# Boosting Text-to-Video Generative Model with MLLMs Feedback

**Xun Wu**[1], **Shaohan Huang**[1✉], **Guolong Wang**[2], **Jing Xiong**[3], **Furu Wei**[1]
[1] Microsoft Research Asia, [2] University of International Business and Economics
[3] The University of Hong Kong
xunwu@microsoft.com, shaohanh@microsoft.com, fuwei@microsoft.com

## Abstract

Recent advancements in text-to-video generative models, such as Sora [3], have showcased impressive capabilities. These models have attracted significant interest for their potential applications. However, they often rely on extensive datasets of variable quality, which can result in generated videos that lack aesthetic appeal and do not accurately reflect the input text prompts. A promising approach to mitigate these issues is to leverage Reinforcement Learning from Human Feedback (RLHF), which aims to align the outputs of text-to-video models with human preferences. However, the considerable costs associated with manual annotation have led to a scarcity of comprehensive preference datasets. In response to this challenge, our study begins by investigating the efficacy of Multimodal Large Language Models (MLLMs) generated annotations in capturing video preferences, discovering a high degree of concordance with human judgments. Building upon this finding, we utilize MLLMs to perform fine-grained video preference annotations across two dimensions, resulting in the creation of VIDEOPREFER, which includes 135,000 preference annotations. Utilizing this dataset, we introduce VIDEORM, the first general-purpose reward model tailored for video preference in the text-to-video domain. Our comprehensive experiments confirm the effectiveness of both VIDEOPREFER and VIDEORM, representing a significant step forward in the field.

## 1 Introduction

Diffusion models have significantly enhanced the quality of generation across various media formats, including images [30, 31, 55] and videos [13, 22, 42]. In the context of text-to-video generation, recent advancements in text-to-video diffusion models, exemplified by Sora [4], have achieved remarkable success in generating high-quality videos from textual prompts. However, despite recent progress, the visual fidelity of generated videos still presents opportunities for refinement [38]. Video generation poses greater challenges than image generation, as it entails modeling a higher-dimensional spatio-temporal output space while remaining conditioned solely on a textual prompt. Consequently, existing text-to-video generative models often produce results that are visually unappealing and inadequately aligned with the provided textual prompts.

To address these limitations, a pivotal solution is Reinforcement Learning from Human Feedback (RLHF), which has proven its efficacy in text-to-image diffusion models [51, 2, 26]. These methods aim to fine-tune a diffusion model to maximize a reward function corresponding to specific aspects of image quality or alignment with text prompts. *Despite this, when a Reward Model trained in the image domain is transferred to the video domain, it can exhibit significant disparities with the objectives of video optimization (e.g., InstructVideo [53]).* Training a reward model directly in the video domain is becoming increasingly urgent, but *the scale of preference datasets in the video field is small* (see in Table 1).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Statistics of existing preference datasets for text-to-video generative models. $^{*}$ denotes annotated by human while $^{\dagger}$ denotes annotated by GPT-4 V.

| Dataset | Prompts | Videos | Preference Choices |
|---|---|---|---|
| VBench [15]$^{*}$ | 1K | 4K | 44K |
| TVGE [56]$^{*}$ | 0.5K | 2.5K | 2.5K |
| T2VQA-DB [18]$^{*}$ | 1K | 45K | 45K |
| VideoPrefer$^{\dagger}$ | **14K** | **54K** | **135K** |

Table 2: Correlations between MLLMs and human judgment on text-video alignment on TVGE datasets [56]. $^{*}$ denotes image-based MLLMs while $^{\dagger}$ denotes Video-based MLLMs.

| Model | Spearman $\rho$ | Kendall $\tau$ | Acc (%) |
|---|---|---|---|
| Video-LLaMA [54]$^{\dagger}$ | 0.288 | 0.206 | – |
| mPLUG-OWL2-V [52]$^{\dagger}$ | 0.394 | 0.285 | 61.87 |
| InstructBLIP [10]$^{*}$ | 0.342 | 0.246 | 54.33 |
| mPLUG-OWL2-I [52]$^{*}$ | 0.358 | 0.257 | 53.36 |
| Gemini pro Vision$^{*}$ | 0.3921 | 0.2993 | 64.71 |
| LLaVA 1.6-34B$^{*}$ | 0.3139 | 0.2278 | 53.20 |
| GPT-4 V$^{*}$ | **0.486** | **0.360** | **69.65** |

Recognizing the importance of addressing these challenges in text-to-video generative models, we first construct a large-scale fine-grained preference benchmark by utilizing MLLMs as annotators, namely VideoPrefer. VideoPrefer contains following strength: (1) VideoPrefer is the largest open-source video preference dataset, containing 135,000 preference choices (see in Table 1). (2) Different from existing video preference datasets, VideoPrefer contains true video captured by human, making VideoPrefer more generalizable. (3) VideoPrefer is annotated using Multimodal Large Language Models, which is cost-effective and easily scalable.

Based on VideoPrefer, we release the first general-purpose text-to-video preference reward model, VideoRM. Unlike the preference rewards in the image domain used by previous methods for video generation alignment [53], our VideoRM automatically captures the temporal information in videos, enhancing the modeling of quality assessment and alignment. Furthermore, we investigate the alignment of text-to-video generative models using VideoRM and conduct extensive experiments that demonstrate its efficacy as a reward model and metric for video alignment. This significantly improves the generation quality of video generation models across multiple aspects. Our main contributions are:

- We systematically identify the challenges for text-to-video human preference annotation. Consequently, we employ MLLMs for preference annotation and construct the largest and most comprehensive video preference dataset, **VideoPrefer**.
- We systematically identify the lack of general-purpose text-to-video preference reward model and propose **VideoRM**, outperforming existing reward models for text-to-video alignment.
- Extensive experimental results validate the effectiveness of both the VideoPrefer and VideoRM. Furthermore, through detailed discussions, we demonstrate that MLLMs are effective and cost-effective as annotators for video preferences, revealing the promising future prospects of RLAIF in the domain of video preference alignment.

## 2 VideoPrefer

We introduce VideoPrefer, a large-scale, fine-grained video preference dataset constructed by collecting feedback from multimodal large language model annotators. In total, VideoPrefer contains 135K pairs of binary preference choices for 54K videos. In § 2.1, we provide evidence demonstrating that GPT-4 V is a human-aligned preference annotator in the video domain. The construction pipeline of VideoPrefer is introduced in § 2.2. Detailed analysis of the statistics of VideoPrefer can be found in Appendix § B.1.

### 2.1 Why GPT-4 V(ision) can act as a human-aligned preference annotator?

GPT-4 V has already demonstrated annotation performance aligned with human consistency in text-to-image generation [7, 46, 44]. To further validate GPT-4 V's ability to provide reliable preference annotations in the video domain, we used different MLLMs to annotate a subset of the TVGE dataset and calculated their correlation with the ground truth annotations, e.g., accuracy and kendall. The results are shown in Table 2. We find that (1) GPT-4 V's annotation accuracy and correlation are the best among all MLLMs. (2) The accuracy of GPT-4 V is 69.65%, which is very close to the previously reported agreement rate between qualified human annotators (approximately 70% [9, 46, 25, 11]). This demonstrates that GPT-4 V is a reliable annotator.

### 2.2 Construction Pipeline

The construction pipeline of VideoPrefer can be split into following three steps:

**Step-1: Prompts Collection.** We collect prompts from VidProM [40] which contains 1.67 million unique text-to-video prompts, as well as from two video-captioning benchmarks: ActivityNet[5] and MSR-VTT[49]. For VidProM, we just randomly sample 12.9k prompts from its corpus. For video-captioning benchmarks, we directly utilize the provided video segment captions from the original dataset as prompts. For instance, in the ActivityNet [5], we utilize the text provided in caption corpus, comprising 0.9K thousand text instances corresponding to distinct video segments. Consequently, we obtain a total of 14K prompts, constituting the prompt set for our VideoPrefer.

**Step-2: Video Collection & Generation.** We generate videos with text-to-video models and collect real-world videos based on prompts collected in Step-1 for preference annotation. Through these two ways, we obtained 54K video candidates:

· *Model-generated videos.* We selected the top-ranked text-to-video generation models on HuggingFace[1] (e.g., ModelScopeT2V [38]), as well as the open-source state-of-the-art text-to-video models (e.g., Open-Sora [57]), to constitute our video generation model pool. Details regarding these models, as well as the proportions and resolutions of the generated videos, can be found in Table 5. For each prompt, we randomly sample models from the model pools, along with class-free guidance scales, to generate four different corresponding videos. This approach results in a high degree of diversity, facilitating the training of a more generalizable and comprehensive preference reward model.

· *Real-world videos.* Given the substantial quality disparity between videos generated by existing text-to-video generation algorithms and real videos, we also incorporate real-world videos to enhance the generalizability and diversity of our VideoPrefer. Specifically, we incorporate video segments from the two datasets introduced in Step-1 (ActivityNet [5] and MSR-VTT [49]), as one of the four candidate videos, with the remaining three being generated as above.

**Step-3: Preference Generation.** Due to the prohibitively high cost of manual annotation, we utilize the state-of-the-art multimodal large language models (MLLMs), i.e., `GPT-4 V` to annotate video preferences, enabling extensive and fine-grained annotation at scale. The reliability of `GPT-4 V` as annotators is thoroughly discussed in § 2.1. Specifically, for each prompt and its corresponding four video candidates, we employ `GPT-4 V` to assign preference scores on a 1-to-5 Likert scale to each video candidate for two aspects: *Prompt-Following* and *Video Quality*. Lower scores indicate lower preference and vice versa. Finally, we obtain 135K preference choices. Detailed input instructions used for `GPT-4 V`'s annotation can be found in the Appendix E.

By performing the above three steps, we finally have 14K data items in VideoPrefer, while each data item contains a textual prompt and four corresponding generated videos. For each video, there are preference annotations for two aspects: Prompt-Following and Video-Quality provided. More details of VideoPrefer, e.g., visualization of example data item (see in Figure 10), can be found in Appendix B.

## 3 VideoRM: A General-Purpose Video Preference Reward Model

Based on VideoPrefer, we implement the reward model training and derive the first general-purpose video preference reward model, VideoRM. We detail the architecture of VideoRM in § 3.1, its optimization objectives in § 3.2, and the methodology for fine-tuning video generative models with VideoRM in § 3.3.

### 3.1 Architecture

Existing text-to-video alignment work [53] adopts HPS v2 [45] as the reward model, which is fine-tuned from CLIP [27] and optimized upon a large image preference benchmark and demonstrates state-of-the-art preference evaluation capabilities for text-to-image generation.

However, we contend that the direct application of HPS v2 may lead to a one-sided assessment of video preferences, as it lacks the capability to evaluate the overall attributes of a video, such

[1] https://huggingface.co/models?pipeline_tag=text-to-video

as temporal coherence and dynamics. Nevertheless, given the strong correlation between video preferences and the preferences of individual frames, the proficiency of HPS v2 [45] in evaluating single-frame image preferences can still aid in enhancing the assessment of video preferences. Thus, to achieve a better evaluation for video preferences, we modify the structure of HPS v2 by integrating several temporal modeling modules, to develop a specialized reward model VideoRM for the video domain.

The full architecture of VideoRM is shown in Figure 1. Specifically, inspired by recent advancements [39, 21] in enhancing the temporal modeling capabilities of the CLIP [27], we add two kinds of temporal modeling modules into HPS v2 to modeling videos:

**Temporal Shift** [20], a parameter-free module which shifts part of the feature channels along the temporal dimension and facilitates information exchange between neighboring input frames. Following [39], we insert the module between every two ViT Layers of the image encoder of HPS v2.

**Temporal Transformer**. The sequence of frame-wise features extracted by the image encoder are then fed into a temporal transformer to modeling the temporal features of the video.

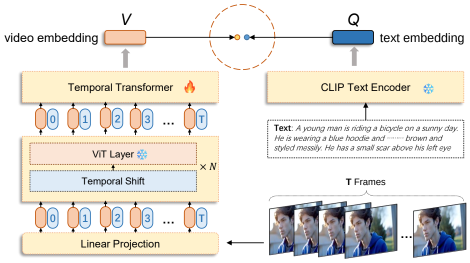

V
video embedding
Q
text embedding
Temporal Transformer 🔥
CLIP Text Encoder ❄
0
1
2
3
...
T
ViT Layer ❄
Temporal Shift
× N
Linear Projection
Text: A young man is riding a bicycle on a sunny day. He is wearing a blue hoodie and ········ brown and styled messily. He has a small scar above his left eye
T Frames
...

Figure 1: **Overview of VideoRM**. By incorporating temporal modeling modules, VideoRM is capable of not only capturing the preference scores of individual frames but also modeling the temporal features of the video, thereby better evaluating the overall preference score of the video.

### 3.2 Optimization

During the optimization process of VideoRM, we freeze the text encoder and ViT layers contained in the image encoder to retain the single-frame preference modeling capability of HPS v2, while only optimizing the temporal transformer module. Thus, VideoRM achieves the capability to model video preferences from both the perspectives of individual frames and temporal dynamics, allowing for a more comprehensive understanding of video content.

Similar to previous works [36, 25], to train VideoRM on VideoPrefer, we first average the scores of all 16 aspects for each video candidate, obtaining the final preference score for that video candidate. Then we formulate the preference score for each pair-wise video candidates as rankings. Thus, for each data item (consisting of one prompt $\mathbf{T}$ and its corresponding four video candidates $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ and $\mathbf{v}_4$), we get at most $C_4^2$ comparison pairs if there are no ties between two video candidates. For each comparison, if $\mathbf{v}^+$ is better and $\mathbf{v}^-$ is worse ($\mathbf{v}^+ \succ \mathbf{v}^-$), the loss function can be formulated as:

$$\text{loss}(\theta) = -\mathbb{E}_{(\mathbf{T},\mathbf{v}^+,\mathbf{v}^-)\sim\mathcal{D}}\left[\log\left(\sigma\left(R_\theta\left(\mathbf{T},\mathbf{v}^+\right) - R_\theta\left(\mathbf{T},\mathbf{v}^-\right)\right)\right)\right] \tag{1}$$

where $R_\theta(\mathbf{T}, \mathbf{v})$ is a scalar value of preference model for prompt $\mathbf{T}$ and generated video $\mathbf{v}$.

### 3.3 Fine-tuning Text-to-Video Models with VideoRM

Current exploration of reward reinforcement learning algorithms for fine-tuning text-to-video generative models is quite limited. The only related work is InstructVideo [53], which utilizes an image-domain reward model (HPS v2 [45]) to fine-tune text-to-video models.

Consequently, InstructVideo may have the following shortcomings: (1) Due to the inherent gap between images and videos, directly using image-domain reward models cannot accurately calculate the reward for generated videos, leading to visual artifacts such as structural twitching and color jittering [53]. (2) Additionally, InstructVideo [53] calculates the reward value for all frames selected from generated videos throughout the full DDIM [34] sampling procedure, making the fine-tuning process highly sample inefficient.

VideoRM is specifically designed for the video domain. It is initialized with the weights of the best image-domain reward model and trained on video preference datasets, utilizing temporal modeling mechanisms (e.g., temporal transformers) to capture temporal features of videos. This enables it to evaluate the reward score of generated videos more effectively from both individual frame and temporal perspectives. Additionally, unlike the image-domain reward model used in InstructVideo, which takes single video frames as input, VideoRM processes the entire video as input. Therefore, our approach is more computational efficient and can directly leverage various effective image-domain reward reinforcement learning algorithms (such as DRaFT [8]) without the need to balance rewards between frames generated at each step, as required by InstructVideo.

Based on the above, we integrated VideoRM into the image-domain DRaFT [8] algorithm and design a reward reinforcement learning algorithm suitable for text-to-video generative models, named DRaFT-V. The specific algorithm details are shown in Algorithm 1. DRaFT-V truncates the backward pass, differentiating and computing reward score from VideoRM through only the last $K$ steps, making the full fine-tuning process more efficient.

**Algorithm 1** DRaFT-V: Reward Reinforcement Learning for Fine-tuning Text-to-Video Models with VideoRM

```
Dataset: Prompt set $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$
Pre-training Dataset: Text-Video pairs dataset $\mathcal{D} = \{(\mathrm{t}_1, \mathrm{v}_1), ...(\mathrm{t}_n, \mathrm{v}_n)\}$
Input: Text-to-video models $\Upsilon$ with pre-trained parameters $\theta_0$, VideoRM $R$, reward-to-loss map function $\theta$, Text-to-video models pre-training loss function $\psi$, reward re-weight scale $\lambda$
Initialization: The number of noise scheduler time steps $T$, and the truncating step $K$
for $y_i \in \mathcal{Y}$ and $(\mathrm{t}_i, \mathrm{v}_i) \in \mathcal{D}$ do
    $\mathcal{L}_{pre} \leftarrow \psi_{\theta_i}(\mathrm{t}_i, \mathrm{v}_i)$
    $w_i \leftarrow w_i$ // Update $\Upsilon_{\theta_i}$ using $\mathcal{L}_{pre}$
    $x_T \sim \mathcal{N}(0, I)$ // Sample noise as latent
    for $j = T, ..., 1$ do
        if $j > K$ then
            no grad: $x_{j-1} \leftarrow \Upsilon_{\theta_i}\{x_j\}$
        else
            with grad: $x_{j-1} \leftarrow \Upsilon_{\theta_i}\{x_j\}$
            $x_0 \leftarrow x_j$ // Predict the original latent by noise scheduler
            $z_i^j \leftarrow x_0$ // From latent to image
            $\mathcal{L}_{reward} \leftarrow \lambda\phi(r(y_i, z_i^j))$ // Reward loss
        end if
    end for
    $\theta_{i+1} \leftarrow \theta_i$ // Update $\Upsilon_{\theta_i}$ using $\mathcal{L}_{reward}$
end for
```

In implementation, rather than fine-tuning the full set of model parameters, we follow [53] to adopt LoRA [14] to further accelerate fine-tuning and circumvent the issue of computational intensity as well as the risk of catastrophic forgetting associated with the reward loss in diffusion models.

# 4 Experiments

We conduct extensive experiments to validate the effectiveness of VideoPrefer and VideoRM. We first train VideoRM and evaluate it on existing human-preference benchmark (§ 4.1). Next, we fine-tune existing text-to-video diffusion models for aligning human preference by utilizing DRaFT-V with VideoRM (§ 4.2). Finally, we present ablation studies (§ 4.3).

## 4.1 Task 1: Reward Modeling

**Setup**. Based on VideoPrefer, we develop VideoRM, an advanced open-source general-purpose reward model that provides preferences for generated videos. Specifically, we train three versions of VideoRM. (1) Firstly, to validate the effectiveness of VideoPrefer, we average the preference scores in each aspect of VideoPrefer to get a final preference score, and train **VideoRM-V** with the merging version of VideoPrefer. (2) Then, we train **VideoRM-H** on several open-source human-crafted preference datasets listed in Table 1. (3) Finally, to build a stronger RM for text-to-video generation, we mix these open-source human-crafted preference datasets with VideoPrefer to train **VideoRM**. The details for dataset processing can be found in Appendix A. All VideoRM series models are trained in half-precision on 8 × 32GB NVIDIA V100 GPUs, with a learning rate of 1e-5 and batch size of 64 in total. We set the input frames $N$ = 8.

**Compared Baselines**. Due to the lack of reward models in the video domain, we compare VideoRM with state-of-the-art reward models from the image domain, i.e., CLIP ViT-H/14 [27], ImageReward [48], PickScore [17] and HPS v2 [45]. For all these reward models from the image domain, we calculate the scores for all video frames and then take the average as the final reward score for the video.

**Preference Accuracy**. The preference prediction accuracy across test benchmarks from three human-crafted preference datasets are reported in Table 3. As we can see, the VideoRM series outperform baseline reward models by a large margin, indicating that VideoRM series are the best open-source reward models for text-to-video domain. We also find that VideoRM-V which does not train on any open-source video preference datasets also surpasses all other baselines. This result validates the high quality of VideoPrefer enables strong out-of-distribution generalization and validate

Table 3: Pair-wise preference prediction accuracy across human-crafted preference datasets. The Aesthetic Classifier (simplified as Aesthetic) makes prediction without seeing the text prompt. The best results are in blod and the second are underlined.

| **Model** | TVGE [56] | VBench [15] | T2VQA-DB [18] | **Avg** |
|---|---|---|---|---|
| CLIP ViT-H/14 [27] | 57.3 | 52.7 | 52.1 | 54.0 |
| Aesthetic [33] | 56.1 | 51.0 | 52.7 | 53.3 |
| ImageReward [48] | 66.8 | 54.3 | 53.9 | 58.3 |
| PickScore [17] | 64.7 | 53.9 | 61.2 | 59.9 |
| HPS v2 [45] | 69.5 | 55.7 | 52.8 | 59.3 |
| **VideoRM-H** | 72.8 | 60.2 | 64.3 | 65.8 |
| **VideoRM-V** | <u>73.0</u> | <u>61.1</u> | <u>65.1</u> | <u>66.4</u> |
| **VideoRM** | **73.7** | **63.5** | **65.4** | **67.5** |

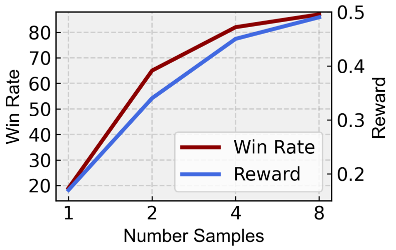


Figure 2: Best-of-$n$ experiments on the T2VQA-DB [18] test benchmark. We sample $n$ generated videos and choose the one with the highest reward.

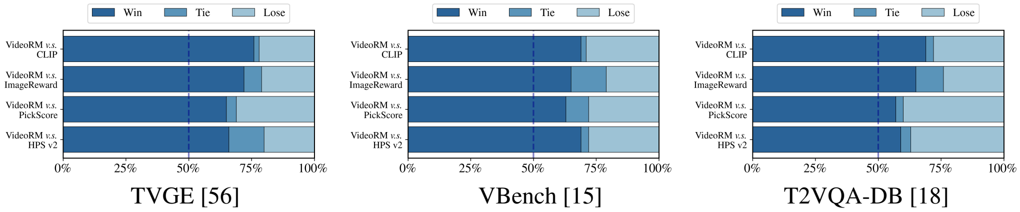


Figure 3: Win rates of VideoRM compared to other reward models across three test benchmarks. On average, 72% to CLIP, 67% to ImageReward, 62% to PickScore and 65% to HPS v2.

the effectiveness of treating state-of-the-art MLLMs, i.e., `GPT-4 V`, as preference annotator for text-to-video generation domain.

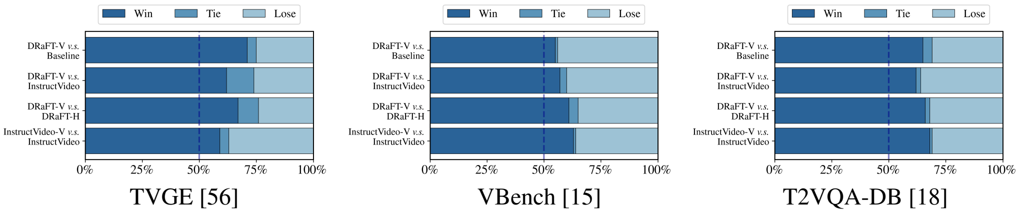
**Best-of-$n$ Experiments**. To verify that our VideoRM could serve as a good indicator of video generation quality, we conduct best-of-$n$ experiments on the T2VQA-DB [18] test benchmark. For each data item, which includes a prompt and ten corresponding videos, we calculate the reward score for each of the ten videos with VideoRM. Thereafter, we select the best-of-$\{1, 2, 4, 8\}$ responses and calculate their scores. The final results are presented at Figure 2, we find the win rate at the test benchmark increases proportionally with rewards, which validates that our VideoRM gives rigorous rewards that reflect the overall preference score.

**Human Evaluation**. To evaluate the ability of VideoRM to select the more preferred videos among large amounts of generated videos, we produce human study. We randomly select 200 textual prompts from these three test benchmarks and generate 32 different videos for each prompt by utilizing ModelScopeT2V [38]. Then, we perform different reward to select from those videos to get top3 results. After that, five annotators are asked to identify which video was superior or if both were of equal quality (denoted as 'Tie') and we show the corresponding win rates against other reward models [45] at Figure 3. Qualitative results can be seen at Figure 11 in Appendix. All of these results show that VideoRM can select videos that are more aligned to text and with higher fidelity and avoid toxic contents.

### 4.2 Task 2: Fine-tuning Text-to-Video Generative Models

**Setup.** Based on VideoRM, we validate the effectiveness of DRaFT-V. We adopt the publicly available text-to-video diffusion model ModelScopeT2V [38] as base model, which is trained on WebVid10M with T = 1000 and is able to generate videos of 16 × 256 × 256 resolution. We random sample 2K prompt-video pair from the mixture of existing video preference datasets (TVGE [56], VBench [15] and T2VQA-DB [18]) and VideoPrefer as the training data. Each group of experiment is trained in half-precision on 4 × 32GB NVIDIA V100 GPUs, with a learning rate of 1e-5, batch size of 8, fine-tuning step of 400 and $K$ adopted in DRaFT-V as 10 in total.

**Compared Baselines.** Since the only existing reward-based fine-tuning algorithm for text-to-video models is InstructVideo [53]. We compare our DRaFT-V with the the baseline text-to-video model (i.e., ModelScopeT2V [38]) and InstructVideo implemented by ourselves. Note that in the aforementioned algorithms, InstructVideo uses HPS v2 [45] as the reward model. Additionally, to verify the effectiveness of using VideoRM as a reward model for fine-tuning video generation models, we

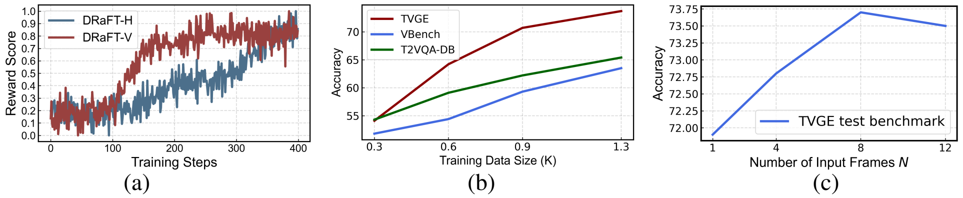


TVGE [56]

VBench [15]

T2VQA-DB [18]

Figure 5: Win rates of text-to-video models fine-tuned with DRaFT-V compared to other baselines. Here baseline denotes the base text-to-video model without any fine-tuning.



(a)

(b)

(c)

Figure 6: (a) Visualization of reward model values change with the training steps for both DRaFT-V and DRaFT-H. (b) Evaluation result across three test benchmarks for the size of training data used in optimizing VIDEORM. (c) Ablation study for the number of input frames $N$ in VIDEORM.

introduced two comparison groups: (1) DRaFT-H, which uses the same algorithm as our DRaFT-V (i.e., DRaFT [8]), but with HPS v2 as the reward model. (2) InstructVideo-V, which uses the same algorithm as InstructVideo [53], but with VIDEORM as the reward model.

**Generation Quality Evaluated by Multiple Reward Models.** At first, we visualize the evolution of model's generation quality across multiple reward models as the model training steps increase when using our DRaFT-V at Figure 4. We observe that with the progress of training, all reward models show an increasing trend, indicating that VIDEORM can serve as a reliable reward model to enable the generated videos to align more closely with human preferences.

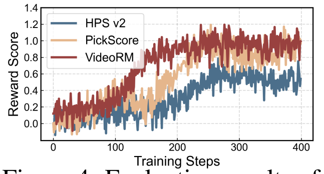


Figure 4: Evaluation results of the text-to-video model's generation quality across multiple reward models when maximizing scores from VIDEORM during the DRaFT-V fine-tuning process.

**Human Evaluation.** To evaluate the generative ability of text-to-video models fine-tuned with DRaFT-V and other compared baselines, we produce human study. We randomly select 100 prompts from these three test benchmarks (TVGE [56], VBench [15] and T2VQA-DB [18]), and for each prompt, we generate videos using different fine-tuned models. Five annotators are asked to identify which video was superior or if both were of equal quality (denoted as 'Tie') and we show the corresponding win rates at Figure 5. We find that: (1) All of these results show that DRaFT-V can generate videos that are more prefered by human. (2) By comparing DRaFT-V with DRaFT-H, as well as InstructVideo-V with InstructVideo, we demonstrate that VIDEORM is a more effective reward model for fine-tuning text-to-video models.

Besides, qualitative results are presented at Figure 12 in Appendix D. We find that: (1) The base generation model often fails to align videos with prompt descriptions, but fine-tuning with a reward model (image or video domain) improves quality and prompt consistency, demonstrating the effectiveness of reward-based fine-tuning. (2) Fine-tuning with VIDEORM (DRaFT-V and InstructVideo-V) significantly outperforms fine-tuning with an image domain reward model HPS v2, indicating VIDEORM's superiority for text-to-video model fine-tuning.

**Efficiency Evaluation.** We compare the efficiency of our algorithm on two aspects: (1) Convergence speed: we visualize the changes in reward values with training steps for DRaFT-H and DRaFT-V in Figure 6 (a). We find that under the same reward-based fine-tuning algorithm, using VIDEORM leads to earlier convergence and better performance (as demonstrated by the results above) compared to HPS v2, further validating the effectiveness of VIDEORM in fine-tuning text-to-video models. (2) Inference speed. For a single video with 8 frames, HPS v2 requires approximately an average of 5 seconds for inference, while VIDEORM only needs an average of 1.3 seconds. Besides, InstructVideo requires 20 DDIM steps for a single fine-tuning step, taking approximately an average of 8.8 seconds, while DRaFT-V only needs an average of 2.3 seconds.

### 4.3 Ablation Study

**Scalability.** To investigate the effect of training dataset sizes on the performance of VideoRM, and to verify the scalability of VideoPrefer, comparative experiments are conducted. Note that VideoRM in this experiment are only trained on VideoPrefer. Figure 6 (b) shows that adding up the scale of the dataset significantly improves the preference accuracy of VideoRM. It's promising that if we employ GPT-4 V to collect more annotation data in the future, VideoRM will get better performance.

**Number of Video Frames $N$ Input to VideoRM.** We investigate the impact of different numbers of input video frames on the performance of the RM model. Specifically, we set the number of input frames to $N$ = 1, 4, 8, and 12 and test corresponding model at TVGE [56] test benchmark. The corresponding results are presented at Figure 6 (c), we find that the VideoRM's performance significantly decreases when $N$ is small. This is promising because too few input frames prevent the model from accurately capturing all the information and temporal features of the video, e.g., the degree of motion and temporal consistency.

**Temporal Feature Modeling method in VideoRM.** We conduct experiments to explore the impact of different video frame temporal feature modeling methods on the performance of the reward model. In VideoRM, we use a temporal transformer trained from scratch to model the temporal features of video frames. Additionally, we conducted ablation experiments by replacing this transformer with 1D Convolutional Layer (denoted as VideoRM$^{\dagger}$) and LSTM (denoted as VideoRM$^{\ddagger}$). Note that we adjusted the number of layers in the temporal feature modeling module to ensure that the number of trainable parameters remains equal. Besides, Temporal Shift are employed for all compared methods. The results are presented at Table 4. We find that: (1) All models outperform the image-domain reward model. (2) Replacing the temporal transformer with LSTM achieved comparable performance, indicating that VideoRM is robust to the choice of model architecture for temporal feature modeling. (3) Replacing the temporal transformer with Conv1D resulted in a significant performance drop. We hypothesize that this is because Conv1D cannot effectively model temporal features, leading to VideoRM$^{\dagger}$'s inability to accurately evaluate video generation quality from a temporal perspective. This highlights the effectiveness of using a temporal transformer for modeling temporal features in VideoRM.

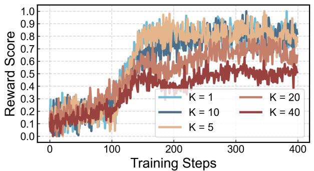


Figure 7: Ablation over $K$ adopted in DRaFT-V during fine-tuning text-to-video model.

**Backbone of VideoRM.** We explore the impact of different backbone on the final performance of VideoRM. In our design, VideoRM is initialized with the weights of HPS v2 [45], which is demonstrated as the best reward model for image domain. Here we replace HPS v2 with other reward models in image domain, e.g., PickScore [17] (denoted as VideoRM$^{\sigma}$) and ImageReward [48] (denoted as VideoRM$^{\beta}$) while keep the count of trainable parameter equal. The corresponding results are presented at Table 4. We find that the final performance of the corresponding VideoRM is proportional to the performance of the initialized backbone weights. For example, ImageReward, which has the lowest performance among the three (see Table 3), corresponds to the worst VideoRM$^{\beta}$, while PickScore, which has the best performance among the three (see Table 3), corresponds to the best VideoRM$^{\sigma}$.

**Step $K$ adopted in DRaFT-V.** We investigate the impact of different $K$ adopted in DRaFT-V on the final generative performance. The results are presented at Figure 7, we find that: (1) Performance degrades as K increases for $K$ > 10. (2) Even with smaller values of $K$, e.g., $K$ = 1, DRaFT-V can achieve good results. Both findings are aligned with the results in DRaFT [8].

## 5 Analysis

In this section, we analyze the impact of different parameter selections—specifically, the number of frames ($N$) sampled from a single video and the temperature setting $\tau$ of the GPT-4 V, on the accuracy of GPT-4 V-generated annotations. Due to GPT-4 V's limitation of processing a maximum of 10 frames, we analyze annotation accuracy for $N$ = 4, 6, 8, and 10 frames. The temperature $\tau$ controls the diversity and randomness of GPT-4 V's outputs. We analyze annotation accuracy for $\tau$ = 0.3, 0.5, 0.7, and 0.9.

Table 4: Ablation study for VIDEORM. The Aesthetic Classifier (simplified as Aesthetic) makes prediction without seeing the text prompt.

| **Model** | TVGE [56] | VBench [15] | T2VQA-DB [18] | **Avg** |
|---|---|---|---|---|
| VIDEORM† | 70.1 | 58.4 | 61.5 | 63.3 |
| VIDEORM‡ | **73.9** | 62.7 | 65.2 | 67.3 |
| VIDEORM$^{\sigma}$ | 72.2 | **64.2** | **66.9** | 67.7 |
| VIDEORM$^{\beta}$ | 72.4 | 63.6 | 64.2 | 66.7 |
| VIDEORM | 73.7 | 63.5 | 65.4 | **67.5** |

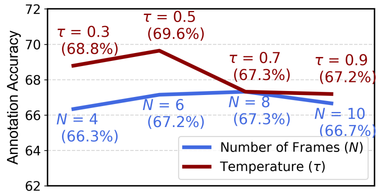


Figure 8: The impact of different hyper-parameter choices on the annotation accuracy of `GPT-4 V`.

The summarized results are shown in Figure 8. For $N$, accuracy initially increases with more frames but decreases beyond $N = 8$, achieving the highest accuracy at $N = 8$. This is because more frames allow for a comprehensive evaluation, but too many frames can lead to overly long contexts, reducing accuracy due to `GPT-4 V`'s limited capacity. For $\tau$, lower values yield higher accuracy, likely because reduced randomness leads to more comprehensive predictions, while higher randomness might cause `GPT-4 V` to overly focus on specific frames or aspects, resulting in less comprehensive outcomes.

## 6 Related Work

Generative models, particularly those based on diffusion techniques, have demonstrated high-quality generation capabilities [32, 28, 31, 30, 35] by training on extensive internet-scale datasets, but the mixed quality of these datasets often leads to visually unappealing and misaligned outputs.

**Aligning text-to-image generative models.** Aligning text-to-image generative models [37, 17, 8, 26, 37] has garnered increasing attention in recent years and has shown promising results in producing outputs that are more aligned with human preference. Imagereward [48], HPS v2 [47, 45], and PickScore [17] are the three most commonly used reward models for align text-to-image models. They are respectively trained on three major image domain preference datasets, namely Imagereward, ImageRewardDB [48], HPD [47, 45] and Pick-a-Pic [17]. The effectiveness of these reward models has been validated across alignment algorithmss like AlignProp [26], DRAFT [8] and ReFL [48].

**Aligning text-to-video generative models.** Compared to text-to-image, exploration related to aligning text-to-video generative models is relatively sparse. InstructVideo [53] instruct text-to-video diffusion models with by fine-tuning with existing image-wise preference reward model HPS v2. We hypothesize that this may restrict effective preference evaluations for generated videos, as image-based reward models cannot adequately capture temporal features, impairing assessments of video coherence and dynamics. Additionally, preference datasets for text-to-video generation are also scarce. The lack of large-scale and effective open-source preference datasets severely restricts the development of research related to align text-to-video generative models.

**Reinforcement Learning from AI Feedback.** LLMs have been extensively used for data generation [43, 24], augmentation [12] and in self-training setups [41, 23]. Some works [1] introduced the idea of reinforcement learning from AI feedback (RLAIF), which used LLMs labeled preferences in conjunction with human labeled preferences to jointly optimize for the two objectives of helpfulness and harmlessness. Recent works have also explored related techniques for generating rewards from LLMs [29, 19, 50]. These works demonstrate that LLMs can generate useful signals for reinforcement learning fine-tuning. However, leveraging the feedback from MLLMs for aligning text-to-image generative model or for text-to-video generative model is less explored.

## 7 Conclusion

In this paper, we identify the current obstacles faced in aligning text-to-video generative models, i.e., lacking of large-scale preference datasets and reward models specifically tailored for videos. Thus, we introduce VIDEOPREFER and VIDEORM to address the aforementioned issues. Experimental validation confirms that `GPT-4 V` can act as a human-aligned preference annotator. We utilized it to label 135K video preference annotation, forming VIDEOPREFER. Based on this, we trained a video-specific reward model, VIDEORM. Extensive analytical usage has demonstrated the effectiveness of both VIDEOPREFER and VIDEORM.

## References

[1] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

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

[3] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024.

[4] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024.

[5] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[6] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[7] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

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

[9] G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[10] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[11] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. S. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, Aug. 2021. Association for Computational Linguistics.

[13] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[16] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators, 2023.

[17] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

[18] T. Kou, X. Liu, Z. Zhang, C. Li, H. Wu, X. Min, G. Zhai, and N. Liu. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024.

[19] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh. Reward design with language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[20] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding, 2019.

[21] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end video clip retrieval, 2021.

[22] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023.

[23] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

[24] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, and J. Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR, 2023.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[26] M. Prabhudesai, A. Goyal, D. Pathak, and K. Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[28] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[29] P. Roit, J. Ferret, L. Shani, R. Aharoni, G. Cideron, R. Dadashi, M. Geist, S. Girgin, L. Hussenot, O. Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023.

[30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[31] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[32] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[33] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[34] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[35] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

[37] Z. Tang, D. Rybin, and T.-H. Chang. Zeroth-order optimization meets human feedback: Provable learning via ranking oracles. *arXiv preprint arXiv:2303.03751*, 2023.

[38] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[39] M. Wang, J. Xing, and Y. Liu. Actionclip: A new paradigm for video action recognition, 2021.

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

[41] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[42] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[43] Z. Wang, A. W. Yu, O. Firat, and Y. Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021.

[44] S. Wen, G. Fang, R. Zhang, P. Gao, H. Dong, and D. Metaxas. Improving compositional text-to-image generation with large vision-language models. *arXiv preprint arXiv:2310.06311*, 2023.

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

[46] X. Wu, S. Huang, and F. Wei. Multimodal large language model is a human-aligned annotator for text-to-image generation. *arXiv preprint arXiv:2404.15100*, 2024.

[47] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[48] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

[49] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[50] K. Yang, D. Klein, A. Celikyilmaz, N. Peng, and Y. Tian. Rlcd: Reinforcement learning from contrast distillation for language model alignment, 2023.

[51] K. Yang, J. Tao, J. Lyu, C. Ge, J. Chen, Q. Li, W. Shen, X. Zhu, and X. Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023.

[52] Q. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[53] H. Yuan, S. Zhang, X. Wang, Y. Wei, T. Feng, Y. Pan, Y. Zhang, Z. Liu, S. Albanie, and D. Ni. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312.12490*, 2023.

[54] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[55] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[56] J. Zhangjie Wu, G. Fang, H. Wu, X. Wang, Y. Ge, X. Cun, D. Junhao Zhang, J.-W. Liu, Y. Gu, R. Zhao, et al. Towards a better metric for text-to-video generation. *arXiv e-prints*, pages arXiv–2401, 2024.

[57] Z. Zheng, X. Peng, and Y. You. Open-sora: Democratizing efficient video production for all, March 2024.

# A Datasets Processing

**T2VQA-DB [18]**. The T2VQA-DB dataset only comprises a single overall preference evaluation score for each data item. We randomly selected 2.3K samples to serve as the final test benchmark.

**TVGE [56]**. The TVGE dataset comprises preference evaluations from 2 aspects, namely: *text alignment* and *video quality*. We just simply average the scores across these two aspects to obtain a final overall evaluation score as with the aforementioned datasets. We randomly selected 15K samples to serve as the final test benchmark.

**VBench [15]**. The VBench dataset comprises preference evaluations from 16 aspects, namely: *subject consistency*, *background consistency*, *temporal flickering*, *motion smoothness*, *dynamic degree*, *aesthetic quality*, *imaging quality*, *object class*, *multiple objects*, *human action*, *color*, *spatial relationship*, *scene*, *temporal style*, *appearance style*, *overall consistency*.

Due to the fact that the videos and prompts evaluated for preference in each aspect are largely non-overlapping, we cannot simply average the scores across all aspects to obtain a final overall evaluation score as with the aforementioned datasets. Instead, we mixed all samples from the 16 aspects and randomly selected 10K samples to serve as the final test benchmark.

# B VideoPrefer

## B.1 Statistic of VideoPrefer

In Table 5, we present the sources and distribution of the videos in our VideoPrefer. We find that VideoPrefer comprises a diverse range of video sources, including videos generated by state-of-the-art text-to-video models as well as real videos. This extensive variety of video sources enhances the robustness and generalization capabilities of VideoPrefer.

We visualize the distribution of preference annotations across two evaluation aspects (prompt-following an video-quality) in VideoPrefer in Figure 9. We find that the overall score distribution of the dataset is close to the normal distribution for the Prompt-Following evaluation perspective. From the perspective of Video-quality evaluation, the overall distribution of data sets tends to be low, which indicates that the generation effect of the existing text-to-video model is not satisfactory.

Besides, we also visualize some data examples in VideoPrefer at Figure 10.

Table 5: Video sources of VideoPrefer. † denotes the realistic videos from existing benchmarks.

| **Source** | Proportion | Video Length | Type | Resolution |
|---|---|---|---|---|
| LaVie [42] | 16.4% | 2.0s | Diffusion | 512×512 |
| ModelScope [22] | 21.0% | 2.0s | Diffusion | 256×256 |
| VideoCrafter2 [6] | 17.2% | 1.6s | Diffusion | 320×320 |
| Open-Sora [57] | 2.4% | - | DiT | 512×512 |
| Pika[2] | 7.2% | 3.0s | DiT | 1088×640 |
| Text2Video-Zero [16] | 15.3% | 2.0s | Diffusion | 512×512 |
| ZeroScope | 13.8% | - | Diffusion | 1024×576 |
| Gen-2 | 0.7% | - | - | 1792×1024 |
| ActivityNet [5]† | 1.7% | - | Human-captured | - |
| MSR-VTT [49]† | 0.03% | - | Human-captured | - |

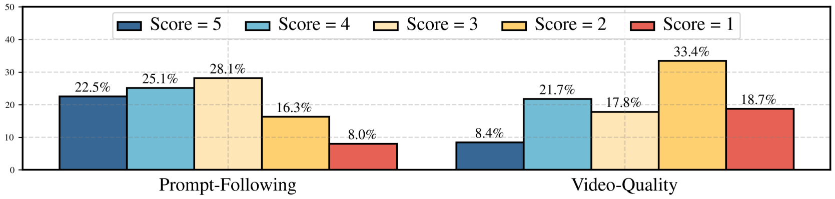
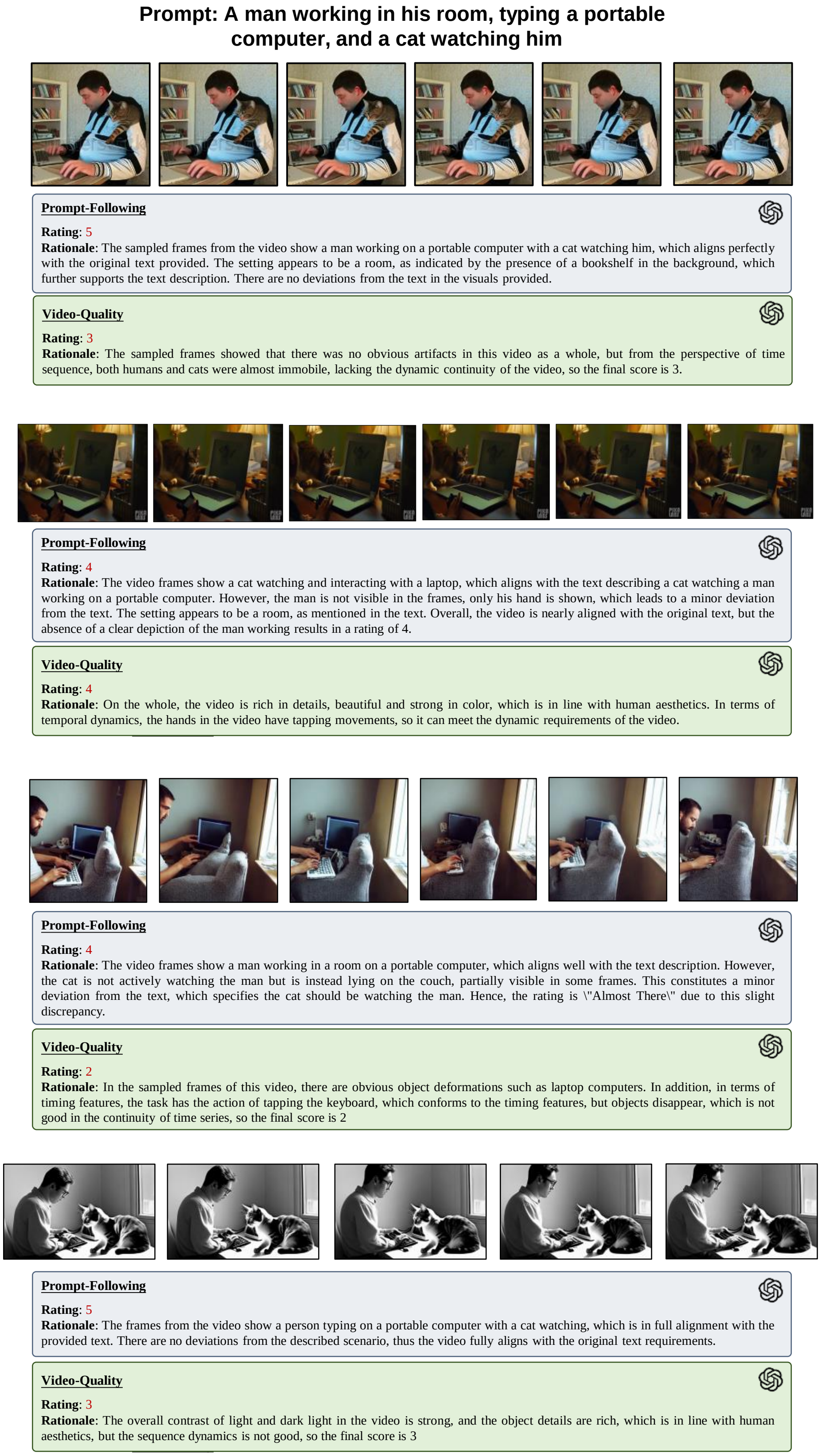


Figure 9: Score distribution across two annotation aspects in VideoPrefer.



**Prompt: A man working in his room, typing a portable computer, and a cat watching him**

**Prompt-Following**

**Rating**: 5
**Rationale**: The sampled frames from the video show a man working on a portable computer with a cat watching him, which aligns perfectly with the original text provided. The setting appears to be a room, as indicated by the presence of a bookshelf in the background, which further supports the text description. There are no deviations from the text in the visuals provided.

**Video-Quality**

**Rating**: 3
**Rationale**: The sampled frames showed that there was no obvious artifacts in this video as a whole, but from the perspective of time sequence, both humans and cats were almost immobile, lacking the dynamic continuity of the video, so the final score is 3.

**Prompt-Following**

**Rating**: 4
**Rationale**: The video frames show a cat watching and interacting with a laptop, which aligns with the text describing a cat watching a man working on a portable computer. However, the man is not visible in the frames, only his hand is shown, which leads to a minor deviation from the text. The setting appears to be a room, as mentioned in the text. Overall, the video is nearly aligned with the original text, but the absence of a clear depiction of the man working results in a rating of 4.

**Video-Quality**

**Rating**: 4
**Rationale**: On the whole, the video is rich in details, beautiful and strong in color, which is in line with human aesthetics. In terms of temporal dynamics, the hands in the video have tapping movements, so it can meet the dynamic requirements of the video.

**Prompt-Following**

**Rating**: 4
**Rationale**: The video frames show a man working in a room on a portable computer, which aligns well with the text description. However, the cat is not actively watching the man but is instead lying on the couch, partially visible in some frames. This constitutes a minor deviation from the text, which specifies the cat should be watching the man. Hence, the rating is \"Almost There\" due to this slight discrepancy.

**Video-Quality**

**Rating**: 2
**Rationale**: In the sampled frames of this video, there are obvious object deformations such as laptop computers. In addition, in terms of timing features, the task has the action of tapping the keyboard, which conforms to the timing features, but objects disappear, which is not good in the continuity of time series, so the final score is 2

**Prompt-Following**

**Rating**: 5
**Rationale**: The frames from the video show a person typing on a portable computer with a cat watching, which is in full alignment with the provided text. There are no deviations from the described scenario, thus the video fully aligns with the original text requirements.

**Video-Quality**

**Rating**: 3
**Rationale**: The overall contrast of light and dark light in the video is strong, and the object details are rich, which is in line with human aesthetics, but the sequence dynamics is not good, so the final score is 3

Figure 10: Visualization of example data item in VIDEOPREFER. Here we show one data item which contains a prompt and four corresponding generated videos. For each video, there are two annotations from different annotation aspects (Prompt-Following and Video-Quality) are provided by `GPT-4 V`

# C Visualization of Selection Results

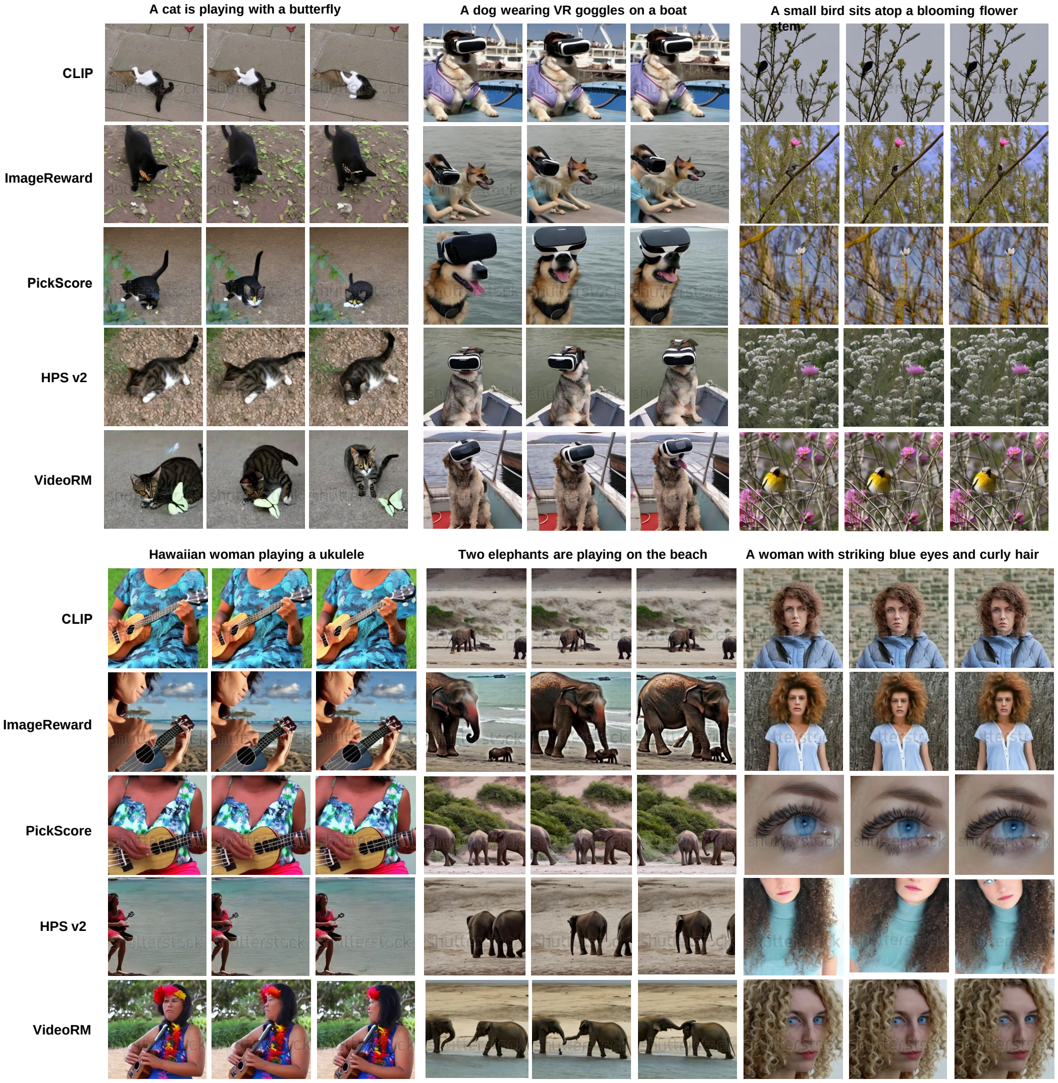

A cat is playing with a butterfly
A dog wearing VR goggles on a boat
A small bird sits atop a blooming flower stem
CLIP
ImageReward
PickScore
HPS v2
VideoRM
Hawaiian woman playing a ukulele
Two elephants are playing on the beach
A woman with striking blue eyes and curly hair
CLIP
ImageReward
PickScore
HPS v2
VideoRM

Figure 11: Top-1 videos from 32 generated videos select by CLIP, ImageReward, PickScore, HPS v2 and VIDEORM. **VIDEORM is capable of selecting higher-quality generated videos, e.g., those that better match the prompt descriptions and exhibit more dynamic content.**

# D Visualization Results for Fine-tuning Text-to-Video Models

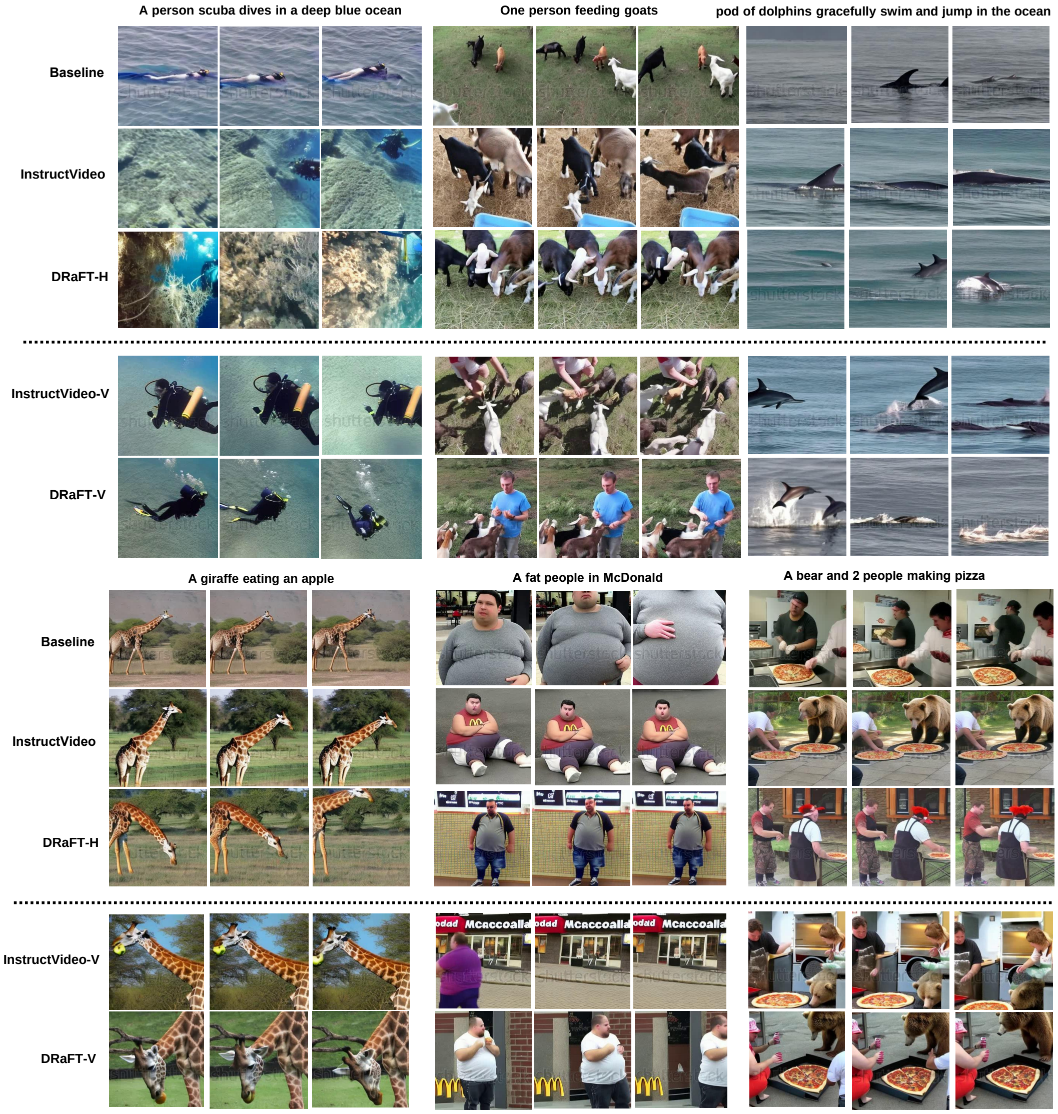

A person scuba dives in a deep blue ocean
One person feeding goats
pod of dolphins gracefully swim and jump in the ocean
Baseline
InstructVideo
DRaFT-H
InstructVideo-V
DRaFT-V
A giraffe eating an apple
A fat people in McDonald
A bear and 2 people making pizza
Baseline
InstructVideo
DRaFT-H
InstructVideo-V
DRaFT-V

Figure 12: Visualization Results for different fine-tuning methods. We find that compared to fine-tuning with an image domain reward model, fine-tuning with VIDEORM significantly enhances the performance of text-to-video models (DRaFT-V and InstructVideo-V).

## E Instruction Template

### Preference Instruction for Prompt-Following

**Prompt-Following:**
You are an AI assistant programmed to assess videos with impartial and balanced standards. A video has been created based on a piece of text. Your task is to evaluate how well The content of the video aligns with the original text provided as ("Input") The video evaluation is based on sampled frames shown in sequence.
**Scoring**: Rating outputs 1 to 5:

1. **Irrelevant**: No alignment.
2. **Partial Focus**: Addresses one aspect poorly.
3. **Partial Compliance**:
   - (1) Meets goal or restrictions, neglecting other.
   - (2) Acknowledges both but slight deviations.
4. **Almost There**: Near alignment, minor deviations.
5. **Comprehensive Compliance**: Fully aligns, meets all requirements.

Please present your assessment as follows:
**Output**
Rating: [Provide the rating]
Rationale: [Explain the reason for your rating in concise sentences]

---

Now, review the following video and its corresponding text.
**Input**:
#### Text: [INSERT PROMPT HERE]
#### Frames sampled from video: [INSERT THE FRAMES OF VIDEO HERE]

**Preference Instruction for Video-Quality:**

**Video-Quality:**
You are an AI assistant trained to impartially assess temporal consistency quality, dynamic quality and aesthetic quality in videos. A video has been created based on a piece of text. Your task is to analyze the quality of this video based on the following guidelines and provide a comprehensive evaluation.
**Scoring**: Rating outputs 1 to 5:

1. **Bad**: blurry, underexposed with significant noise, indiscernible subjects, exhibits significant inconsistencies and noticeable discrepancies in appearance of subjects.
2. **Poor**: Noticeable blur, poor lighting, washed-out colors, and awkward composition with cut-off subjects, suffers from noticeable issues in maintaining uniformity of subjects and backgrounds.
3. **Fair**: In focus with adequate lighting, dull colors, decent composition but lacks creativity. Subjects and backgrounds maintain a reasonable degree of uniformity throughout most of the video, with only minor discrepancies.
4. **Good**: Sharp, good exposure, vibrant colors, thoughtful composition with a clear focal point. Good video dynamics and temporal consistency.
5. **Excellent**: Exceptional clarity, perfect exposure, rich colors, masterful composition with emotional impact. perfect temporal consistency and excellent dynamics.

Please present your assessment as follows:
**Output**
Rating: [Provide the rating]
Rationale: [Explain the reason for your rating in concise sentences]

---

Now, review the following video and its corresponding text.
**Input**:
#### Text: [INSERT PROMPT HERE]
#### Frames sampled from video: [INSERT THE FRAMES OF VIDEO HERE]

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper includes our motivation, preference data construction details, experimental settings, quantitative experimental results, and qualitative visual examples that reflect and justify the claims in our abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The analysis section contains a discussion of limitations in our work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We state the reference before illustrating formulations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of the dataset construction process, implementation for training the reward model, and offer a multitude of visualization examples.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will public our data and code upon paper acceptance, due to the management regulations of our institution.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Full training and testing details are in §4 and Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: It would take too long with our available computational resources to repeat all experiments multiple times.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the name of GPU we used for experiments in Setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All relevant human questions and relevant datasets have been checked for privacy compliance prior to experiments and submission.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work supports the alignment of generative models with human values, and its societal impact is discussed in the introduction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our contribution does not include new datasets or pre-trained models that pose a risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Code that we derive from earlier work is properly licensed and referenced.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed illustration of new data benchmark and visualized examples.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: We only recruited participants for user experiments to validate the effectiveness of our model, where they were asked to choose from generated images. No human participants were involved in the dataset construction or model training process.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our experiment solely involves measurement and does not entail behavioral manipulation; therefore, we did not apply for IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.