VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation

Jiazheng Xu[1*†], Yu Huang[1*†], Jiale Cheng[1†], Yuanming Yang[1†], Jiajun Xu[1†], Yuan Wang[1†], Wenbo Duan[1†], Shen Yang[1†], Qunlin Jin[1†], Shurun Li[1†], Jiayan Teng[1†], Zhuoyi Yang[1†], Wendi Zheng[1†], Xiao Liu[1†], Dan Zhang[1†], Ming Ding[2], Xiaohan Zhang[2], Shiyu Huang[2], Xiaotao Gu[2], Minlie Huang[1] , Jie Tang[1] , Yuxiao Dong[1]

[1]Tsinghua University
[2]Z.AI
{xjz22, h-y22}@mails.tsinghua.edu.cn, yuxiaod@tsinghua.edu.cn

Abstract

Visual generative models have achieved remarkable progress in synthesizing photorealistic images and videos, yet aligning their outputs with human preferences across critical dimensions remains a persistent challenge. Though reinforcement learning from human feedback offers promise for preference alignment, existing reward models for visual generation face limitations, including black-box scoring without interpretability and potentially resultant unexpected biases. We present VisionReward, a general framework for learning human visual preferences in both image and video generation. Specifically, we employ a hierarchical visual assessment framework to capture fine-grained human preferences, and leverages linear weighting to enable interpretable preference learning. Furthermore, we propose a multi-dimensional consistent strategy when using VisionReward as a reward model during preference optimization for visual generation. Experiments show that VisionReward can significantly outperform existing image and video reward models on both machine metrics and human evaluation. Notably, VisionReward surpasses VideoScore by 17.2% in preference prediction accuracy, and text-to-video models with VisionReward achieve a 31.6% higher pairwise win rate compared to the same models using VideoScore.

Code — https://github.com/THUDM/VisionReward

1 Introduction

Visual generative models, including text-to-image (Ding et al. 2021; Ramesh et al. 2021; Saharia et al. 2022; Rombach et al. 2022; Betker et al. 2023; Podell et al. 2023) and text-to-video (Hong et al. 2022; Ho et al. 2022; Villegas et al. 2022; Zheng et al. 2024; Chen et al. 2024a; Yang et al. 2024) generation, have recently experienced rapid developments. Through large-scale pretraining, these models can effectively translate textual descriptions into photorealistic images or temporally coherent videos. To further align them with human preferences, reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022)—initially introduced in large language models—has recently been adapted to visual generation tasks (Xu et al. 2023; He et al. 2024).

A key bottleneck in applying RLHF to visual generation lies in developing effective visual reward models. Recent studies (Xu et al. 2023; Kirstain et al. 2023; Wu et al. 2023) have explored training reward models to predict human visual preferences, enabling automatic evaluation and preference optimization for visual generative models. For evaluation, reward models function as automated metrics that quantitatively measure the alignment between generated outputs and human preference criteria (Li et al. 2023). For optimization, reward models identify reliable directions for improving visual generation models. Essentially, they can provide feedback in reinforcement learning or generate preference pairs, thus reducing dependence on human annotation (Black et al. 2023; Fan et al. 2023; Clark et al. 2023).

Despite recent progress in reward models (RMs) for visual generation, two primary challenges remain: First, *lack of interpretability and risk of unexpected bias.* Current RMs for visual generation often suffer from limited interpretability. These models inherently involve complex trade-offs among multiple factors, yet their scoring mechanisms lack transparency regarding how such trade-offs are performed. This opacity raises concerns about *potential unexpected biases*. Though multimodal LLMs like Gemini (Team et al. 2024) and GPT-4o (Achiam et al. 2023) enhance interpretability through explainable rating rationales, their general-purpose architectures usually underperform specialized black-box models in fine-grained assessments (Chen et al. 2024c). This raises a key dilemma: how to design preference prediction method to be interpretable while maintaining accuracy.

[*]Equal contributions. Core contributors: Jiazheng, Yu, Jiale, Yuanming, Jiajun, Yuan, Wenbo, Shen and Qunlin. Corresponding author: Yuxiao (yuxiaod@tsinghua.edu.cn)

[†]Work done while these authors interned at Z.AI.

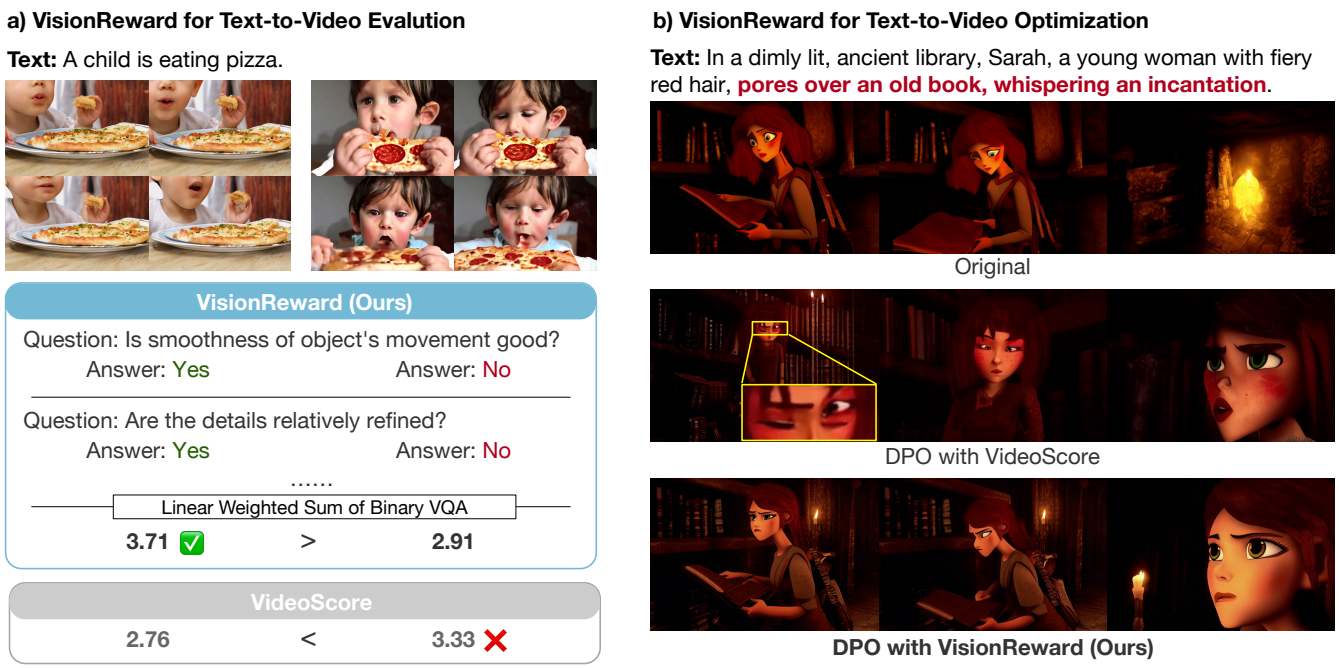

Figure 1: Illustration of how VisionReward works for evaluation and optimization of visual generation. a) **Evaluation**: VisionReward performs comprehensive evaluation through dimension-specific binary visual QA testing, producing human-aligned, fine-grained assessment scores. b) **Optimization**: VisionReward enable better preference optimization, enhancing multiple key aspects.

Second, *lack of effective reward models for video generation.* The rapid development of text-to-video generative models has intensified the demand for video reward models. Although image reward models can assess individual frame quality, their frame-level nature inherently neglects essential temporal dependencies in video sequences. While VideoScore (He et al. 2024) has pioneered direct video evaluation through learnable metrics, it still suffers from limitations such as insufficient accuracy in preference prediction and optimization in video generation.

Contribution. To address these challenges, we propose a general framework VisionReward to build accurate reward models for both image and video generation. VisionReward is trained with two steps: *fine-grained visual assessment* and *interpretable preference learning*. First, to capture human visual preferences, we identify nine major dimensions and decouple preferences into 64 fine-grained questions. Second, to ensure interpretable preference learning, we propose to use the classical linear weighting mechanism on the question outcomes. It enables intuitive visualization of each question's impact.

To apply it as a reward model for visual generation models, we propose *a multi-dimensional consistent strategy* during preference optimization. The goal of this strategy is to mitigate unintended and unquantifiable biases. Specifically, a pair of visual samples is used for preference optimization (e.g., DPO (Wallace et al. 2024)) only if one sample is consistently preferred over the other across all dimensions.

To summarize, we present VisionReward as a general framework for visual preference learning. Empirically, we show that VisionReward makes the following contributions:

- We design fine-grained multi-dimensional preference annotation and build the most fine-grained dataset which contain 81K samples and 5M binary annotation, which enable the training of VisionReward.
- For visual preference prediction, VisionReward achieves state-of-the-art performance across multiple benchmarks, while maintaining interpretability via hierarchical diagnostic QA and explicit linear weighting. For instance, VisionReward outperforms VideoScore (He et al. 2024) by 17.2% in accuracy on preference prediction.
- For visual generation, VisionReward can serve as an effective reward model for preference optimization (e.g., DPO), significantly enhancing the text-to-image and text-to-video models. For example, video generation models with VisionReward achieve a 31.6% higher pairwise win rate compared to the same models using VideoScore.

2 Related Work

Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al. 2020; Nakano et al. 2021; Ouyang et al. 2022) refers to optimizing models with reinforcement learning based on human feedback, which is also explored in image and video generation.

Preference Learning for Visual Generation. There are many works learning from human preferences, which collect human annotation for text-to-image (Xu et al. 2023; Kirstain

Dataset	Image	Video	#Samples	#Dimensions	#Fine-Grained	#Annotation
ImageReward (Xu et al. 2023)	✓		9K	-	-	0.1M
Pick-a-Pic (Kirstain et al. 2023)	✓		38K	-	-	0.6M
HPDv2 (Wu et al. 2023)	✓		430K	-	-	0.8M
RichHF-18K (Liang et al. 2024)	✓		18K	4	-	0.1M
MPS (Zhang et al. 2024b)	✓		608K	4	-	2.4M
VideoScore (He et al. 2024)		✓	38K	5	-	0.2M
VisionReward (Ours)	✓	✓	81K	**18–20**	**61–64**	**5.0M**

Table 1: Comparison of dataset of VisionReward and other datasets.

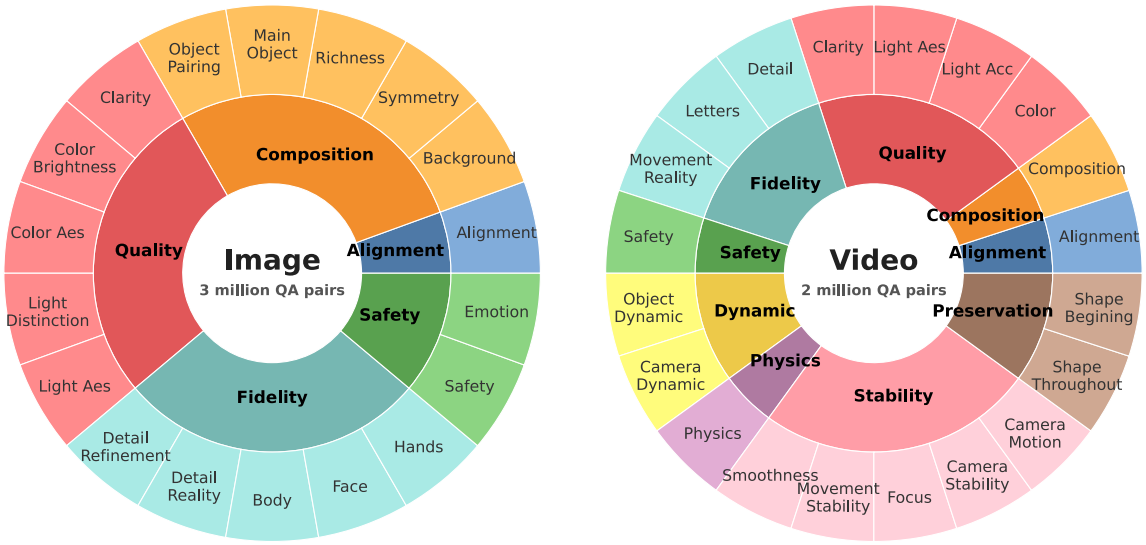

Figure 2: Illustration of fine-grained multi-dimensional design. (Left) For image: 5 dimensions, 18 sub-dimensions, and 61 binary questions. (Right) For video: 9 dimensions, 20 sub-dimensions, and 64 binary questions.

et al. 2023; Wu et al. 2023) and text-to-video (He et al. 2024). Note that existing approaches (Zhang et al. 2024b; Liang et al. 2024) have attempted to augment human annotations or expand dimensions of human preferences in visual generation. Different from them, VisionReward defines fine-grained multi-dimensional human preferences with the goal of disentangling distinct factors to decouple human preferences, to build more accurate and interpretable RM.

RLHF for Visual Generation. For visual generation tasks, several works have explored RLHF, optimizing from the gradient (Xu et al. 2023; Wu et al. 2024) or using a policy-based RL approach (Black et al. 2023; Fan et al. 2023; Clark et al. 2023). All these methods require a reward model (RM) to provide feedback for online learning. Diffusion-DPO (Wallace et al. 2024) has proposed to optimize the diffusion model directly using human-labeled preference data. However, most RLHF methods face the issue of biased-optimization. By employing a multi-dimensional method, VisionReward achieves robust RLHF.

Preliminary for Diffusion-DPO. Given a data distribution $q(x_0)$, Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2020) contains forward process and reverse process. Forward process $q(x_{1:T}|x_0)$ gradually add noise to the data x_0 and reverse process $p_\theta(x_{0:T})$ learns transitions to recover data. Training diffusion model can be performed by evidence lower bound (Kingma et al. 2021; Song et al. 2021):

$$L_{\mathrm{DM}} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}),t}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,t\right)\|_2^2\right], \quad (1)$$

with $t \sim \mathcal{U}(0,T)$ and $\boldsymbol{x}_t \sim q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right)$.

Diffusion-DPO (Wallace et al. 2024) introduces direct preference optimization based on preference pairs. We denote the "win" and "lose" samples as x_0^w, x_0^l, and the objective is as follows:

$$\begin{aligned}\mathcal{L}(\theta) = &-\mathbb{E}_{t\sim\mathcal{U}(0,T),\boldsymbol{x}_t^w\sim q\left(\boldsymbol{x}_t^w|\boldsymbol{x}_0^w\right),\boldsymbol{x}_t^l\sim q\left(\boldsymbol{x}_t^l|\boldsymbol{x}_0^l\right)}\\ &\log\sigma\left(-\beta T\omega\left(\lambda_t\right)\right(\\ &\|\boldsymbol{\epsilon}^w-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t^w,t\right)\|_2^2-\|\boldsymbol{\epsilon}^w-\boldsymbol{\epsilon}_{\mathrm{ref}}\left(\boldsymbol{x}_t^w,t\right)\|_2^2\\ &-\left(\left\|\boldsymbol{\epsilon}^l-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t^l,t\right)\right\|_2^2-\left\|\boldsymbol{\epsilon}^l-\boldsymbol{\epsilon}_{\mathrm{ref}}\left(\boldsymbol{x}_t^l,t\right)\right\|_2^2\right)\Big)\Big). \quad (2)\end{aligned}$$

DPO is ordinarily based on overall preference, which may be biased. VisionReward enables Multi-Dimensional Preference Optimization to enhance it.

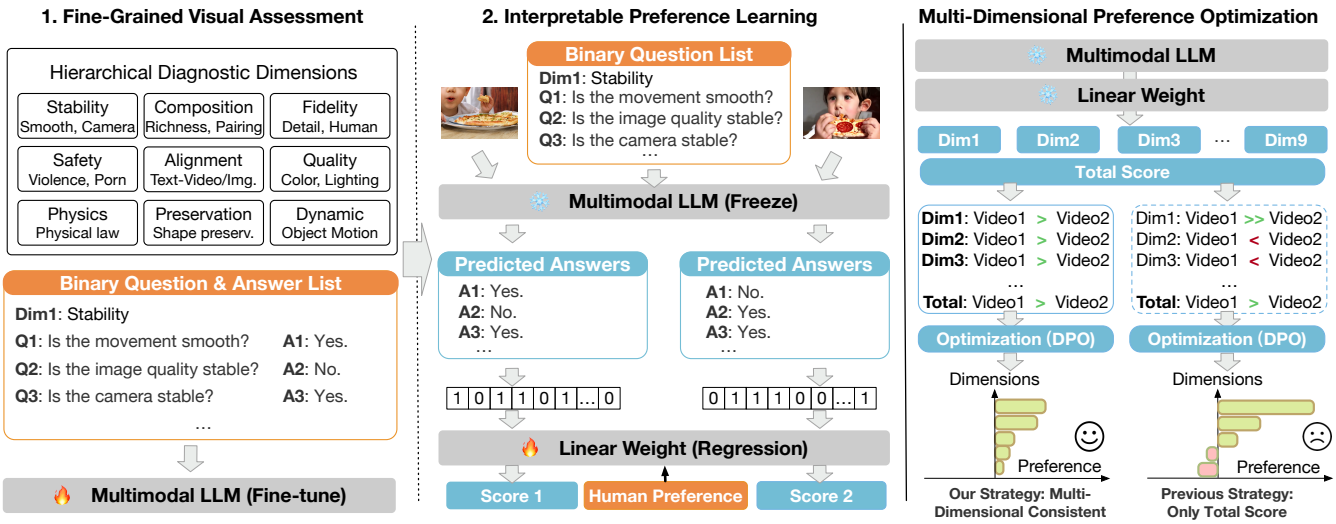

Figure 3: Overall framework of VisionReward. 1) Fine-grained Visual Assessment: fine-tune multimodal LLM to perform binary visual question-answering through hierarchical dimensions. 2) Interpretable Preference Learning: utilize visual QA outputs to predict preferences through linear weighted summation. 3) Multi-Dimensional Preference Optimization: optimization strategy across multiple dimensions.

3 VisionReward

3.1 VisionReward Annotation

Fine-Grained Design. Human preferences are often a result of the interplay of multiple factors (Palmer, Schloss, and Sammartino 2013; Ibarra et al. 2017), necessitating a balance among various considerations. To deconstruct human preferences systematically, we develop a fine-grained multi-dimensional framework, as shown in Table 1 and Fig. 2. For each sub-dimension, we set options that vary gradually in degree, and decompose these options into a series of binary questions (Cf. Tables 15 to 18 in Appendix).

Dataset Preparation. For images, we sample images from multiple popular datasets, including ImageRewardDB (Xu et al. 2023), HPDv2 (Wu et al. 2023), and Pick-a-Pic (Kirstain et al. 2023), and obtain 48k images after filtering. For videos, we sample prompts from VidProM (Wang and Yang 2024). To ensure diversity of prompts, we use Rouge-L (Lin 2004) for initial filtering, follow UniFL (Zhang et al. 2024a) to perform a semantic-based filtering, and use ChatGPT (Achiam et al. 2023) for data cleaning, finally get 10k prompts. Then we use CogVideoX (Yang et al. 2024), VideoCrafter2 (Chen et al. 2024a) and OpenSora (Zheng et al. 2024) to generate 30k videos, sample from Panda-70M (Chen et al. 2024b) to get 3k real videos, leading to 33k videos for annotation. More details are provided in Appendix Section 6.1.

Annotation Management. To avoid bias of annotators, our annotation management includes professional management and standard document. Cooperating with a specialized company, we strictly conduct annotation training for annotators, select qualified annotators, and perform quality inspection of annotation results. Our annotation document gives clear definitions and provides more than 10 examples for each judgment, to align the standard among annotators. Due to these efforts, the consistency of annotators in the binary results reaches **89.29%** (images) and **89.33%** (videos).

Annotation Analysis. Through specialized annotation, we obtain an image dataset containing 48k images and **3 million** question-answer pairs, while a video dataset with 33k videos and **2 million** pairs. More statistical analysis of the annotation results is in Appendix Section 6.2.

3.2 VisionReward Training

The complete training process of VisionReward and its application methodology during preference optimization are illustrated in Fig. 3.

Fine-grained Visual Assessment. Specifically, we use CogVLM2 (Hong et al. 2024b) as the base model for image understanding, and CogVLM2-Video (Hong et al. 2024b) as the base model for video understanding. In terms of data, we have obtained millions of annotated binary question-answering pairs. Initially, we performed a balanced sampling on each binary question by addressing the imbalance between positive and negative examples, ensuring a roughly equal number of positive and negative instances associated with each binary question. Then we use balanced instruction tuning dataset consisting of binary questions to fine-tune base VLM.

Interpretable Preference Learning. After trained on fine-grained dataset, VisionReward can be adopt to give a series of binary response answers ("yes" or "no") $\{A_i\}_{i=1}^N$, where N represents the number of binary questions. We define reward of every binary question as $\{x_i\}_{i=1}^N$:

$$x_i = \mathbb{1}\left[A_i = \text{“yes”}\right]. \tag{3}$$

We construct a feature vector $X = (x_1, \ldots, x_N)$, and use a set of linear weights $W = (w_1, \ldots, w_N)$ to obtain the final reward R:

$$R = \sum_{i=1}^{N} w_i \mathbb{1}\left[A_i = \text{“yes”}\right]. \tag{4}$$

In order to learn linear weights W, we collect human preferences for pairs of $\{(X_i, X_j)\}$. Specifically, we compute the feature difference for each pair, given by $\Delta X = X_i - X_j$, and the corresponding label is assigned as $y = 1$ or $y = 0$ depending on the human preference. We then perform logistic regression $y = \Delta X W^T$ to learn linear weights W:

$$\begin{aligned}\mathcal{L}(W) = -\,\mathbb{E}\big[y \log\left(\sigma(\Delta X W^T)\right) \\ +(1-y)\log\left(1-\sigma(\Delta X W^T)\right)\big].\end{aligned} \tag{5}$$

By calculating dimension-specific scores through intra-dimensional weighting, VisionReward facilitates multi-dimensional preference prediction. We note dimensions as $\{\text{dim}_k\}_{k=1}^{K}$ where dim_k contains questions belonging to the dimension. Then we define reward for certain dimension as:

$$R(\text{dim}_k) = \sum_{i \in \text{dim}_k} w_i \mathbb{1}\left[A_i = \text{“yes”}\right]. \tag{6}$$

3.3 Multi-Dimensional Preference Optimization

To empirically validate the model's capacity, we leverage Direct Preference Optimization (DPO) (Wallace et al. 2024) for Diffusion Models in our experiments, where VisionReward generates multi-dimensional preference pairs to guide the optimization process while maintaining inter-dimensional balance.

Challenges. We replicate the Diffusion-DPO training procedure using SDXL (Podell et al. 2023) on the Pick-a-Pic (Kirstain et al. 2023) dataset, employing VisionReward for comprehensive data analysis and model evaluation. As demonstrated in Fig. 4, both the preference data and optimized model exhibit biases across several fine-grained dimensions. These findings not only underscore VisionReward's capability for fine-grained analysis but also emphasize the necessity for optimization approaches that account for multi-dimensional representation.

MPO: Insight and Solution. Compared to ordinary DPO method which select pairs using overall preference, we propose MPO-enhanced DPO that take account fine-grained multi-dimensional preference. For reward of two samples R^i and R^j, we define R^i as dominating R^j if $R^i(\text{dim}_k) \geq R^j(\text{dim}_k)$ holds for every dimension dim_k. The key differences between the MPO strategy and standard DPO are:

- **Ordinary DPO**: During DPO optimization, we directly select the pair based on the total reward R.
- **MPO-enhanced DPO**: MPO strategy introduces an additional constraint: we only select pairs that R^i dominates R^j, then proceed with standard DPO.

We analyze the effects of MPO in Section 4.3. The MPO strategy can also be applied to other algorithms, which we leave for future exploration.

(a) Data analysis. (b) DPO analysis.

Figure 4: (a) We sample 10,000 human preference pairs from Pick-a-Pic dataset and analyze score deviations across 18 sub-dimensions (represented by the average yes-proportion of checklist questions within each sub-dimension). (b) We show score deviations for images generated by SDXL after Diffusion-DPO, using the same 10,000 prompts.

4 Experiments

4.1 VisionReward for Text-to-Vision Evaluation

Dataset & Training Setting. After balanced sampling, we obtain 40,743 images and corresponding 97,680 judgment questions for training, leaving 6,910 images for subsequent validation and test. For videos, we obtain 28,605 videos and corresponding 89,473 judgment questions for training, with 3,080 videos reserved.

To fine-tune CogVLM2 (Hong et al. 2024b), we set a batch size of 64, a learning rate of 1e-6, and train for 1,500 steps. For CogVLM-Video, we set a batch size of 64, a learning rate of 4e-6, and train for 1,500 steps.

To learn linear weights for preference prediction, we sample human preference pairs and perform logistic regression. For images, We sample 44k pairs (24k from HPDv2 (Wu et al. 2023) and 20k from ImageRewardDB (Xu et al. 2023)); and for videos, we sample prompts from VidProM (Wang and Yang 2024) and generate videos (using CogVideoX (Yang et al. 2024), VideoCrafter2 (Chen et al. 2024a) and OpenSora (Zheng et al. 2024)), getting 1,795 annotated video pairs with preference.

To establish a comprehensive evaluation benchmark for both image and video generation, we construct **MonetBench**, which contains separate test sets for images and videos, each consisting of 1,000 prompts. More details are introduced in Appendix Section 12.

Main Results: Preference Accuracy. Preference accuracy means the probability that a reward model has the same judgment as humans about which image is better. We use MonetBench to construct our test set for human preference,

Method	Image			Video			
	HPDv2	MonetBench		GenAI-Bench		MonetBench	
		tau*	diff**	tau	diff	tau	diff
task-specific discriminative models							
ImageReward (Xu et al. 2023)	74.0	48.8	56.5	48.4	72.1	55.8	58.4
PickScore (Kirstain et al. 2023)	79.8	49.8	57.6	52.4	75.4	57.7	61.6
HPSv2 (Wu et al. 2023)	83.3	48.4	55.6	49.3	73.0	59.3	62.5
MPS (Zhang et al. 2024b)	83.5	44.2	50.7	46.9	67.6	55.8	58.9
generative models							
GPT-4o (Achiam et al. 2023)	77.5	38.9	52.7	41.8	54.3	45.7	48.3
Gemini (Team et al. 2024)	60.7	27.4	55.1	46.9	61.7	52.2	56.8
VQAScore (Lin et al. 2025)	69.7	49.4	56.5	45.2	68.0	56.1	59.5
VideoScore (He et al. 2024)	76.8	45.8	52.5	47.8	71.4	49.1	54.9
VisionReward (Ours)	**81.7**	**51.8**	**59.5**	**51.8**	**74.4**	**64.0**	**72.1**

Table 2: Preference accuracy on multiple dataset. **Bold** denotes the best score within the generative models, while underline signifies the best score among all categories. Tau* means taking account of ties (Deutsch, Foster, and Freitag 2023), and diff** means dropping ties in labels (we drop ties both in labels and responses for GPT-4o and Gemini in diff** because too many ties are given by them).

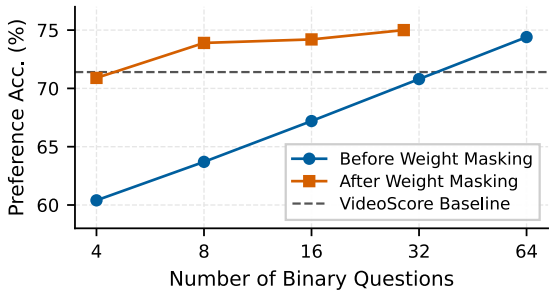

Figure 5: The accuracy of VisionReward on GenAI-Bench improves as the number of binary questions increases. After masking weights from full regression, VisionReward maintains high performance.

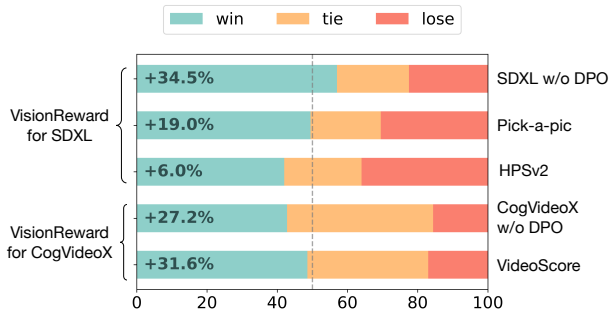

Figure 6: Human evaluation results of DPO using different reward models or preference datasets. We require five annotators to comprehensively evaluate two samples and select the better one. VisionReward achieve the best performance.

using SDXL (Podell et al. 2023) to generate images and CogVideoX (Yang et al. 2024) / VideoCrafter2 (Chen et al. 2024a) / OpenSora (Zheng et al. 2024) to generate videos, resulting in 500 pairs for image and 1,000 pairs for video. We employ annotators to assess the generated images using a preference rating scale from 1 to 5 (with 3 indicating no preference). The average preference score is used as the final preference label. We also take HPDv2 (Wu et al. 2023) and GenAI-Bench (Jiang et al. 2024) as test set.

Table 2 shows that VisionReward obtains state-of-the-state results in multiple datasets. Notably, in video evaluation, image reward models demonstrate competitive performance when the video duration is within 2 seconds (GenAI-Bench). However, when **the video duration reaches 6 seconds (MonetBench)**, only VisionReward is capable of accurately predicting human preference, being twice (**22.1% over random**) as high as the best (12.5% over random) among other methods. This indicates that dynamic information in longer videos poses a challenge for RMs, while VisionReward can effectively address this issue after fine-grained visual learning.

Ablation Study: Scalability of Question Scale. Fig. 5 shows that accuracy of preference prediction exhibits significant improvement as the question scale increases, and masking minimal weights maintains performance (Details in Appendix Section 7). Scalability validates the effects of decomposing preferences via fine-grained questions. We do more analysis of VisionReward in Appendix Section 7 and analysis of fine-grained results in Section 10.

4.2 VisionReward for Preference Optimization

To evaluate the efficacy of VisionReward in preference optimization for visual generative systems, we conduct a series of comparative experiments against current state-of-the-art reward models and established preference datasets. We use SDXL as text-to-image base model and CogVideoX as text-to-video base model. The empirical results presented in Fig. 6 demonstrate VisionReward's superior performance,

achieving statistically significant improvements in human preference metrics over competing approaches.

This section focuses on preference optimization for text-to-video using different reward models. Details of text-to-image are provided in Appendix Section 8.2.

Dataset & Training Settings. For our backbone model, we select CogVideoX-2B. The training prompts are sampled from VidProM (Wang and Yang 2024) (details in Appendix Section 8.1). To adapt these prompts for video generation, we have optimized them following guidelines from CogVideoX (Yang et al. 2024), which results in roughly 22,000 samples. We generate 4 videos for each prompt, and use VisionReward to score these videos and apply the MPO strategy to select approximately 9,400 effective preference pairs. In all our experiments, we maintain a batch size of 32, a learning rate of 5e-6, and employ 100 warmup steps followed by linear decay. We set the DPO parameter β to 500. The MPO training process spans around 500 steps, equivalent to about 2 epochs. During training, we save a checkpoint every 40 steps and use a validation set split from the training set to pick the checkpoint with the highest reward.

Evaluation Settings. To comprehensively assess the MPO models, we have conducted both automatic and human evaluations. The automatic evaluation is conducted across various benchmarks, including VBench (Huang et al. 2024) and our Video-MonetBench. For VBench, we focus on commonly reported key metrics, including *Human Action*, *Scene*, *Multiple Objects*, and *Appearance Style*. In all these experiments, we utilize prompt optimization recommended in CogVideoX. Our baseline comparisons include the original CogVideoX-2B and DPO with VideoScore.

Methods	Human Action	Scene	Multiple Objects	Appear. Style
Original	98.20	55.60	68.43	**24.20**
VideoScore	97.60	56.25	68.66	23.96
VisionReward	**98.40**	**57.57**	**71.54**	24.02

Table 3: Evaluation results on VBench.

Experimental Results. The main results are shown in Table 3. When compared to the original CogVideoX-2B, optimization with VisionReward significantly enhances model performance across these benchmarks. In contrast, optimization with VideoScore tends to degrade performance. The empirical evidence substantiates VisionReward's advanced capacity for multi-dimensional optimization. (Case study in Appendix Section 8.3.)

4.3 Ablation Study of MPO

To comprehensively illustrate how MPO addresses the factor optimization bias inherent in DPO, We conduct experiments based on CogVideoX-5B. We set the threshold of the total score to 0.8 for DPO and 0.6 for MPO, ensuring that the number of pairs obtained through all three strategies is 5k. We use a batch size of 64, a learning rate of 2e-6, the DPO parameter β of 500, and the training steps of 300. VisionReward is employed to evaluate scores across various dimensions, with the detailed results presented in Table 4. Through the implementation of the MPO strategy, CogVideoX is optimized in such a manner that it avoids the degradation of certain factors (e.g., alignment), thereby achieving improved trade-offs, such as maintaining good preservation while avoiding excessively slow dynamic changes. The empirical evidence further substantiates VisionReward's capacity for algorithm-agnostic preference alignment, as evidenced by comparative testing with other approaches like MaPO (Hong et al. 2024a).

Method	Align.	Quality	Dynamic	Physics	Preserv.	Overall
Original	1.733	0.660	0.053	0.344	0.653	4.303
DPO	1.697	0.680	0.034	0.356	**0.741**	4.515
DPO w/ MPO	**1.766**	**0.688**	0.042	0.356	0.721	**4.573**
MaPO	1.736	0.660	0.052	0.345	0.645	4.295
MaPO w/ MPO	**1.737**	0.656	**0.055**	**0.349**	0.649	**4.321**

Table 4: Ablation Study of MPO strategy. Scores are given by VisionReward on MonetBench.

For efficiency discussion, the experimental results in Table 5 reveal the comparative effectiveness of the MPO and DPO methods. We analyze the pairs selected in perspective of "R^i dominating R^j" mentioned in Section 3.3. These results suggest that MPO outperforms the DPO approach in terms of both efficiency and effectiveness. These findings highlight promising directions for future research in developing novel optimization algorithms and adapting VisionReward for multi-dimensional optimization.

Method	#Dom.	#Not-Dom.	Reward
Original	-	-	4.303
DPO w/o MPO	3814	1456	4.515 (+0.212)
DPO w/ MPO	5028	0	4.573 (+0.270)
Δ (*vs* w/o MPO)	+31.8%	-100%	+27.4%

Table 5: Comparison of MPO and DPO on MonetBench. "#Dom." means the number of pairs that match the rule of "R^i dominating R^j", while "#Not-Dom." pairs not match.

5 Conclusion

We introduce VisionReward, a reward model for visual generation, which is fine-grained and multi-dimensional. By enabling Vision-Language Model (VLM) to perform binary assessments and applying linear summation with weighting coefficients derived from preference learning, VisionReward achieves highly accurate and interpretable. For visual generative optimization, VisionReward surpasses other reward models and enable multi-dimensional strategy.

Acknowledgments

This research was supported by Natural Science Foundation of China (NSFC) No. 62276148, NSFC No. 62495063. The authors would like to thank Z.AI for sponsoring the computation resources used in this work.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; et al. 2024b. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13320–13331.

Chen, Z.; Du, Y.; Wen, Z.; Zhou, Y.; Cui, C.; Weng, Z.; Tu, H.; Wang, C.; Tong, Z.; Huang, Q.; et al. 2024c. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? *arXiv preprint arXiv:2407.04842*.

Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*.

Deutsch, D.; Foster, G.; and Freitag, M. 2023. Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12914–12929.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2023. DPOK: reinforcement learning for fine-tuning text-to-image diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 79858–79885.

He, X.; Jiang, D.; Zhang, G.; Ku, M.; Soni, A.; Siu, S.; Chen, H.; Chandra, A.; Jiang, Z.; Arulraj, A.; et al. 2024. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2105–2123.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, J.; Paul, S.; Lee, N.; Rasul, K.; Thorne, J.; and Jeong, J. 2024a. Margin-aware preference optimization for aligning diffusion models without reference. *arXiv preprint arXiv:2406.06424*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; et al. 2024b. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Ibarra, F. F.; Kardan, O.; Hunter, M. R.; Kotabe, H. P.; Meyer, F. A.; and Berman, M. G. 2017. Image feature types and their predictions of aesthetic preference and naturalness. *Frontiers in Psychology*, 8: 632.

Jiang, D.; Ku, M.; Li, T.; Ni, Y.; Sun, S.; Fan, R.; and Chen, W. 2024. GenAI Arena: An Open Evaluation Platform for Generative Models. *arXiv preprint arXiv:2406.04485*.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in neural information processing systems*, 34: 21696–21707.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 6833–6846.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19401–19411.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 366–384. Springer.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.;

et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Palmer, S. E.; Schloss, K. B.; and Sammartino, J. 2013. Visual aesthetics and human preference. *Annual review of psychology*, 64(1): 77–107.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. IEEE Computer Society.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, Y.; Durkan, C.; Murray, I.; and Ermon, S. 2021. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H.; Saffar, M. T.; Castro, S.; Kunze, J.; and Erhan, D. 2022. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

Wang, W.; and Yang, Y. 2024. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *arXiv preprint arXiv:2403.06098*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Wu, X.; Hao, Y.; Zhang, M.; Sun, K.; Huang, Z.; Song, G.; Liu, Y.; and Li, H. 2024. Deep Reward Supervisions for Tuning Text-to-Image Diffusion Models. *arXiv preprint arXiv:2405.00760*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 15903–15935.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Zhang, J.; Wu, J.; Ren, Y.; Xia, X.; Kuang, H.; Xie, P.; Li, J.; Xiao, X.; Zheng, M.; Fu, L.; and Li, G. 2024a. UniFL: Improve Stable Diffusion via Unified Feedback Learning. arXiv:2404.05595.

Zhang, S.; Wang, B.; Wu, J.; Li, Y.; Gao, T.; Zhang, D.; and Wang, Z. 2024b. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8018–8027.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-Sora: Democratizing Efficient Video Production for All.