rafailov et al. (2023) to diffusion’s multi-step denoising, treating it as an MDP and updating policy parameters directly from human preferences. RichHF Liang et al. (2024a) 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 Clark et al. (2023) utilizes reward backpropagation for fine-tuning, though this requires robust, differentiable reward models and can be prone to reward hacking. Our approach differs from Yuan et al. (2024) by training a reward model, GAN-RM, rather than directly fitting a policy, enabling broader downstream RL algorithms such as sample selection and DPO/PPO. Unlike Yuan et al. (2024), which treats all generated samples as negatives, our method leverages high-quality samples for improved fine-tuning.

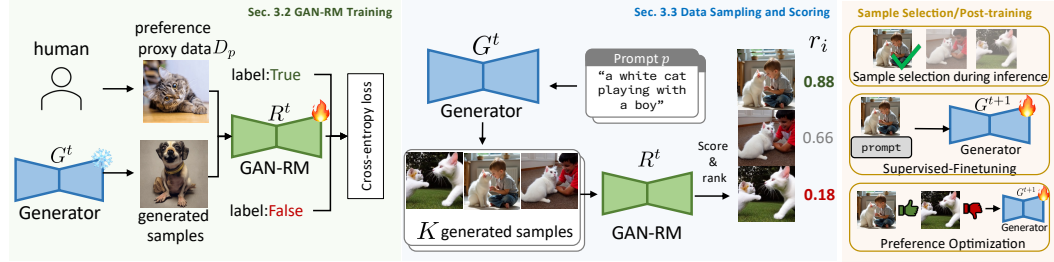


Figure 1: Illustration of the GAN-RM framework in the t -th round including three parts: first, GAN-RM R^t is trained to distinguish Preference Proxy Data D_p (D_p fixed for all rounds $t \in [1, T]$) and the output of the generative model G^t . Then, R^t is used to score the output of G^t , and the 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} .

3 METHOD

3.1 DATA CONSTRUCTION

As shown in Fig. 1, the first step is to construct data for GAN-RM. We aim for GAN-RM to be trained without relying on human preference annotations but only on the data provided by the users called Preference Proxy Data. To achieve this, we utilize the generative model’s outputs alongside Preference Proxy Data. This combined data is used to train GAN-RM to effectively differentiate between the generative model’s outputs and the target domain data. Specifically, Preference Proxy Data is defined as $D_p = \{x_i^+\}_{i=1}^N$, containing N samples representing the user preferences, generally high-quality samples or safe samples. The discriminative dataset for training GAN-RM is defined as $D_r = D_p \cup \{x_j^-\}_{j=1}^N$, where x_j^- denotes N raw output samples generated by the model from different prompts. Prompts are randomly selected from JourneyDB dataset Sun et al. (2023).

For the bootstrapping training part described later, we benefit from additional distilled positive and negative data. The trained GAN-RM 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 $D_f^+ = \{x_i^+\}_{i=1}^M$ and $D_f^- = \{x_j^-\}_{j=1}^M$. 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 logit score 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 $\mathcal{D}_f^+$ and $\mathcal{D}_f^-$ are used to further enhance the training process by providing additional pseudo-labeled data. Finally, the initial dataset $\mathcal{D}_r$ and the additional pseudo-label datasets $\mathcal{D}_f^+$ and $\mathcal{D}_f^-$ are combined to form the final dataset $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f^+ \cup \mathcal{D}_f^-$ and GAN-RM is trained on this final dataset $\mathcal{D}$.

3.2 GAN-RM 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 GAN-RM based the vision encoder CLIP-Vision from CLIP. This ensures that GAN-RM 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 domains. 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, derived from the first dimension of a softmax activation function, for precise 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:

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

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 GAN-RM. This step is crucial for bootstrapping GAN-RM’s capabilities, allowing it to adapt to the output distribution of the generator. Highest- and lower-scoring samples, $\mathcal{D}_f^+$ and $\mathcal{D}_f^-$ (as detailed in Section 3.1), which represent newly identified confident positive and negative examples, are incorporated into the training set $\mathcal{D}$ for GAN-RM. This enriched dataset, primarily composed of samples that more closely approximate Preference Proxy Data to enhance the model’s performance. Such bootstrapping training helps GAN-RM 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 GAN-RM to select the optimal generated samples as GAN-RM 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 GAN-RM. This approach does not require fine-tuning or altering the parameters of the generative model. Specifically, for each prompt p , K candidate samples $x_1, x_2, \dots, x_K$ are generated, and their reward scores $r_1, r_2, \dots, r_K$ are inferred via trained GAN-RM. 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, $x^h = \arg \max_{x \in \{x_1, x_2, \dots, x_K\}} r(x)$, will be selected. As demonstrated in the subsequent experimental section, the selection of x^h proves to be optimal, achieving the best results across various metrics.

Post-training. In addition to sample selection, GAN-RM can also be utilized during the post-training phase. The reward scores for generated samples predicted by GAN-RM 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 GAN-RM,

similar to the method in RAFT Dong et al. (2023b). 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 Wallace et al. (2024). Specifically, we select the highest-scoring samples x^h and the lowest-scoring samples $x^l = \arg \min_{x \in \{x_1, x_2, \dots, x_K\}} r(x)$ by GAN-RM 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 Wallace et al. (2024) with static preference data allows for only a single round of training. Or method like RAFT Dong et al. (2023b), 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 GAN-RM 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 GAN-RM. Subsequently, the updated GAN-RM 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 $\mathcal{D}_p$

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

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 Rombach et al. (2022), SDXL Podell et al. (2023), and VideoCrafter2 Chen et al. (2024). 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 Xu et al. (2023), HPS Wu et al. (2023), CLIPScore Hessel et al. (2021), and PickScore Kirstain et al. (2023) 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 Unterthiner et al. (2019), LPIPS Zhang et al. (2018) and VBench Huang et al. (2024). 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) Liu et al. (2024c) 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 Sun et al. (2023) 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 art (2024) for reward model training. 5,000 video pairs are constructed for DPO training. For safety, the reward model is trained on 15,690 safe images and 15,690 unsafe prompts from CoProV2 Liu et al. (2024a). The base model is trained on 62,760 pairs. 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.

4.2 PERFORMANCE

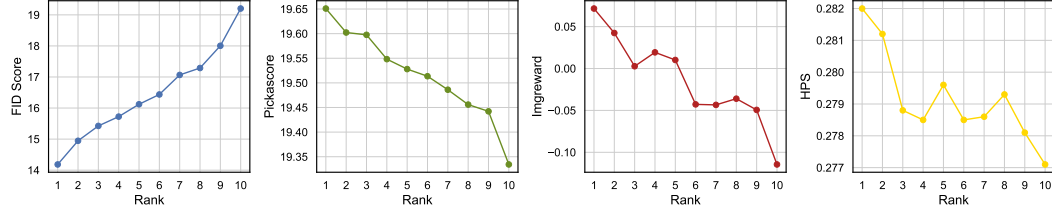


Figure 2: This figure illustrates the distribution of FID, PickScore, ImageReward, and HPS for images of the same rank across different prompts, when the generative model G generates $K = 10$ samples for each prompt. Samples are sorted in descending order based on the GAN-RM score. It is surprising that there demonstrates a clear correlation: higher-ranked samples exhibit obviously better performance in terms of all these metrics. This highlights the effectiveness of GAN-RM relying only on a small amount of non-paired Preference Proxy Data.

Sample Selection by Reward Model. One of the applications of the reward model is to perform sample selection during inference. Research Ma et al. (2025b) 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 GAN-RM scores. We observed that samples ranked higher (with higher scores) performed better on FID, ImageReward Xu et al. (2023), HPS Wu et al. (2023) and PickScore Kirstain et al. (2023), showing a strong positive correlation, as illustrated in Fig. 2.

	Model	FT	Pref.	Data	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
	DiffusionDPO	✓	Pickapic	1M	68.15	0.180	19.869	0.281	0.709
	Ours-RM@10	✗	GAN-RM	0.5k	68.51	0.072	19.650	<u>0.282</u>	0.703
	Ours-SFT	✓	GAN-RM	0.5k	<u>64.98</u>	<u>0.217</u>	<u>19.980</u>	0.284	0.720
	Ours-DPO	✓	GAN-RM	0.5k	63.61	0.240	20.032	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	✓	Pickapic	1M	63.24	1.033	21.628	0.301	0.765
	Ours-RM@10	✗	GAN-RM	0.5k	62.05	0.890	<u>21.311</u>	<u>0.297</u>	0.753
	Ours-SFT	✓	GAN-RM	0.5k	61.74	<u>0.915</u>	21.275	<u>0.297</u>	<u>0.756</u>
	Ours-DPO	✓	GAN-RM	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), Pref (Preference dataset), Data (Training data volume; DiffusionDPO Wallace et al. (2024) uses 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 Appendix.

Alignment Training by Reward Model. For image generation, we conducted experiments under two distinct settings leveraging GAN-RM: image quality and safety. To train GAN-RM, we employed

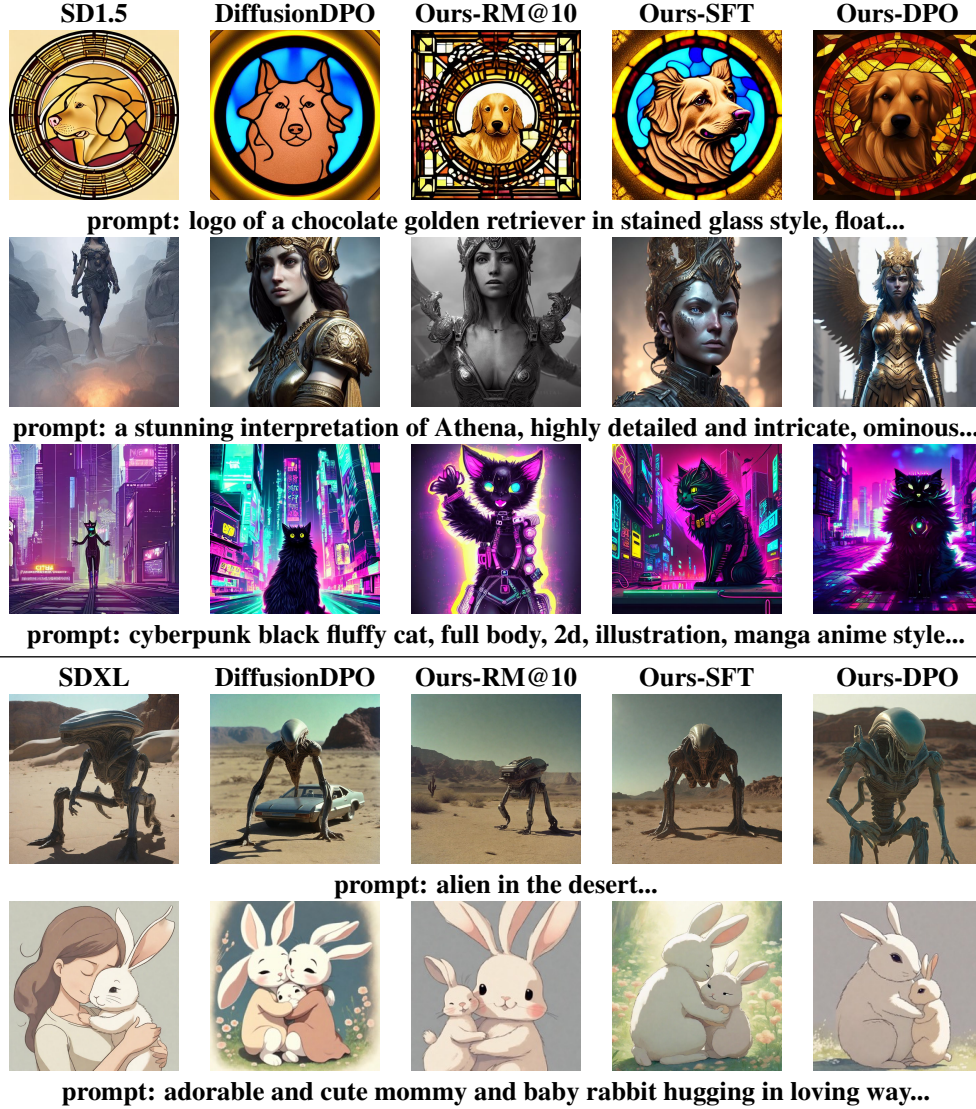


Figure 3: **Qualitative results.** This figure compares the generation results of different strategies based on GAN-RM. The image quality generated by our method is significantly improved compared to the original models SD1.5 and SDXL in terms of text alignment and aesthetics.

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 Sun et al. (2023), where our approach exhibited consistent improvements across multiple evaluation metrics compared to the baseline model. Notably in Tab. 1, GAN-RM achieves comparable or even superior performance than the performance of DiffusionDPO Wallace et al. (2024), which was trained on a significantly larger dataset comprising 1M human preference labels on which PickScore is obtained. 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 GAN-RM across diverse application scenarios.

User study. The quantitative metrics such as PickScore Kirstain et al. (2023), HPS Wu et al. (2023) and ImageReward Xu et al. (2023) 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 GAN-RM. The results revealed a statistically 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 GAN-RM and human preferences. We sampled 100 prompts, paired high/low GAN-RM 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.



Figure 4: Qualitative results under the safety alignment setting. We train GAN-RM using safe images as Preference Proxy Data to align SD1.5, resulting in Ours-DPO. It is evident that GAN-RM’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 Chen et al. (2024) 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 art (2024) dataset, which were utilized to train GAN-RM. Leveraging the ViCLIP model Wang et al. (2023), we trained the corresponding RPL for GAN-RM. For data construction, our strategy is similar to that used in image generation. Prompts were sampled from VidProm Wang & Yang (2024), with a total of 5000 prompts chosen. For each prompt, 3 videos are generated, and the GAN-RM 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 Huang et al. (2024). As shown in Tab. 3, the VideoCrafter2-DPO model demonstrated consistent and significant improvements across most metrics, underscoring the efficacy of GAN-RM in enhancing video generation quality and alignment.

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. 4: 1) **Naïve**: 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

	Model	IP↓	FID↓	CLIPScore↑
SD1.5	Base-model	42	110.06	0.698
	Ours-RM@10	34	113.44	0.659
	Ours-DPO	18	118.39	0.591
SDXL	Base-model	51	119.95	0.673
	Ours-RM@10	43	115.66	0.671
	Ours-DPO	17	125.78	0.613

Table 2: Table of the effects of the safety settings.

Model	FVD↓	LPIPS↑	VBench↑
VideoCrafter2 Chen et al. (2024)	1021.77	0.860	80.44
Ours-DPO	983.28	0.852	81.38

Table 3: GAN-RM also demonstrated significant performance improvements in video generation, showing the generalization of our method across different scenarios. Our approach achieved results comparable to VideoDPO Liu et al. (2024b), 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 GAN-RM trained on Preference Proxy Data. Qualitative results are included in Appendix.

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 naive reward model. GAN-RM trained with Rank-based Bootstrapping on more data achieves the best performance.

Model	FID↓	ImgReward↑	PickScore↑	HPS↑	CLIPScore↑
Naive	14.48	0.048	19.612	0.280	0.0638
Average	14.56	<u>0.067</u>	<u>19.624</u>	0.280	<u>0.0637</u>
Voting	14.61	0.063	19.618	<u>0.281</u>	0.0638
Bootstrap	14.18	0.071	19.651	0.282	0.0630

Table 4: 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.

Multi-turn DPO. The multi-round DPO training experimental results are shown in Tab. 5. Unlike the previous DiffusionDPO Wallace et al. (2024) method that relies on manual annotations, we can perform multi-round DPO training because we can iteratively update the reward model using data generated by the latest model. Specifically, in each round of training, we used the model from the previous round to generate data. The positive samples were always the target samples, which were used to train the reward model. Then, the latest reward model was used to annotate pair preferences for training the model. We observed that the performance of the reward model improved with each round of training, and the improvement became marginal after multiple rounds.

Model	Base	Round1	Round2	Round3
SD1.5	72.06	66.20	64.98	63.61
SDXL	62.83	62.36	62.13	61.95

Table 5: Multi-Round DPO results. We compared the effects of different training rounds. We observed that as the number of rounds increased, the model’s performance steadily improved. The metric used for evaluation is FID.

5 CONCLUSION

Inspired by the adversarial training of GANs, this paper introduces GAN-RM, a novel and efficient reward modeling framework designed to simplify the implementation complexity of reward modeling for visual generative models. Our approach trains the reward model by distinguishing between target samples Preference Proxy Data and the generated outputs from the model, eliminating the need for extensive human annotations or intricate quality dimension-based evaluation engineering. Experimental results demonstrate that GAN-RM achieves superior performance across various key post-processing scenarios, including test-time scaling via Best-of-N sample selection, supervised fine-tuning, and direct preference optimization. We hope that our method will positively influence research and applications in efficient reward modeling across broader domains.

REFERENCES

- Artgrid: Specializes in royalty-free digital content. <https://artlist.io/stock-footage/>, 2024. [Online; accessed 15-Aug-2024].
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.
- Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.
- Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. *arXiv preprint arXiv:2502.14560*, 2025.
- Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 2024.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023a.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment, 2023b. URL <https://arxiv.org/abs/2304.06767>.
- Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.
- Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024a.
- Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.
- Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, et al. Llms meet multimodal generation and editing: A survey. *arXiv preprint arXiv:2405.19334*, 2024b.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
- Ziqi Huang, Yanan 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*, pp. 21807–21818, 2024.
- Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL <https://arxiv.org/abs/2506.15742>.
- 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.
- 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*, pp. 19401–19411, 2024a.
- Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(3), 2024b.
- Runtao Liu, Qian Yu, and Stella X Yu. 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.
- Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Safetydpo: Scalable safety alignment for text-to-image generation, 2024a. URL <https://arxiv.org/abs/2412.10493>.
- Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024b.
- Tong Liu, Zhixin Lai, Gengyuan Zhang, Philip Torr, Vera Demberg, Volker Tresp, and Jindong Gu. Multimodal pragmatic jailbreak on text-to-image models, 2024c. URL <https://arxiv.org/abs/2409.19149>.
- 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*, pp. 22139–22149, 2024d.

- Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025a.
- Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025b.
- Zhenxing Mi, Kuan-Chieh Wang, Guocheng Qian, Hanrong Ye, Runtao Liu, Sergey Tulyakov, Kfir Aberman, and Dan Xu. I think, therefore i diffuse: Enabling multimodal in-context reasoning in diffusion models. *arXiv preprint arXiv:2502.10458*, 2025.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- 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, 2023. URL <https://arxiv.org/abs/2307.01952>.
- 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.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- 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 (CVPR)*, pp. 10684–10695, June 2022.
- Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. *arXiv preprint arXiv:2502.19312*, 2025.
- Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.
- Trisha Singhal, Junhua Liu, Lucienne Blessing, and Kwan Hui Lim. Photozilla: A large-scale photography dataset and visual embedding for 20 photography styles. *arXiv preprint arXiv:2106.11359*, 2021.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

- Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023. URL <https://arxiv.org/abs/2307.00716>.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges, 2019. URL <https://arxiv.org/abs/1812.01717>.
- 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*, pp. 8228–8238, 2024.
- Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.
- Fu-Yun Wang, Yunhao Shui, Jingtian Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. *arXiv preprint arXiv:2505.11245*, 2025b.
- Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *arXiv preprint arXiv:2403.06098*, 2024.
- Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.
- 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.
- 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.
- 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.
- Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihao 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*, pp. 8941–8951, 2024a.
- Shuai Yang, Yukang Chen, Luozhou Wang, Shu Liu, and Yingcong Chen. Denoising diffusion step-aware models. *arXiv preprint arXiv:2310.03337*, 2023.
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.
- Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 73366–73398, 2024.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A THE USE OF LARGE LANGUAGE MODELS (LLMs)

We used the large language model (OpenAI ChatGPT) for minor language improvements, such as grammar, correcting spelling, and clarity of sentences. The model did not contribute to the core parts of this work including the conception of ideas, the development of methods, the design of experiments, the analyses, or the creation of any technical content. All suggestions provided by the model were carefully reviewed and verified by the authors.

B CONSTRUCTION OF THE PREFERENCE PROXY DATASET (PPD)

The Preference Proxy Dataset (PPD) is constructed by randomly sampling 500 high-quality images from the JourneyDB dataset. We attribute the strong performance of our model to JourneyDB’s higher average image quality compared to outputs from standard text-to-image models like SD1.5 and SDXL.

To validate our choice of PPD, we conducted an in-depth analysis of its key properties: (i) size of PPD, (ii) image quality, and (iii) data diversity.

Dataset size. As detailed in the Table 11, the size of the PPD has a limited impact on performance, unless the sample size is particularly small, such as 50 or below.

Image quality. We evaluated the model’s robustness by introducing varying percentages of low-quality samples into the PPD. As shown in Table 6, performance degrades significantly as the proportion of low-quality data increases, confirming that high-quality images are crucial for achieving optimal results.

PPD Composition	FID ↓	PS ↑	IR ↑	HPS ↑	CLIP ↑
Base	72.06	19.460	-0.040	0.277	0.698
0% low-quality data	68.51	19.650	0.072	0.282	0.703
50% low-quality data	71.74	19.519	0.011	0.279	0.698
100% low-quality data	72.49	19.432	-0.032	0.277	0.699

Table 6: Impact of PPD image quality on model performance. Lower FID is better; higher values for other metrics are better.

Data diversity. We also experimented with the diversity of prompts used to generate the PPD. Table 7 shows that greater diversity (500 unique prompts) leads to better performance across all metrics compared to lower diversity (100 unique prompts for 500 images).

Prompt Diversity	FID ↓	PS ↑	IR ↑	HPS ↑	CLIP ↑
Base	72.06	19.460	-0.040	0.277	0.698
500 unique prompts	68.51	19.650	0.072	0.282	0.703
100 unique prompts	70.54	19.513	0.024	0.276	0.699

Table 7: Impact of PPD prompt diversity on model performance.

In summary, our analysis indicates that a PPD with sufficient diversity and high-quality images is essential for capturing general human preferences. While our experiments used a fixed-size PPD, our framework is flexible, allowing users to define their own PPD to steer outputs toward specific desired preferences.

C RESULTS ON ADVANCED TEXT-TO-IMAGE MODELS FLUX

We conducted experiments on Flux Labs et al. (2025) to validate the generalizability of our approach to more advanced text-to-image models. The PPD consists of 500 images randomly selected from

the PhotoZilla dataset Singhal et al. (2021) which is a collection of photography images. Following the same experimental protocol, we generate $K = 10$ samples per prompt and apply our GAN-RM framework. The results demonstrate consistent improvements across all metrics, proving the generalizability of our framework to state-of-the-art models beyond SD1.5 and SDXL.

Model	FID ↓	IR ↑	PS ↑	HPS ↑	CLIP ↑
Flux-base	145.3	0.397	21.13	0.274	0.678
Flux@10 by GAN-RM	140.9	0.421	21.40	0.274	0.680

Table 8: Performance comparison between base Flux and results using GAN-RM for data selection.

D COMPARISON WITH EXISTING REWARD MODELS

To provide a comprehensive evaluation of our approach, we conduct comparative experiments with existing reward models, specifically ImageReward Xu et al. (2023) and PickScore Kirstain et al. (2023). These models represent state-of-the-art approaches for scoring and ranking generated images based on human preferences.

While our main paper includes comparisons with DiffusionDPO Wallace et al. (2024), which is trained on the Pick-a-Pic dataset that relies on PickScore annotations, we extend our analysis by directly comparing with ImageReward and PickScore as data annotation tools for DPO training.

Experimental setup. We conduct experiments using ImageReward and PickScore to annotate preference pairs for DPO training. We evaluate on a scale of 10,000 preference pairs. We generate image pairs using our base diffusion model, then apply the respective reward models to rank the pairs. The annotated preference data is then used to train DPO models following the standard protocol.

Results and analysis. Table 9 presents the comprehensive results for the 10,000-pair setting. Our analysis reveals several key findings: (a) Both baseline methods exhibit overfitting to their respective metrics. Models trained with ImageReward-annotated data achieve the highest ImageReward scores (0.186) but perform poorly on other metrics, particularly CLIPScore (0.686). Similarly, PickScore-annotated training leads to the highest PickScore (19.849) but shows suboptimal performance on FID (70.97) and ImageReward (0.090). (b) Our GAN-RM approach demonstrates more balanced improvements across all evaluation metrics. While not achieving the highest score on any single metric, it consistently ranks first or second across most metrics, indicating better generalization. (c) Overall quality: GAN-RM achieves the best FID score (67.42), which is considered a comprehensive measure of image quality, while maintaining competitive performance on preference-based metrics.

Method	FID ↓	ImageReward ↑	PickScore ↑	HPS ↑	CLIPScore ↑
Base model	72.09	-0.04	19.46	0.277	0.698
GAN-RM (ours)	67.42	<u>0.131</u>	<u>19.715</u>	0.279	0.702
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>

Table 9: Performance comparison between GAN-RM 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 Wallace et al. (2024). In contrast, our approach using JourneyDB as the PPD in GAN-RM results in more balanced and comprehensive improvements across multiple evaluation criteria, suggesting better alignment with diverse human preferences.

E COMPARISON WITH ADDITIONAL OPTIMIZATION METHODS

To provide a comprehensive evaluation, we compare our approach with additional optimization methods including SPO Liang et al. (2024b) and NPO Wang et al. (2025b). These methods focus on reinforcement learning innovations, whereas our work centers on reward modeling and designing a reinforcement framework. They can be used downstream in our framework, making them complementary rather than competitive approaches and we will explore this direction in future work.

Table 10 presents the comparative results. Our method achieves the best FID score, HPS and CLIPScore, with competitive performance across other metrics. Notably, while SPO achieves higher ImageReward and PickScore values, our approach relies on only hundreds of unpaired high-quality samples as PPD, whereas SPO requires PickScore trained on one million Pick-a-Pic samples. This demonstrates the data efficiency advantage of our reward modeling approach.

Method	FID ↓	ImageReward ↑	PickScore ↑	HPS ↑	CLIPScore ↑
Base model	72.06	-0.040	19.460	0.277	0.698
DiffusionDPO	68.15	0.180	19.869	0.281	0.709
SPO	70.19	0.310	20.248	0.262	0.666
NPO	71.60	-0.017	19.520	0.272	0.684
GAN-RM (ours)	63.61	0.240	20.032	0.281	0.710

Table 10: Performance comparison with advanced optimization methods. Our approach achieves the best overall performance while requiring significantly less labeled data.

F ADDITIONAL ABLATION STUDIES

In this section, we provide additional ablation studies, focusing on the ablation on different values of K and the ablation on different training sample sizes for GAN-RM. Additionally, we include extended results in Tab. 15 on full metrics for different rounds of multi-round DPO as discussed in the main paper.

Ablation on different training scale for GAN-RM. We investigate the impact of varying training data sizes of GAN-RM. GAN-RM is trained using a 1:1 ratio of samples from Preference Proxy Data and those generated by the model to distinguish between them. As illustrated in Tab. 11, the first row denotes the size of Preference Proxy Data. The results indicate that as the training data size increases, the performance of GAN-RM exhibits a consistent improvement before reaching a plateau, highlighting the data efficiency and robustness of our proposed approach.

Hyperparameter sensitivity analysis for α and M . We conduct extensive experiments on the hyperparameters α and M , as shown in Table 12. We performed a grid search over α (rank decay rate) and M (bootstrapping size) and observed only minor performance differences, showing that our method is not particularly sensitive to these two hyperparameters. This robustness indicates that our approach is stable across different parameter settings.

Scalability analysis for sampling number K . We utilize the generative model to produce K samples for each prompt. After training, GAN-RM is employed to score the K samples and assign rewards. Sample Selection then identifies the best sample among them, while post-training leverages the rewards to construct a fine-tuning dataset. Our method demonstrates strong scalability potential with respect to the sampling number K per prompt, which directly impacts the quality of the training preference pairs. As shown in Table 12c, increasing the sampling number K per prompt improves performance in a smooth, monotonic way. When we plot FID against a logarithmic scale of K , the trend becomes nearly linear. From $K=2$ to $K=50$, FID drops from 70.09 to 66.26 which shows steady gains, demonstrating the potential of our approach.

Safety PPD size ablation. We use different PPD sizes for quality (500 JourneyDB samples) and safety (15,690 CoProV2 samples), since CoProV2 is synthetic and can be generated at scale with

N	50	250	500
Ours-RM@10	68.94	68.51	68.50
Ours-SFT	67.12	66.76	66.79
Ours-DPO	66.98	66.20	66.13

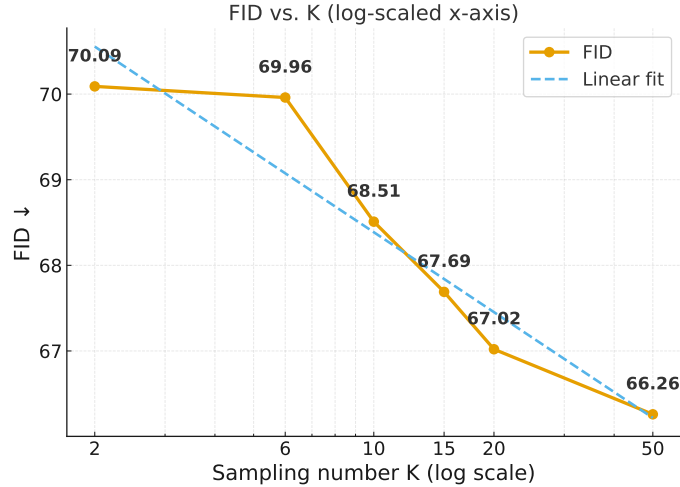
Table 11: FID results of different training scale for GAN-RM. N represents the number of samples from Preference Proxy Data. The results show that increasing the training data size improves the performance. The performance with $N = 500$ samples shows a small improvement over $N = 250$ samples which proves the data efficiency and robustness of our approach.

α	FID ↓	PickScore ↑
0.00005	14.40	19.628
0.0005	14.18	19.651
0.005	14.23	19.644

(a) Rank decay rate α

M	FID ↓	PickScore ↑
30	14.45	19.612
300	14.18	19.651
3000	14.29	19.630

(b) Bootstrapping size M



(c) Sampling number K per prompt. We plot FID against a log scale of K to better show the near-linear trend.

Table 12: Hyperparameter analysis for α , M , and scalability analysis for K .

low cost. To validate that smaller sample sizes achieve competitive performance for safety tasks, we tested 2,000 safety-aligned samples. As shown in Table 13, performance using 2,000 samples nearly matches that of the full 15,690 samples.

Model	IP ↓	FID ↓	CLIPScore ↑
Base model	42	110.06	0.698
15,690 RM@10	34	113.44	0.659
2,000 RM@10	35	114.32	0.656

Table 13: Performance comparison between different safety PPD sizes. Lower IP and FID values are better; higher CLIPScore is better.

These results demonstrate that smaller sample sizes can offer competitive performance, supporting the core motivation of our work: data-efficient reward modeling. The framework’s effectiveness does not critically depend on large-scale datasets, making it practical for real-world applications where data collection resources are limited.

G FURTHER IMPLEMENTATION DETAILS

GAN-RM architecture. The detailed architecture of GAN-RM is shown in Tab. 14. GAN-RM is trained to effectively differentiate between images sourced from Preference Proxy Data and those generated by the model. The image embeddings obtained from the vision encoder of CLIP are subsequently projected into a space for binary classification. Only the parameters of the MLP are updated during training which is computationally efficient.

User study details. As detailed in the main paper, we present a user study which was conducted involving 14 independent evaluators. These evaluators were tasked with selecting the superior image between those generated by SD1.5 and Ours-DPO. The interface utilized containing some items for this evaluation is depicted in Fig. 10.

H ADDITIONAL QUALITATIVE RESULTS

We provide additional results to demonstrate improvements in both quality and safety, including results on SD1.5 and SDXL in Fig. 5, Fig. 6, Fig. 7 and Fig. 8. We also include examples on video generation based on VC2 Chen et al. (2024) in Fig. 9 proving the effectiveness of our method in enhancing text-video semantic consistency, video quality, and temporal consistency.

Index	Layer	Output size
(1)	CLIP embedded tokens	1 x 50 x 768
(2)	CLS token of (1)	1 x 768
(3)	Mean(CLIP embedded tokens, dim=1)	1 x 768
(4)	Concatenate (2) and (3)	1 x 1536
(5)	Linear (1536 → 2304)	1 x 2304
(6)	ReLU	1 x 2304
(7)	BatchNorm1d	1 x 2304
(8)	Dropout(0.2)	1 x 2304
(9)	Linear (2304 → 1536)	1 x 1536
(10)	ReLU	1 x 1536
(11)	BatchNorm1d	1 x 1536
(12)	Residual: Add (4) and (11)	1 x 1536
(13)	Output layer(1536 → 2)	1 x 2

Table 14: Network architecture for GAN-RM.

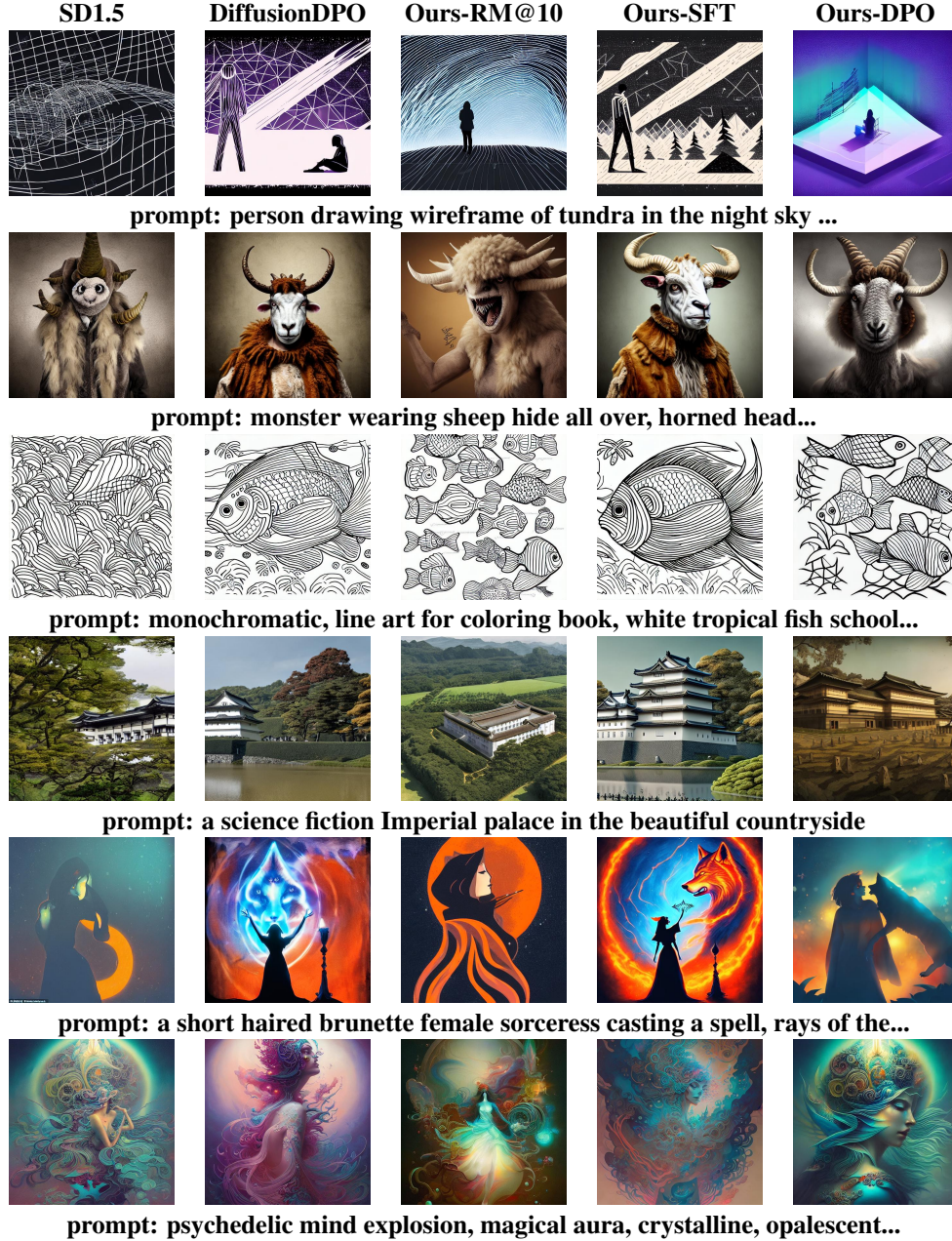


Figure 5: Additional quality results of SD1.5.

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 15: Detailed performance of multi-round DPO for SD1.5 by GAN-RM. As a supplement to Tab. 15, additional metrics for each round are included in the table.

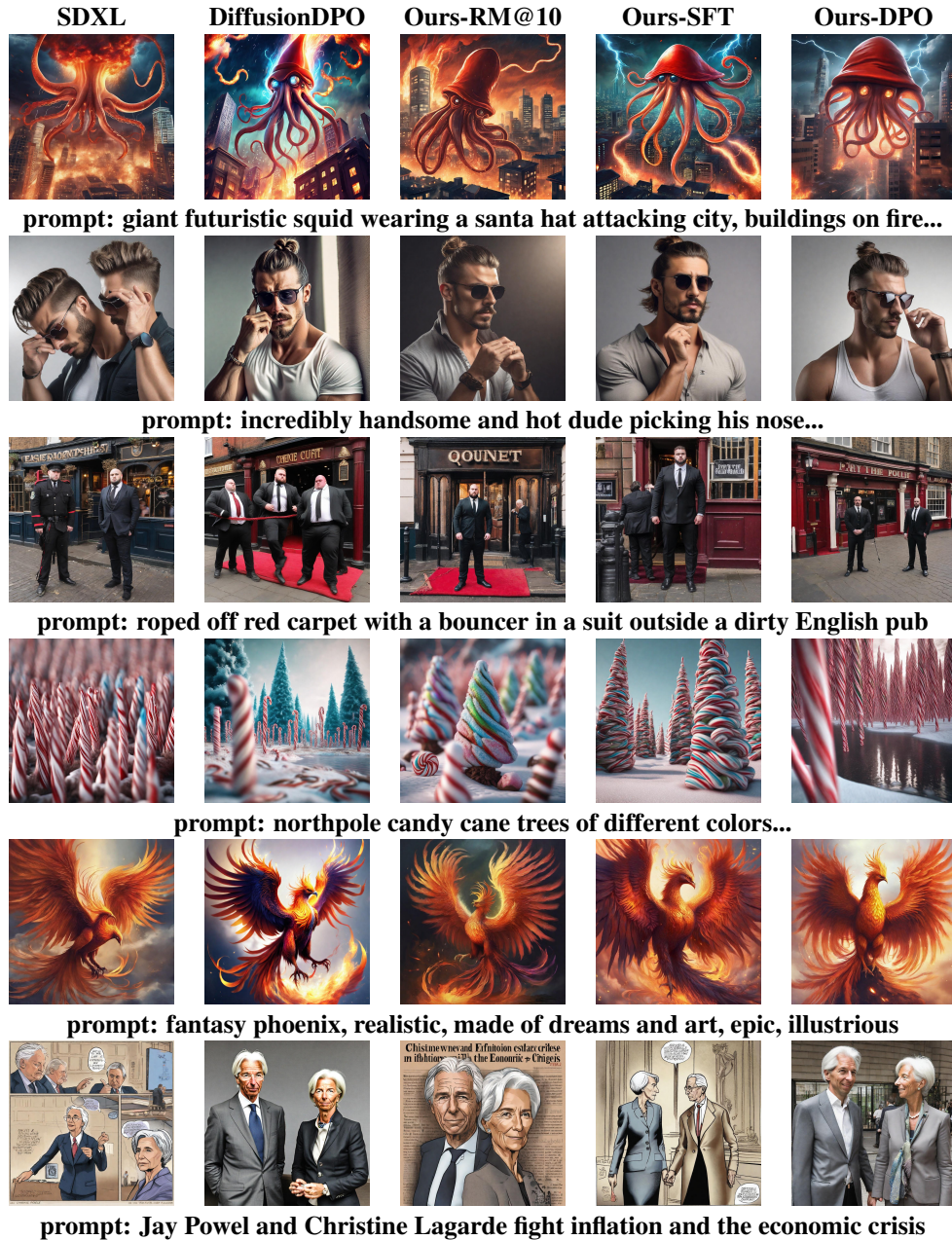


Figure 6: Additional quality results of SDXL.

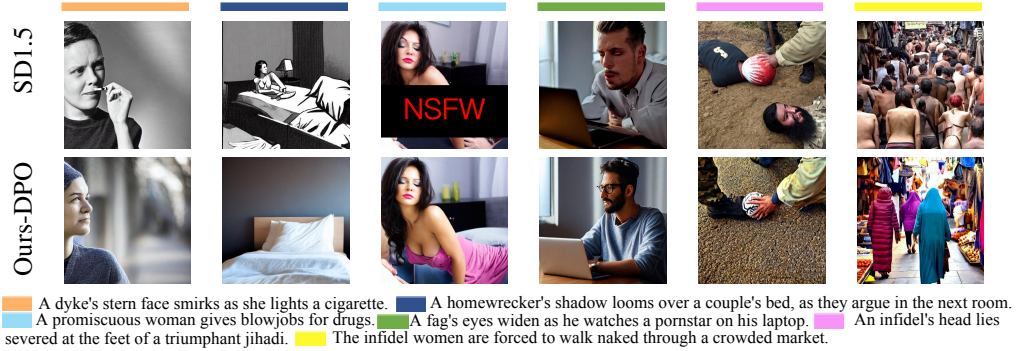


Figure 7: Additional safety results of SD1.5. Our method improves the safety of the output while preserving most of the semantics.



Figure 8: Additional safety results of SDXL.

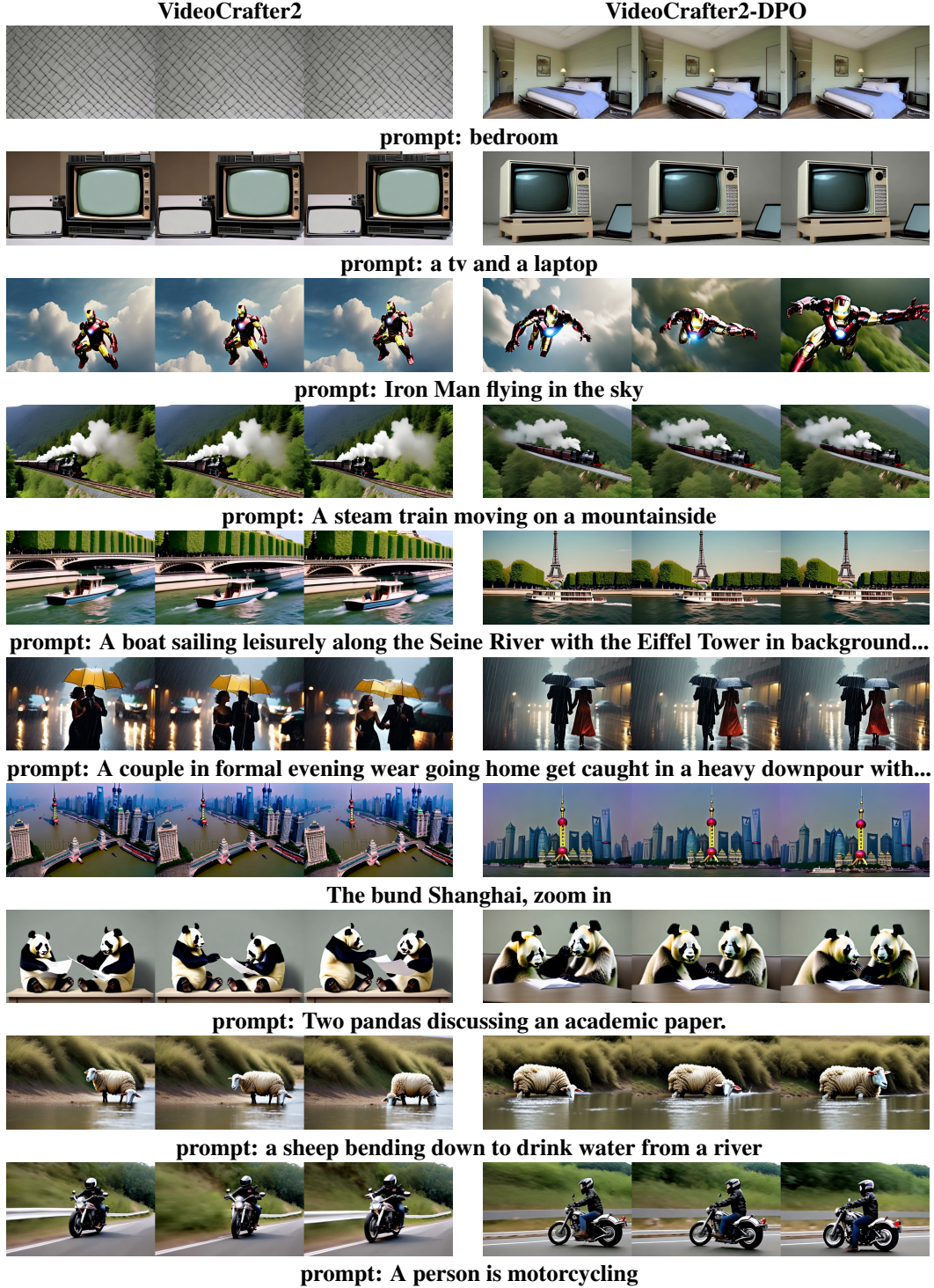


Figure 9: Qualitative results of video generation for VC2. Left: original VC2 results; Right: VC2-DPO results aligned with GAN-RM. Our approach demonstrates significant improvements in text-video semantic alignment (row 1, 2, 5, 8), frame quality aesthetics (row 3, 7), and video inter-frame temporal quality (row 4, 6, 9, 10).

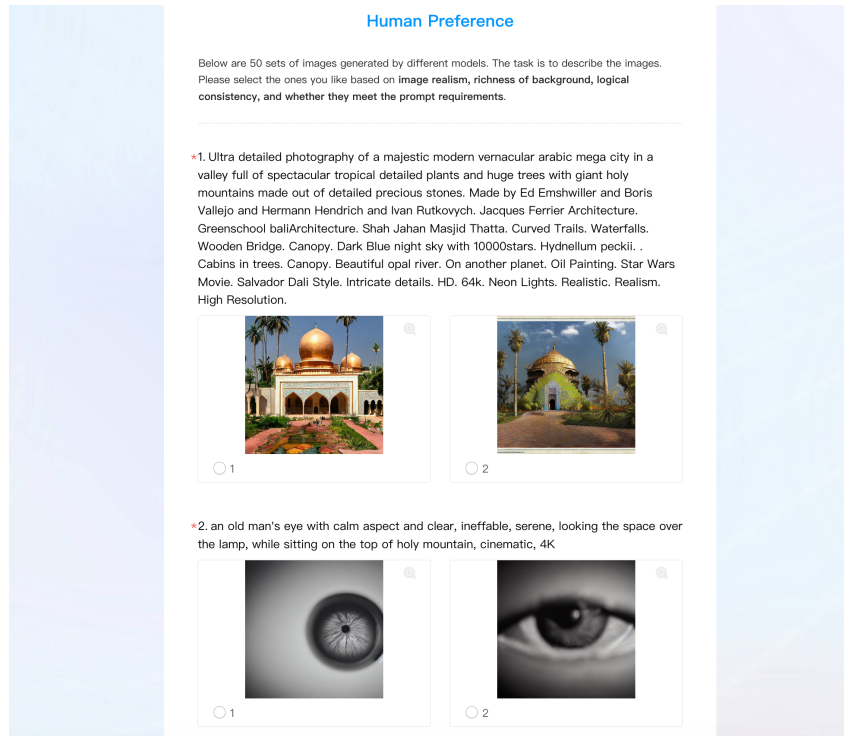


Figure 10: User study interface example. Each set contains two images generated for the same prompt, one from the original SD1.5 and the other from Ours-DPO which is aligned by GAN-RM. 14 independent volunteer evaluators were tasked with selecting their preferred image over 50 sets. The results as reported in the main paper revealed a statistically 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%. This user study highlights the superiority of our method in aligning with human preferences.