Hallucination-Aware Prompt Optimization for Text-to-Video Synthesis

Jiapeng Wang[1*], **Chengyu Wang**[2†], **Jun Huang**[2], **Lianwen Jin**[1†]
[1]South China University of Technology, Guangzhou, China
[2]Alibaba Cloud Computing, Hangzhou, China
eejpwang@mail.scut.edu.cn, eelwjin@scut.edu.cn
{chengyu.wcy, huangjun.hj}@alibaba-inc.com

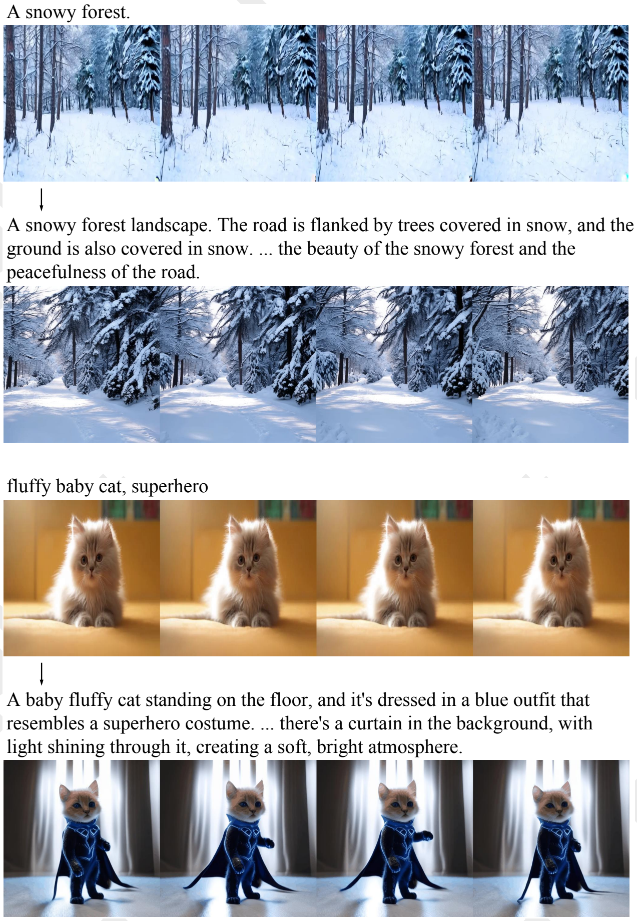
Abstract

The rapid advancements in AI-generated content (AIGC) have led to extensive research and application of deep text-to-video (T2V) synthesis models, such as OpenAI's Sora. These models typically rely on high-quality prompt-video pairs and detailed text prompts for model training in order to produce high-quality videos. To boost the effectiveness of Sora-like T2V models, we introduce **VidPrompter**, an innovative large multi-modal model supporting T2V applications with three key functionalities: (1) generating detailed prompts from raw videos, (2) enhancing prompts from videos grounded with short descriptions, and (3) refining simple user-provided prompts to elevate T2V video quality. We train **VidPrompter** using a hybrid multi-task paradigm and propose the hallucination-aware direct preference optimization (HDPO) technique to improve the multi-modal, multi-task prompt optimization process. Experiments on various tasks show our method surpasses strong baselines and other competitors.

1 Introduction

With the rapid development of the AI-generated content (AIGC) field, there has been a significant increase in research and applications of text-to-video (T2V) models [Zheng *et al.*, 2024; Xu *et al.*, 2024; Guo *et al.*, 2024; Chen *et al.*, 2024a], exemplified by Sora [Brooks *et al.*, 2024]. The goal of these models is to generate high-quality videos based on user input prompts that align with user intent. They provide a highly convenient way for users, even those without foundational knowledge of aesthetics or art, to engage with corresponding applications.

One of the key elements to enhance the performance of T2V is *prompts*, which are textual inputs to these models that express user needs. As shown in Figure 1, more detailed and informative prompts often lead to more vivid videos generated by T2V models, necessitating better *prompt optimization*. In the entire lifecycle of T2V, the functionality of prompt optimization is twofold: (1) on-demand prompt refinement for users to improve the video quality for T2V-based applications; and (2) quality improvement for video captions to achieve better alignment between textual input and video output for training T2V models, as it remains challenging for many existing models to overcome hallucination issues [Li *et al.*, 2023c; Jiang *et al.*, 2024].


A snowy forest.

A snowy forest landscape. The road is flanked by trees covered in snow, and the ground is also covered in snow. ... the beauty of the snowy forest and the peacefulness of the road.

fluffy baby cat, superhero

A baby fluffy cat standing on the floor, and it's dressed in a blue outfit that resembles a superhero costume. ... there's a curtain in the background, with light shining through it, creating a soft, bright atmosphere.

Figure 1: Illustration of how more detailed and informative prompts affect the performance of the T2V results.

In this paper, we propose a novel model, **VidPrompter**, which is designed to perform three core tasks: (1) generat-

[*] Contribution during internship at Alibaba Cloud Computing.
[†] Co-corresponding authors.

ing detailed prompts from raw videos; (2) producing more comprehensive prompts from raw videos that are grounded with short descriptions; and (3) enriching user-provided simple prompts into more detailed prompts. Our model integrates these highly relevant capabilities required by T2V models into a single framework, thereby avoiding redundant training. Additionally, we have found that mixed training can positively impact the performance of these sub-tasks, leading to mutual reinforcement.

To train the **VidPrompter** model, we build upon existing large multi-modal models (LMMs) and further enhance specific tasks during alignment learning. Direct preference optimization (DPO) [Rafailov *et al.*, 2023] has emerged as the predominant method for aligning large language models (LLMs) with human preferences. However, existing work [Li *et al.*, 2023b; Wang *et al.*, 2024a] has found that merely replacing text preference data with multi-modal preference data does not consistently yield positive results; instead, it can exacerbate issues such as hallucinations. To address these issues, we propose to enhance vanilla DPO for our **VidPrompter** model by incorporating multi-type hallucination identification, named **H**allucination-aware **D**irect **P**reference **O**ptimization (HDPO). Specifically, our hallucination-aware enhancements comprehensively consider the following three aspects: (1) **Response hallucination:** The model is required to demonstrate a significant preference for the chosen response over a hallucinated one. (2) **Input hallucination:** We require the model to clearly demonstrate a positive preference for the chosen response under the condition of the original input, while showing a negative preference for the same chosen response under the condition of a new hallucinated input. (3) **Task hallucination:** The model needs to develop a positive preference for the chosen response corresponding to a specific data example, while exhibiting a negative preference when the task is incorrectly replaced. For evaluation, our experiments across each task demonstrate that our method can significantly outperform baseline models and other competitors, showing substantial improvements and advantages.

Our main contributions are as follows:

- We propose a novel model named **VidPrompter**, which improves the prompt generation process through multi-task learning, leading to more detailed and contextually rich outputs.
- We enhance the vanilla DPO method by incorporating multi-aspect hallucination identification, referred to as HDPO, which effectively leverages model capabilities across multiple tasks and reduces output hallucinations.
- We provide a comprehensive evaluation of **VidPrompter**, highlighting the efficacy and versatility of the model. The results demonstrate the superiority of our approach in enhancing the quality and reliability of T2V outputs.

2 Related Work

Recent developments in LMMs have achieved significant milestones. LMMs are typically composed of three essential components: (i) a vision encoder for extracting visual features, (ii) a modality alignment module that incorporates visual features into the language model's embedding space, and (iii) an LLM backbone responsible for decoding multi-modal contexts. When it comes to LMMs designed for video processing, the primary distinction lies in their methods for encoding video into vision tokens compatible with LLMs. Video-LLaMA [Zhang *et al.*, 2023] utilizes a Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] along with an image Q-Former to encode individual frames before applying a video Q-Former for temporal modeling. VideoChat2 [Li *et al.*, 2024] employs a video Transformer for feature encoding and then uses a Q-Former [Li *et al.*, 2023a] to reduce the number of video tokens. TimeChat [Ren *et al.*, 2024] constructs datasets for time-sensitive instruction tuning, embedding timestamp knowledge into visual tokens. Additionally, VTimeLLM [Huang *et al.*, 2024a] proposes a three-stage training method similar to LLaVA. Recently, VILA [Lin *et al.*, 2024] has demonstrated that re-blending text-only instruction data with image-text data during instruction fine-tuning not only alleviates the decline in performance on text-only tasks but also enhances the accuracy of LMM tasks.

After completing the supervised fine-tuning (SFT) phase, a reinforcement learning stage can be introduced to further improve the model's performance or refine it towards specific objectives. Reinforcement learning from human feedback (RLHF) [Christiano *et al.*, 2017; Ouyang *et al.*, 2022] has proven effective in aligning large language models (LLMs) with human values. Another notable approach is DPO [Rafailov *et al.*, 2023], which optimizes LLMs based on human preferences, achieving impressive results without the need for a separate reward model.

Significant efforts have been dedicated to enhancing the efficacy and efficiency of DPO. Techniques such as SimPO [Meng *et al.*, 2024] streamline DPO by eliminating the reference model, which reduces both computational and memory demands. IPO [Azar *et al.*, 2024] addresses the issue of reward overfitting in DPO. Approaches such as KTO [Ethayarajh *et al.*, 2024] and NCA [Chen *et al.*, 2024b] seek to fulfill DPO's requirement for paired preference data by developing optimization goals that leverage unpaired data. Iterative DPO [Xu *et al.*, 2023; Yuan *et al.*, 2024] and SPPO [Wu *et al.*, 2024] advocate for on-policy sampling of preference data, outperforming off-policy DPO methods.

In multi-modal contexts, recent studies have concentrated on generating multi-modal preference data. These efforts include gathering human preferences [Sun *et al.*, 2024; Yu *et al.*, 2024a], deriving preferences from LMMs [Li *et al.*, 2023b; Yu *et al.*, 2024b], and aligning a model's preferences with its outputs [Deng *et al.*, 2024]. In terms of learning objectives, recent works generally focus on DPO tailored to LLMs [Li *et al.*, 2023b; Zhou *et al.*, 2024], or employ reinforcement learning [Sun *et al.*, 2024] and contrastive learning [Jiang *et al.*, 2024]. Our study investigates an often-overlooked but critical issue (i.e., hallucination) within the multi-modal DPO learning process.

3 Methodology

In this section, we first introduce the high-level pipeline of **VidPrompter**. Then, we present the preliminary knowledge

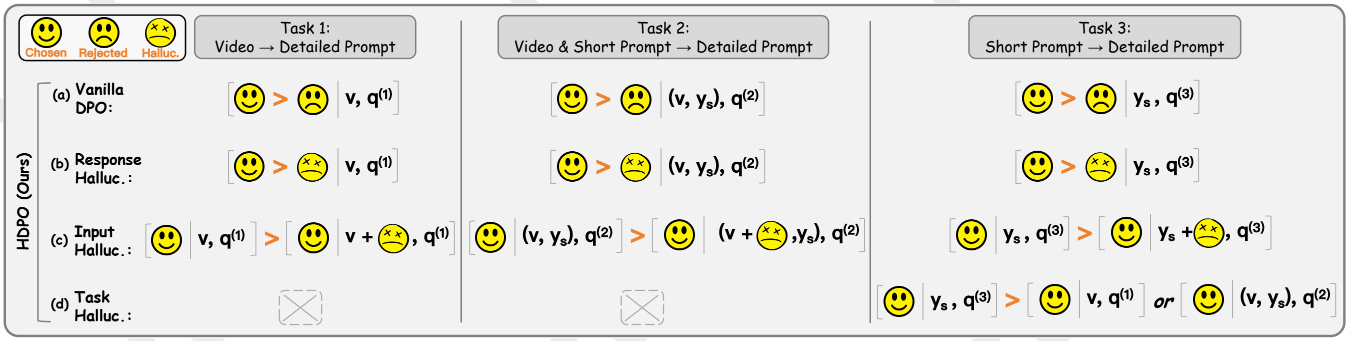


Figure 2: Illustration of our proposed hallucination-aware direct preference optimization method. “Halluc.” is short for “Hallucination”.

of DPO. Next, we propose our Hallucination-aware Direct Preference Optimization (HDPO) method, as shown in Figure 2. Finally, we provide the details of our training and testing resources.

3.1 VidPrompter

VidPrompter is designed to perform three tasks: (1) *Task 1*: generating detailed prompts from raw videos; (2) *Task 2*: producing more comprehensive prompts from raw videos with additional short descriptions; (3) *Task 3*: enriching user-provided simple prompts into detailed prompts. Our model integrates these highly relevant capabilities using a single framework, thereby avoiding redundant training and achieving mutual reinforcement.

For *Task 1*, given a video v and the task-specific instruction q, the model π should generate a detailed prompt y to accurately and comprehensively describe the video v: $y = \pi(v, q)$. For *Task 2*, it additionally uses the short description y_s as information guidance: $y = \pi((v, y_s), q)$. For *Task 3*, it needs to expand on the short prompt y_s based on its own imagination and strive for a comprehensive detailed prompt y, without referencing or relying on any specific video content: $y = \pi(y_s, q)$.

Figure 3 shows the instructions for our different tasks.

3.2 Preliminaries of DPO

Preference optimization aims to align LLMs with human preferences, thereby improving their ability to meet human needs. In our context, it encourages the model to recognize that, for a specific input x (e.g., the video v, the user-provided simple prompt y_s) and a specific instruction q, the good response y_g selected by the evaluator is preferred over the rejected response y_b. A popular approach for achieving this is DPO [Rafailov *et al.*, 2023]. Grounded in reward modeling from RLHF [Ouyang *et al.*, 2022], DPO seeks to maximize the difference in rewards between the chosen response $r(y_g|x, q)$ and the rejected response $r(y_b|x, q)$. More concretely, given a model to optimize, denoted as π_θ, and a reference model π_{ref} which is typically initialized from a SFT model, DPO defines the reward as follows:

$$r(y|x, q) = \beta \log \frac{\pi_\theta(y|x, q)}{\pi_{\text{ref}}(y|x, q)} + Z(x, q), \tag{1}$$

where $Z(x, q)$ is a partition function, β is a hyperparameter that controls the deviation from the reference model. Then,

Task 1: Video → Detailed Prompt

Elaborate on the visual and narrative elements of the video in detail.

Task 2: Video & Short Prompt → Detailed Prompt

Elaborate on the visual and narrative elements of the video in detail. And the following is a brief descriptive information about the video for your reference (which may not always be accurate): {$\mathbf{y_s}$}

Task 3: Short Prompt → Detailed Prompt

Directly expand the following Input (brief description of a video) to Output (detailed description).
1. The Output should be a detailed description in more than three sentences.
2. It should contain description of the main subject actions or status sequence, including the main subjects (person, object, animal, or none) and their attributes, their action, their position, and movements.
3. It should contain summary of the background, include the objects, location, weather, and time.
4. It should contain summary of the view shot, camera movement and changes in shooting angles.
5. It should contain briefly summary of the visual, photographic and artistic style.
6. Do not describe each frame individually. Do not reply with words like 'first frame'. Do not output repetitive, redundant, or too long content. The description should be useful for AI to generate the video.
Input: A woman is walking.
Output: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.
Input: {$\mathbf{y_s}$}
Output:

Figure 3: Instructions for different tasks in **VidPrompter**. y_s indicates the short prompt of the video.

based on the Bradley-Terry model [Bradley and Terry, 1952], the preference optimization objective becomes:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_g|x, q)}{\pi_{\text{ref}}(y_g|x, q)} - \beta \log \frac{\pi_\theta(y_b|x, q)}{\pi_{\text{ref}}(y_b|x, q)} \right), \tag{2}$$

which is essentially equivalent to maximizing the following term: $\sigma(r(y_g|x, q) - r(y_b|x, q))$.

3.3 Hallucination-aware Direct Preference Optimization (HDPO)

In this section, we present HDPO, an enhanced version of DPO specifically designed to address the hallucination issue during multi-modal preference alignment. The hallucinations generated in the responses of LLMs have become a widely recognized issue [Huang *et al.*, 2024b; Jiang *et al.*, 2024; Li *et al.*, 2023c; Guan *et al.*, 2024]. This problem is particularly common when generating longer answers, and it can have serious negative impacts on practical applications, such as in the field of autonomous driving. One possible reason for hallucinations is that the model may over-rely on certain tokens from intermediate summaries during text generation, or it might neglect to use sufficient factual and visual evidence when forming answers [Huang *et al.*, 2024b].

As illustrated in Figure 2, HDPO incorporates three new preference optimization objectives into the vanilla DPO: response hallucination, input hallucination, and task hallucination. The direct application of the original DPO to various tasks is illustrated in Figure 2(a). For example, for *Task 1*, we should optimize to maximize the following expression:

$$\sigma(r(y_g^{(1)}|x^{(1)}, q^{(1)}) - r(y_b^{(1)}|x^{(1)}, q^{(1)})), \tag{3}$$

where the "(1)" in the upper right corner represents the task number. Similarly, we also have:

$$\sigma(r(y_g^{(2)}|x^{(2)}, q^{(2)}) - r(y_b^{(2)}|x^{(2)}, q^{(2)})), \tag{4}$$

$$\sigma(r(y_g^{(3)}|x^{(3)}, q^{(3)}) - r(y_b^{(3)}|x^{(3)}, q^{(3)})). \tag{5}$$

Based on the direct application of these baselines, we propose our hallucination-based improvements.

Response Hallucination

As shown in Figure 2(b), we first introduce response hallucination, which enables the model to perceive and identify hallucinations in the responses. Our preparatory steps involve employing syntactic analysis tools, such as NLTK [Bird *et al.*, 2009] and spaCy [Honnibal *et al.*, 2020], to execute part-of-speech tagging and parse syntactic structures within the chosen response. Then, we selectively replace specific words (*nouns, numerals, colors, terms related to direction, verbs*) with their **semantically disparate counterparts within the same syntactic category** (e.g., *boys* to *girls*, *white* to *blue*, *throwing* to *lifting*). Through this simulated generation approach, we can obtain responses containing hallucinations, which are denoted as g_h. Next, in a manner similar to the vanilla DPO, we aim for the model to undergo the corresponding preference optimization:

$$\sigma(r(y_g^{(1)}|x^{(1)}, q^{(1)}) - r(y_h^{(1)}|x^{(1)}, q^{(1)})), \tag{6}$$

$$\sigma(r(y_g^{(2)}|x^{(2)}, q^{(2)}) - r(y_h^{(2)}|x^{(2)}, q^{(2)})), \tag{7}$$

$$\sigma(r(y_g^{(3)}|x^{(3)}, q^{(3)}) - r(y_h^{(3)}|x^{(3)}, q^{(3)})). \tag{8}$$

Input Hallucination

We also approach the training of the model by adding hallucinations from this innovative perspective. We can consider the model's process of generating detailed prompts as a conditional generation process. The model not only needs to learn to distinguish the quality of different responses under given conditions, but it must also understand that the quality of the same response can vary under different input conditions. This means that a particular response is not always good or bad; rather, it is determined by the given input conditions. Therefore, for *Task 1*, we aim for the model to maximize:

$$\sigma(r(y_g^{(1)}|x^{(1)}, q^{(1)}) - r(y_g^{(1)}|x^{(1)} + \mathcal{H}^{(1)}, q^{(1)})). \tag{9}$$

Here, $\mathcal{H}^{(1)}$ represents the hallucination that we add to the input $x^{(1)}$, specifically the video v. Similarly, for the other two tasks, we have also designed:

$$\sigma(r(y_g^{(2)}|x^{(2)}, q^{(2)}) - r(y_g^{(2)}|x^{(2)} + \mathcal{H}^{(2)}, q^{(2)})), \tag{10}$$

$$\sigma(r(y_g^{(3)}|x^{(3)}, q^{(3)}) - r(y_g^{(3)}|x^{(3)} + \mathcal{H}^{(3)}, q^{(3)})). \tag{11}$$

The degree of hallucinations requires careful consideration. If the hallucinations are too subtle, the modified input conditions may not differ significantly from the original, and asking the model to maximize the reward gap between the two preference sets might have an adverse effect. Conversely, if the hallucinations are too pronounced, it can render the input unreasonable, making it difficult for the model trained on such inputs to perform well on normal inputs.

For the short prompt y_s, we use the same modification approach as in response hallucination, which involves replacing words with semantically disparate ones within the same syntactic category. Meanwhile, for the video v, we randomly shuffle the order of frames in the input video and randomly paste patches from another random frame onto each frame image. We have found that this method can appropriately and moderately add hallucinations to the video to enhance model performance, which will be verified in the experiments.

Task Hallucination

A good response for a data sample of a specific task is most likely not optimal for another task. For example, the long prompt $y_g^{(3)}$ produced by the model based on the user-provided short prompt $y_s^{(3)}$ can be correct, but it may contain many hallucinations when evaluated in the context of describing a certain video v for *Task 1* or *Task 2*. In this regard, we optimize the model to maximize:

$$\sigma(r(y_g^{(3)}|x^{(3)}, q^{(3)}) - r(y_g^{(3)}|x^{(1)}, q^{(1)})), \tag{12}$$

$$\sigma(r(y_g^{(3)}|x^{(3)}, q^{(3)}) - r(y_g^{(3)}|x^{(2)}, q^{(2)})). \tag{13}$$

Here, the model's alignment with relationships between input and output can be strengthened without damaging the input. To summarize, our HDPO integrates Eq. (3) to Eq. (13) as the maximization goal. These objectives work together to ensure that the LMM can capture preferences based on the multimodal clues.

3.4 Resource

We randomly select 60K videos from the Panda-70M [Chen *et al.*, 2024d] dataset to form the training set, and 1K videos for the testing set. The dataset initially includes high-quality, short prompts generated by several teacher models and a fine-tuned prompt selection model. Subsequently, we utilize the advanced VILA-34B [Lin *et al.*, 2024] model to generate detailed video captions, which serve as the selected responses for *Task 1* and *Task 2*. VILA-34B has achieved the top ranking among open-source LMMs on the authoritative Video-MME [Fu *et al.*, 2024] leaderboard (as of July 2024). Additionally, we employ Llama-3-70B [Touvron *et al.*, 2023a; Touvron *et al.*, 2023b] to generate detailed prompts for *Task 3*. Finally, we adopt the SFT model as the baseline and implement a beam search to sample diverse responses for all tasks. We also incorporate additional evaluator models to assess whether the responses should be labeled as "chosen" or "rejected".

3.5 Evaluator

CLIP-style models [Radford *et al.*, 2021] are often effective in evaluating the degree of match between a video and its corresponding prompt. Consequently, we select the widely-used CLIP [Radford *et al.*, 2021] model along with the promising Long-CLIP [Zhang *et al.*, 2024] and VideoCLIP-XL [Wang *et al.*, 2024b] models, which aim to enhance the capability for processing long prompts for images and videos, respectively. Since the CLIP and Long-CLIP models are originally designed for images, we utilize the averaged feature from multiple frames to represent the overall feature of the video. At this stage, we have obtained all three evaluators, and we use their average similarity scores for the final evaluation to identify the "rejected responses".

4 Experiments

4.1 Implementation Details

We select VILA-1.5-8B [Lin *et al.*, 2024] as our backbone. For the sampling of rejected responses, we set the beam search width to 5 and select two samples with the lowest average evaluator scores as the rejected responses. In the HDPO training, we set β to 0.1, the batch size to 64, and the learning rate to 1e-6. We conduct HDPO training for 2 epochs on the training set using 8 NVIDIA A100 GPUs.

4.2 Experimental Settings

We have selected the following commonly used, robust, and competitive methods and models for a fair comparison in our application scenario: (1) ShareCaptioner-Video [Chen *et al.*, 2024c] is fine-tuned on self-collected video caption data. For flexible usage, it is designed with consideration for application scenarios similar to our *Task 1* and *Task 3*. (2) VILA-1.5-8B [Lin *et al.*, 2024] serves as our selected baseline model. (3) SFT indicates that we directly fine-tune the baseline model with the training set, utilizing the chosen responses as ground truth. This approach allows us to compare whether reinforcement learning-based methods yield more significant improvements over simple SFT. (4) DPO denotes the vanilla DPO method. (5) Hinge-DPO [Liu *et al.*, 2023] proposes using a hinge loss instead of the sigmoid function for optimization. (6) KTO [Ethayarajh *et al.*, 2024] defines the loss function entirely in terms of individual examples that are labeled as "good" or "bad". (7) IPO [Azar *et al.*, 2024] adds a regularization term to the DPO loss, enabling the training of models to convergence without the need for techniques such as early stopping. (8) mDPO [Wang *et al.*, 2024a] designs a multi-modal DPO objective that prevents the over-prioritization of language-only preferences by also optimizing image preferences.

For *Task 1* and *Task 2*, to evaluate the consistency between the generated detailed captions and the input video content, we employ CLIP [Radford *et al.*, 2021], Long-CLIP [Zhang *et al.*, 2024], and VideoCLIP-XL [Wang *et al.*, 2024b] to assess the video-caption similarity scores. These are referred to as the "CLIP Score", "Long-CLIP Score", and "VideoCLIP-XL Score", respectively. For *Task 3*, we utilize the powerful Llama-3-70B [Touvron *et al.*, 2023a; Touvron *et al.*, 2023b] to evaluate the qualities of expanded detailed captions/prompts (denoted as the "Llama-3-70B Eval Score") from multiple perspectives. The prompt template is:

Description: {Detailed Caption}

Please evaluate the quality of video description displayed above. A perfect (full-marks) video description needs to meet the following factors:

1. It should be a detailed description in more than three sentences.

2. It should contain description of the main subject actions or status sequence, including the main subjects (person, object, animal, or none) and their attributes, their action, their position, and movements.

3. It should contain summary of the background, include the objects, location, weather, and time.

4. It should contain summary of the view shot, camera movement and changes in shooting angles.

5. It should contain brief summary of the visual, photographic and artistic style.

6. It should be natural and coherent, rather than weird or confused.

7. It should NOT describe each frame individually. It should NOT contain repetitive, redundant, or too long content.

Your evaluation should consider these factors. You should give an overall score on a scale of 0 to 10, where a higher score indicates better overall quality. Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias. Then, output the score with the following format: Evaluation evidence: <your evaluation explanation here> Score: <score>

It is important to note that the scores for CLIP-style evaluation metrics typically range from 20 to 30, based on actual usage feedback. In contrast, the evaluation scores obtained from Llama-3-70B fall within the range of 0 to 10.

4.3 General Comparison

The experimental results are presented in Table 1, from which we can draw the following insights: (1) Our method clearly outperforms ShareCaptioner-Video, which is the state-of-the-art (SOTA) captioner designed for T2V applications. (2) The baseline model VILA-1.5-8B demonstrates poor performance, particularly for *Task 3*, as it was initially designed for multi-modal tasks. (3) The performance of hinge-DPO on *Task 1* and *Task 2* is comparable to that of DPO, while KTO and IPO show significantly poorer performance. In *Task 3*, DPO exhibits substantial superiority over hinge-DPO, KTO,

Method	*Task 1*			*Task 2*			*Task 3*
	CLIP Score	Long-CLIP Score	VideoCLIP-XL Score	CLIP Score	Long-CLIP Score	VideoCLIP-XL Score	Llama-3-70B Eval Score
ShareCaptioner-Video [Chen *et al.*, 2024c]	26.0	25.5	24.2	-	-	-	7.39
VILA-1.5-8B [Lin *et al.*, 2024]	28.2	26.3	25.0	29.0	26.6	24.9	4.92
+ SFT	28.8	26.5	25.3	28.8	26.6	25.3	8.47
+ DPO [Rafailov *et al.*, 2023]	28.7	26.8	26.5	29.1	26.8	26.5	8.56
+ Hinge-DPO [Liu *et al.*, 2023]	28.8	26.8	26.7	29.3	26.9	26.6	8.21
+ KTO [Ethayarajh *et al.*, 2024]	28.2	26.3	26.1	28.3	26.3	26.0	8.16
+ IPO [Azar *et al.*, 2024]	27.3	25.7	25.2	26.5	25.4	25.0	8.25
+ mDPO [Wang *et al.*, 2024a]	27.9	26.8	26.7	28.8	26.9	26.2	8.59
+ HDPO (Ours)	**29.7**	**27.2**	**27.3**	**30.1**	**27.3**	**27.4**	**8.90**

Table 1: Performance comparison on the testing set with competitors.

Method	*Task 1*	*Task 2*	*Task 3*
	Averaged CLIP Score	Averaged CLIP Score	Llama-3-70B Eval Score
VILA-1.5-8B (Baseline)	26.5	26.8	4.92
DPO	27.3	27.5	8.56
+ Response Hallucination			
1-3 words	27.4	27.5	8.58
1-4 words	27.7	27.6	8.62
1-5 words	27.8	27.8	8.62
1-6 words	27.9	27.7	8.60
+ Input Hallucination			
Random Crop	27.2	27.4	8.56
Random Crop and Paste (1/4-1/3)	27.5	27.6	8.57
Random Crop and Paste (1/3-1/2)	27.7	27.8	8.59
Random Crop and Paste (1/2-1)	27.6	27.6	8.59
Shuffle Frame	27.5	27.7	8.58
+ Task Hallucination	27.8	28.0	8.64
HDPO + Separate-Training	27.9	27.6	**8.90**
HDPO (Ours)	**28.1**	**28.3**	**8.90**

Table 2: The ablation results of various components in our method. The averaged CLIP score is the arithmetic mean of CLIP, Long-CLIP and VideoCLIP-XL scores.

and IPO. This suggests that different training methods may exhibit varying strengths in pure text and multi-modal tasks. (4) Our method significantly surpasses mDPO, which is also tailored for multi-modal scenarios. This advantage arises because our approach incorporates hallucination-aware optimization strategies from a more comprehensive perspective. (5) *Task 2* provides additional support with short captions compared to *Task 1*. However, the majority of methods fail to achieve consistent improvements across all CLIP scores in *Task 2* relative to *Task 1*. In contrast, our method demonstrates relatively consistent enhancements in *Task 2* compared to *Task 1* across all CLIP scores.

4.4 Ablation Study

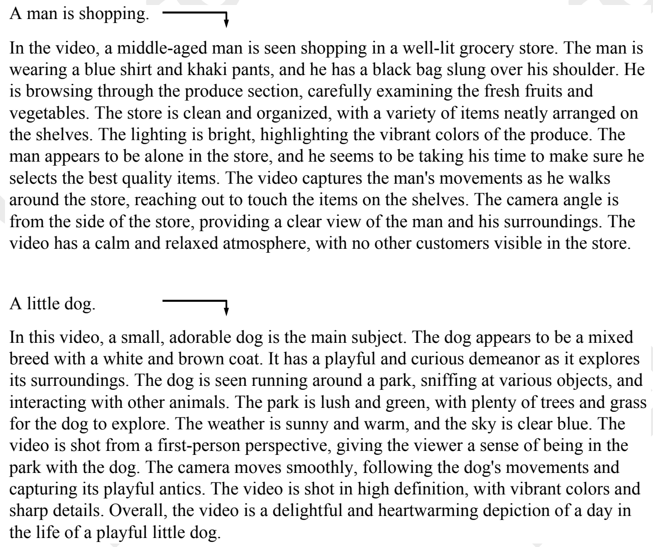
To verify the effectiveness of the components of our method, we conduct ablation experiments, as shown in Table 2. The results indicate several key findings. (1) The vanilla DPO training demonstrates a significant improvement compared to the baseline model. (2) Building upon DPO, the incorporation of **Response Hallucination**, **Input Hallucination**, or **Task Hallucination** each yields notable enhancements in performance. (3) For **Response Hallucination**, we experiment with randomly replacing 1 to n words in the original captions to generate hallucinations. The optimal performance was achieved by expanding the replacement range from 1-3 to 1-5, with the performance at 1-6 being generally comparable to that at 1-5. Thus, we ultimately settle on a range of 1-5. (4) In the context of **Input Hallucination**, we evaluate various methods and degrees of video hallucination strategies. The Random Crop technique retains between 80% to 100% of each frame, preserving both height and width dimensions. This approach results in minimal content modification and is associated with poorer performance. Conversely, the Random Crop and Paste (1/4-1/3) method involves pasting a segment from another frame at a random location within each frame, covering between 1/4 and 1/3 of the length and width dimensions. This approach significantly alters the content of the video, improving the model's ability to learn the relationship between conditional information and responses. The observed performance improvements with varying modification levels further reinforce our earlier assertion that the method and degree of adding hallucinations necessitate careful study. (5) Additionally, within **Input Hallucination**, we find that randomly shuffling the frames of the video can yield significant effects, allowing the model to learn how variations in the temporal information of the input video influence its responses. (6) Lastly, the separate training approach (which trains independent models for each task) demonstrates that mixed training improves the metrics for *Task 1* and *Task 2*, while maintaining performance on *Task 3*. This partially validates the notion that incorporating pure text data into multimodal tasks can enhance the model's robustness and cross-modal alignment. Overall, we prove that each of the components proposed in this paper shows positive effects in improving the performance of the **VidPrompter** model.

4.5 Evaluation for Prompt Beautification

To better evaluate the performance of our model on *Task 3* (Short Prompt to Detailed Prompt), we use VBench [Huang *et al.*, 2024c] to assess the quality of T2V generation results. We randomly sample 100 short captions from the test set, enhance them into detailed captions using our model, and then provide both sets to the T2V model EasyEnimate [Xu *et al.*, 2024] to obtain synthetic videos. Finally, we conduct a multi-dimensional evaluation of the generated content, as shown in Table 3. From the comparison, it can be observed that the detailed prompts generated by our model can create more vivid videos, exhibiting higher motion smoothness and

Dimension	Original	Beautified (Ours)
Motion Smoothness	0.992	**0.993**
Dynamic Degree	0.24	**0.36**
Aesthetic Quality	0.526	**0.529**
Imaging Quality	63.8	**64.3**
Subject Consistency	*0.972*	*0.968*
Background Consistency	*0.961*	*0.958*

Table 3: Evaluation for *Task 3* (Short Prompt to Detailed Prompt) using VBench.



A man is shopping. →

In the video, a middle-aged man is seen shopping in a well-lit grocery store. The man is wearing a blue shirt and khaki pants, and he has a black bag slung over his shoulder. He is browsing through the produce section, carefully examining the fresh fruits and vegetables. The store is clean and organized, with a variety of items neatly arranged on the shelves. The lighting is bright, highlighting the vibrant colors of the produce. The man appears to be alone in the store, and he seems to be taking his time to make sure he selects the best quality items. The video captures the man's movements as he walks around the store, reaching out to touch the items on the shelves. The camera angle is from the side of the store, providing a clear view of the man and his surroundings. The video has a calm and relaxed atmosphere, with no other customers visible in the store.

A little dog. →

In this video, a small, adorable dog is the main subject. The dog appears to be a mixed breed with a white and brown coat. It has a playful and curious demeanor as it explores its surroundings. The dog is seen running around a park, sniffing at various objects, and interacting with other animals. The park is lush and green, with plenty of trees and grass for the dog to explore. The weather is sunny and warm, and the sky is clear blue. The video is shot from a first-person perspective, giving the viewer a sense of being in the park with the dog. The camera moves smoothly, following the dog's movements and capturing its playful antics. The video is shot in high definition, with vibrant colors and sharp details. Overall, the video is a delightful and heartwarming depiction of a day in the life of a playful little dog.

Figure 4: Qualitative examples for prompt beautification.

dynamic degree, with a slight decrease in subject and background consistency. These two consistency metrics are not always better when higher, as completely static videos can achieve high consistency, but this is not what we aim for. In addition, the detailed prompts can generate more visually appealing videos, as reflected in the improvements seen in aesthetic and image quality. Examples can be seen in Figure 4.

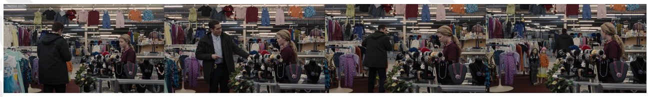
4.6 Case Study of Video Captioning

Figure 5 shows qualitative examples of detailed video captions generated by ShareCaptioner-Video [Chen *et al.*, 2024c] and our HDPO method. In example (a), we observe that ShareCaptioner-Video inaccurately describes the two characters in the video as appearing one after the other, whereas they actually move alongside each other. In contrast, our model provides a more accurate description, including the two characters passing by others who are also wearing uniforms. In example (b), ShareCaptioner-Video offers a general overview of the video content but overlooks critical details. For instance, the two characters are not merely engaged in conversation; specifically, the man is making a payment to the woman. Additionally, the description only mentions that there are two individuals without clarifying their genders. In contrast, our model accurately describes both the man and the woman, noting the payment detail as well. Overall, our description is more precise and concise.

ShareCaptioner-Video:
The video captures a sequence of events within a formal indoor setting, likely a government or institutional building, ... Initially, the focus is on a man in a dark suit, who is seen walking through a corridor, ... As the video continues, the man in the dark suit is joined by another individual, also in formal attire, who appears to be walking alongside him. ... The interaction between the two individuals becomes more pronounced as they engage in a handshake, indicating a greeting or agreement. This handshake is brief, as the individuals soon part ways, continuing their respective journeys down the corridor. The video concludes with the individuals continuing their separate paths, with one moving further down the corridor and the other turning slightly, indicating a change in direction. ... the formal interactions and movements of the individuals within the ceremonial setting.

HDPO (Ours):
The video opens with a scene of two men walking side by side down a hallway. The man on the left is dressed in a dark suit and tie, ... As they walk, they pass by several other men who are standing in the hallway, dressed in suits and uniforms. The man on the left seems to be checking his watch, while the man on the right is possibly holding a document or a device in his hands. ... ends with the two men seemingly still talking in the hallway.

(a)



ShareCaptioner-Video:
The video captures a scene inside a clothing store, where two individuals are engaged in a conversation. The person on the left, dressed in a dark coat and a lighter-colored shirt, stands with their back to the camera, facing the person on the right. This individual, wearing a maroon top, is positioned in front of a clothing rack filled with various garments. ... suggesting a static shot that captures the ongoing conversation between the two individuals.

HDPO (Ours):
In the video, a man and a woman engaged in a transaction or a conversation. The man is standing and appears to be handing over a card or a piece of paper to the woman, who seems to be an employee or a vendor. The man is wearing a black coat, and the woman is dressed in a red top and is surrounded by an array of items that are likely for sale or display, as suggested by the numerous mannequins and garments in the background. There are no other customers visible in the immediate vicinity of the man and woman.

(b)

Figure 5: Qualitative examples of ShareCaptioner-Video and our model for video captioning.

5 Conclusion

In conclusion, we propose **VidPrompter** to optimize prompt generation for T2V models. By integrating its multi-modal capabilities, **VidPrompter** enhances both the training and inference processes, enabling T2V models to produce higher-quality outputs. Our experimental results offer valuable resources for research and development in this rapidly evolving field, thereby contributing to broader applications of T2V.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China (Grant No.:62476093). It is also partially supported by Alibaba Cloud Computing, through Research Talent Program with South China University of Technology.

References

[Azar *et al.*, 2024] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, pages 4447–4455, 2024.

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

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

[Brooks *et al.*, 2024] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[Chen *et al.*, 2024a] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024.

[Chen *et al.*, 2024b] Huayu Chen, Guande He, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards. *NeurIPS*, 2024.

[Chen *et al.*, 2024c] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. ShareGPT4Video: Improving video understanding and generation with better captions. *NeurIPS*, 37:19472–19495, 2024.

[Chen *et al.*, 2024d] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70M: Captioning 70M videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331, 2024.

[Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017.

[Deng *et al.*, 2024] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *NeurIPS*, 37:131369–131397, 2024.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Ethayarajh *et al.*, 2024] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *ICML*, 2024.

[Fu *et al.*, 2024] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[Guan *et al.*, 2024] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusion-Bench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385, 2024.

[Guo *et al.*, 2024] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

[Honnibal *et al.*, 2020] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. 2020.

[Huang *et al.*, 2024a] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *CVPR*, pages 14271–14280, 2024.

[Huang *et al.*, 2024b] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427, 2024.

[Huang *et al.*, 2024c] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024.

[Jiang *et al.*, 2024] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*, pages 27036–27046, 2024.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[Li *et al.*, 2023b] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.

[Li *et al.*, 2023c] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023.

[Li *et al.*, 2024] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024.

[Lin *et al.*, 2024] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024.

[Liu *et al.*, 2023] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *ICLR*, 2023.

[Meng *et al.*, 2024] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. *NeurIPS*, 37:124198–124235, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36, 2023.

[Ren *et al.*, 2024] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024.

[Sun *et al.*, 2024] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang Yan Gui, Yu Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented RLHF. In *Findings of ACL*, pages 13088–13110, 2024.

[Touvron *et al.*, 2023a] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Touvron *et al.*, 2023b] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2024a] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *EMNLP*, pages 8078–8088, 2024.

[Wang *et al.*, 2024b] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. VideoCLIP-XL: Advancing long description understanding for video CLIP models. In *EMNLP*, pages 16061–16075, 2024.

[Wu *et al.*, 2024] Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *ICML Workshops*, 2024.

[Xu *et al.*, 2023] Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023.

[Xu *et al.*, 2024] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. EasyAnimate: A high-performance long video generation method based on Transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.

[Yu *et al.*, 2024a] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *CVPR*, pages 13807–13816, 2024.

[Yu *et al.*, 2024b] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. RLAIF-V: Aligning MLLMs through open-source AI feedback for super GPT-4V trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[Yuan *et al.*, 2024] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, pages 543–553, 2023.

[Zhang *et al.*, 2024] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *ECCV*, pages 310–325, 2024.

[Zheng *et al.*, 2024] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all, 2024.

[Zhou *et al.*, 2024] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference finetuning. In *ICLR Workshops*, 2024.