Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward

Anonymous authors
Paper under double-blind review

Abstract

Preference modeling techniques, such as direct preference optimization (DPO), has shown effective in enhancing the generalization abilities of large language model (LLM). However, in tasks involving video instruction-following, providing informative feedback, especially for detecting hallucinations in generated responses, remains a significant challenge. Previous studies have explored using large large multimodal models (LMMs) as reward models to guide preference modeling, but their ability to accurately assess the factuality of generated responses compared to corresponding videos has not been conclusively established. This paper introduces a novel framework that utilizes detailed video captions as a proxy of video content, enabling language models to incorporate this information as supporting evidence for scoring video Question Answering (QA) predictions. Our approach demonstrates robust alignment with OpenAI GPT-4V model's reward mechanism, which directly takes video frames as input. Furthermore, we show that applying this tailored reward through DPO significantly improves the performance of video LMMs on video QA tasks.

1 Introduction

This paper addresses the challenge of aligning LMMs, particularly in tasks that involve video instruction following. Despite recent advancements in reinforcement learning (RL) (Ouyang et al., 2022; Bai et al., 2022; Lee et al., 2023; Sun et al., 2023b) and DPO (Rafailov et al., 2024; Chen et al., 2024b; Hosseini et al., 2024), which have been effective in guiding LLMs towards generating more honest, helpful, and harmless content, their effectiveness in multimodal contexts remains limited. The critical obstacle lies in developing a robust reward system capable of distinguishing preferred responses from less preferred ones, especially when such responses are generated based on video inputs. The challenge is further complicated by the presence of hallucinations in generated content, stemming from the scarcity of alignment data across different modalities (Liu et al., 2023b; Sun et al., 2023a).

While human preference data is valuable, it is challenging to scale due to its cost and labor-intensive nature, as highlighted by the LLaVA-RLHF (Sun et al., 2023a) paper, which collected 10k human-evaluated instances at a considerable cost of $3000. Existing approaches for distlling preferences, such as those for image data using GPT-4V (Li et al., 2023d), encounter scalability issues, especially for video inputs that require analyzing multiple frames. While Ahn et al. (2024) leverage a supervised finetuning (SFT) model for self-evaluation, the efficacy of the SFT model remains uncertain, particularly in accurately assessing the factuality of responses in relation to their corresponding videos.

To tackle the aforementioned challenges, we introduce a cost-effective reward mechanism aimed at reliably evaluating the quality of responses generated by video (LLMs), serving as a basis for further preference optimization. We propose the use of detailed video captions as a proxy for video content, enabling a language model analyze video content and assess the accuracy of an LMM's response to a related question and determine the presence of hallucinations. The language model provides natural language feedback as a chain-of-thought step, and generates a numerical score for reward, facilitating a cost-effective feedback system.

However, high-quality video captions are essential for this process. To mitigate the shortage of high-quality video captions, we have developed a comprehensive video caption dataset, SHAREGPTVIDEO, using a novel prompting technique with the GPT-4V model, comprising 900k captions that encompass a wide range of video content, including temporal dynamics, world knowledge, object attributes, and spatial relationships. With this video caption dataset available, we verify that our reward mechanism, which utilizes video captions as a proxy, is well-aligned with evaluations derived from the more powerful, albeit costlier, GPT-4V model-generated rewards. Employing this reward mechanism as the basis for DPO algorithm, we train LLAVA-HOUND-DPO that achieves an 8.1% accuracy improvement over the SFT counterpart. This marks a significant advancement in video LMM alignment and represents the first successful application of a DPO method in this domain.

Our contributions are outlined as follows:

1. We develop a large-scale, detailed video caption dataset, covering a wide array of content. This dataset serves as a foundational resource for LMM model training and research, facilitating advancements in video understanding tasks.

2. We introduce a cost-effective method for evaluating video instruction-following tasks, serving as enhanced evaluation of model performance.

3. We demonstrate the effective application of DPO to improve model performance by leveraging the language model feedback as reward, which substantially improves the alignment of video LMM, establishing a new benchmark for SOTA performance in video QA tasks.

2 Related Work

2.1 Large Multi-Modal Models

LMMs (Liu et al., 2023b;a; Bai et al., 2023; Chen et al., 2023; Li et al., 2023a) have enabled instruction following across modalities by utilizing LLM as backbones. In the context of video understanding, LLMs have been adapted to process video content (Lin et al., 2023a; Zhang et al., 2023a; Maaz et al., 2023; Li et al., 2023b; Luo et al., 2023; Liu et al., 2023c; Jin et al., 2024; Ahn et al., 2024). Our work adots Video-LLaVA backbone, focusing on model enhancement through preference modeling with the DPO technique.

2.2 Video-text Datasets

Existing video-text datasets typically provide brief sentences or mere keywords as captions, as indicated by Bain et al. (2021); Wang et al. (2023); Yu et al. (2019); Jang et al. (2017); Xu et al. (2016). Shvetsova et al. (2023) uses automatic speech recognition to extract textual content from videos, but it encounters alignment issues when the audio does not match or is absent from the visual content. Video-ChatGPT (Li et al., 2023b) employs human effort to create high-quality video instructions, albeit limited to the ActivityNet domain with only 100k instruction pairs. Our work leverages the GPT-4V model with specifically crafted prompts to produce detailed video captions as community resource for LMM training.

2.3 Preference Modeling for LMMs

Preference modeling techniques are employed to enhance the utility of LMMs while mitigating the issue of hallucination. Sun et al. (2023a) leveraged Reinforcement Learning with Human Feedback (RLHF) and incorporated caption information into the reward model to improve the assessment of factuality. More recently, Ahn et al. (2024) used RL on AI feedback to improve video LMM performance. For the image understanding, Li et al. (2023d); Gunjal et al. (2023) introduced the application of DPO on the distilled rewards from GPT-4V on a group of model outputs, while Zhao et al. (2023) created preference data using ChatGPT to generate positive and negative pairs informed by detailed image descriptions. Our contribution extends DPO to the video LMM alignment, with the use of detailed captions as factual evidence for reward modeling.
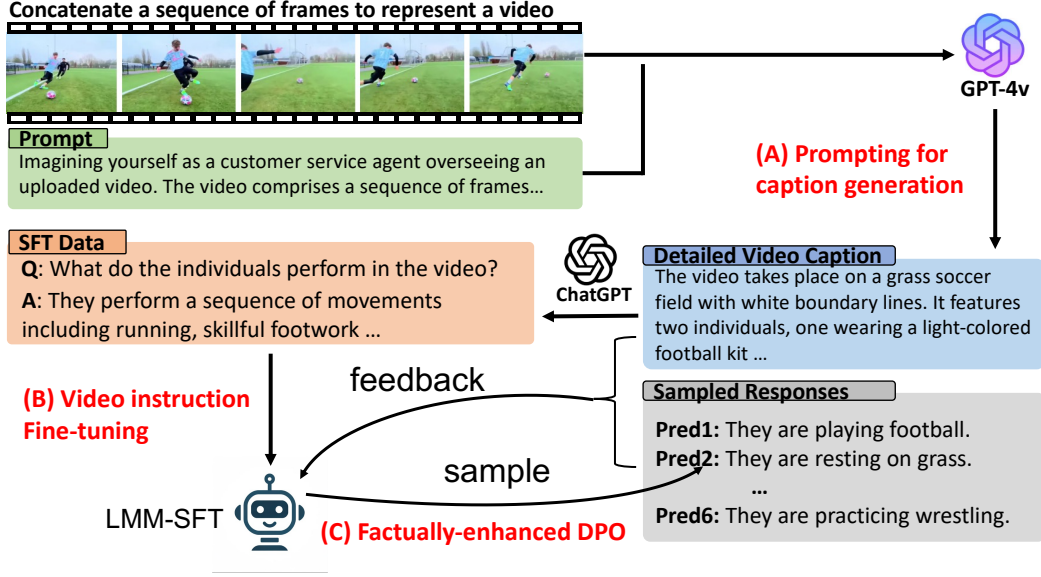
Figure 1: Workflow diagram showing: a) the use of GPT-4V for creating a detailed caption dataset for videos; b) generating video instruction data for SFT; c) integrating captions into a feedback loop for factually-enhanced DPO, improving the model's performance on video instruction-following tasks.

3 Method

As shown in fig. 1, our methodology enhances video LMM alignment through DPO method using rewards from a language model. We elaborate on constructing a video caption dataset in section 3.1. Subsequently, in section 3.2, we discuss the generation of video instruction data and the fine-tuning process of our model. Lastly, section 3.3 details the incorporation of generated captions as a feedback mechanism for DPO method to refine our model's factual alignment in video instruction-following tasks.

3.1 Prompting GPT-4V Model for Detailed Video Caption Distillation

The selection of dataset includes videos from three sources: the WebVid and VIDAL datasets, which are general domain videos sourced from YouTube with 400k and 450k sampled videos respectively, and the ActivityNet dataset, which adds 50k videos focusing on human activities. The three datasets together result in a comprehensive collection of 900k videos. To accommodate the requirement that GPT-4V only takes images as input, we preprocess videos by uniformly extracting ten frames per video content. These frames are then concatenated into a sequence to serve as a proxy for the video. This sequence is the input into GPT-4V to generate a coherent caption for the represented video based on the frame sequence. The prompt adheres to guidelines covering temporal dynamics, world knowledge, object attributes, spatial relationships, aesthetic assessments, etc., with the goal of comprehensively understanding the video contents.

3.2 SFT with Generated Video Instruction Data from Detailed Caption

To generate video instruction-following data for SFT, we adopt a similar methodology outlined in Video-ChatGPT (Li et al., 2023b). Specifically, we first randomly sample 20k, 30k, 30k captions in our dataset from ActivityNet, WebVid and VIDAL respective and then employ ChatGPT to generate three question-answer pairs given each detailed video caption, resulting in a total of 240k instruction data for finetuning. This approach ensures that the instructional data remains factually consistent with the content of the detailed captions. The specific prompting strategy used for this instruction generation process is detailed in fig. 13.

(A) Sample Multiple Outputs from LLM with Temperate=1.0

Query
What is the second symbol drawn on the paper?

LMM-SFT

Sampled Output No. 1
The second symbol is a pound sign.

Sampled Output No. 2
The second symbol that is drawn is a dollar sign.

...

Sampled Output No. 6
The second symbol that is drawn on the blank piece of paper is "¥" which stands for Japanese Yen.

(B) Language-based Feedback from ChatGPT as Reward

```
Given the following inputs:
1. **Ground Truth Video Caption**: {caption}
2. **Question Related to the Caption**: {query}
3. **Ground Truth Answer**: {answer}
4. **Model Predicted Answer**: {sampled_output}

Follow the guidelines to generate reward …
```

ChatGPT

Explanation: In the caption of the video, the second symbol drawn is a Japanese Yen, so the "dollar sign" in the model prediction is not accurate ….
Reward: 2/5

(C) Build Preference Dataset (highest paired with lowest, skip if all ≥ 3 or all <3)

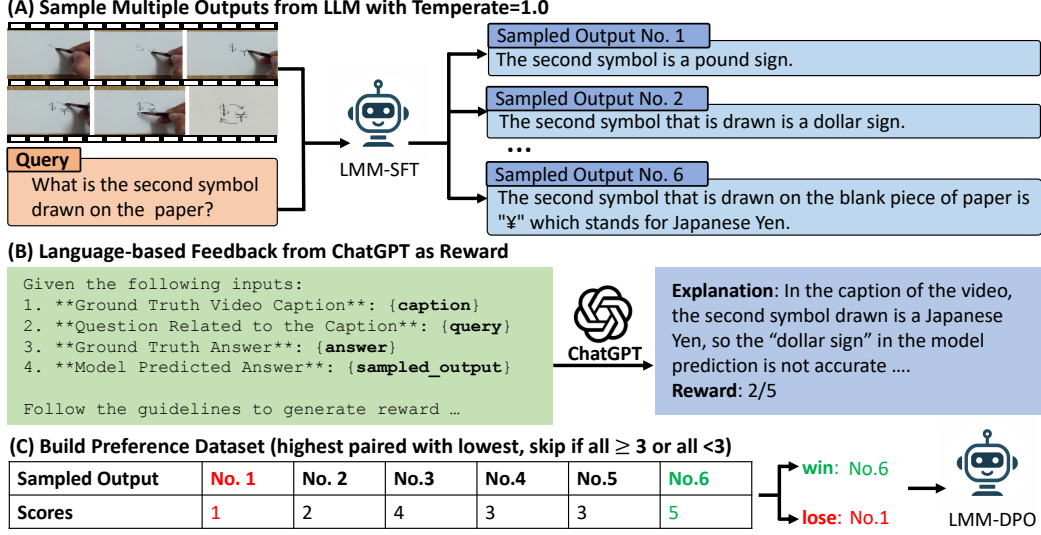| Sampled Output	No. 1	No. 2	No.3	No.4	No.5	No.6
Scores	1	2	4	3	3	5

win: No.6
lose: No.1

LMM-DPO

Figure 2: Detailed illustration of the proposed factually-enhanced DPO method.

3.3 DPO with Feedback from Language Model as Reward

Acquiring high-quality preference data is both costly and labor-intensive. Although GPT-4V is an effective model for reward distillation, its high cost, slow performance, and limited accessibility hinder scalability, especially for video inputs with multiple frames. We propose a cost-efficient method to generate reward data for DPO using detailed video captions as supporting evidence, as shown in fig. 2.

Initially, we randomly select a subset of 20k instruction pairs from the dataset described in section 3.2. The SFT model uses these sampled questions and their corresponding videos to generate six responses per input pair at a temperature of 1.0. This procedure results in 120k question-answer pairs, which will be evaluated. Subsequently, we employ ChatGPT to process inputs including a question, the ground truth answer, the model's prediction, and a detailed description serving as supportive evidence, with the prompt in fig. 15. This generates an output that includes a natural language explanation as chain-of-thought step, followed by a numerical reward score on a scale from 1 to 5, indicating the level of factual alignment and overall quality.

For each video and question pair, we randomly select an answer with a score ≥ 3 as positive example, and an answer with a score below 3 as negative. Cases where all responses are uniformly scored above or below 3 are excluded from the dataset. After the selection process, approximately 17k training instances are compiled for DPO training. Formally, the dataset is denoted as $\mathcal{D}_{DPO} = \{(\mathcal{V}, x, y_w, y_l)\}$, where $\mathcal{V}$ is the video, x is the question, y_w and y_l are the positive and negative responses. The DPO objective is defined as below:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(\mathcal{V}, x, y_w, y_l) \sim \mathcal{D}_{DPO}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x, \mathcal{V})}{\pi_{\text{ref}}(y_w | x, \mathcal{V})} - \beta \log \frac{\pi_\theta(y_l | x, \mathcal{V})}{\pi_{\text{ref}}(y_l | x, \mathcal{V})} \right) \right],$$

where π_θ is the policy model to be optimized and π_{ref} is the base reference model, both models are initialized with SFT weights. σ is the logistic function and β is set to 0.1.

Our approach to reward assignment leverages detailed captions as a proxy for video frames, offering both cost-effectiveness and efficiency. This method incurs costs of less than $20, under a pricing model of $1.5 per million tokens. In comparison, previous methods of preference data collection, such as in Sun et al. (2023a), required an expenditure of $3,000 to gather 10k human preference data points. Additionally, the method proposed by Li et al. (2023d), which employs GPT-4V for reward data labeling, incurs a significantly higher cost—$30 per million tokens—and demonstrates considerably slower inference speeds.
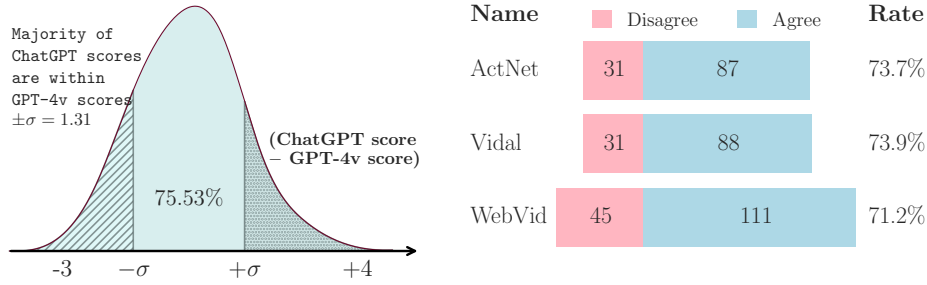
Figure 3: Assessing Evaluator Quality Using Captions in Place of Frames. The left figure shows the distribution of evaluation score differences between ChatGPT (with caption as proxy) and GPT-4V (directly on frames) evaluations. The right figure shows the rate of preference agreement between ChatGPT and GPT-4V as evaluators.

4 Assessment of Evaluator with GPT-4V Caption as Evidence

To assess the effectiveness of our proposed reward assignment method, which utilizes detailed captions as a proxy of actual video frames, we conducted a comparative analysis with the GPT-4V, used as a video QA evaluator. The latter reward system employs GPT-4V evaluation directly taking in video frames, a question, and the model prediction as inputs, with detailed prompt in fig. 16. Both reward systems follow the same set of guidelines for scoring reward.

To compare the two methods, we sample 200 videos from each of the WebVid, VIDAL, and ActivityNet datasets, each associated with one question and two model predictions from our SFT model, with one preferred and one dispreferred by ChatGPT. This results in 1,200 examples, for which we used GPT-4V (with the "gpt-4-vision-preview" version) version to assign scores. Filtering through the Azure API backend resulted in 196, 151, and 143 videos from each dataset, respectively, having both answers evaluated. The average scores of all examples from ChatGPT and GPT-4V evaluations were 2.9 and 3.5 respectively, indicating a tendency of GPT-4V to yield slightly positive evaluations. The Pearson Correlation Coefficient (PCC) of 0.47 ($p < 0.01$) suggests a moderate positive correlation. In fig. 3 (left), the distribution of the difference between ChatGPT and GPT-4V scores reveals that majority ($> 75\%$) of ChatGPT scores fall within one standard deviation ($\sigma = 1.31$) of GPT-4V scores. Additionally, in fig. 3 (right), the agreement on preference between ChatGPT and GPT-4V, excluding ties, exceeded 70%. These findings cautiously support our benchmark's applicability in video QA evaluation. Further refinements for better alignment—such as incorporating Likert scales Zhou et al. (2023) or GPT-4 evaluation—are areas for future research.

5 Experimental Results

5.1 Model Architecture, Image Data Mix-up and Training Pipelines

We adopt Video-LLaVA (Lin et al., 2023a) as the backbone of our video LMM, but our dataset and method can be applied to any other architectures as well. Specifically, Video-LLaVA employs LanguageBind (Zhu et al., 2023) encoder for image and video frame inputs, a MLP projector with 2 fully connected layers to map visual embeddings into text space, and Vicuna Chiang et al. (2023) as large language model. During training, we first initialize the projection MLP layer with the same Video-LLaVA MLP weight. Then we follow the training stages below:

Caption Pre-training Stage (LLAVA-HOUND-PT): At pretraining stage, we use captioning data including 650k image caption data from ALLaVA (Chen et al., 2024a) and our distilled 900k video caption. We freeze the LanguageBind visual encoder and fine-tune the MLP projector and LLM, with learning rate 2e-5 and batch size 128.

| Methods | LLM Size | Existing Video QA Benchmark from Maaz et al. (2023) | | | | | |
| | | MSVD-QA | | MSRVTT-QA | | TGIF-QA | |
		Acc.	Score	Acc.	Score	Acc.	Score
FrozenBiLM (Yang et al., 2022)∗	1B	32.2	-	16.8	-	41.0	-
VideoLLaMA (Zhang et al., 2023a)∗	7B	51.6	2.5	29.6	1.8	-	-
LLaMA-Adapter (Zhang et al., 2023b)∗	7B	54.9	3.1	43.8	2.7	-	-
VideoChat (Li et al., 2023b)∗	7B	56.3	2.8	45.0	2.5	34.4	2.3
BT-Adapter Liu et al. (2023c)∗	7B	67.5	3.7	57.0	3.2	-	-
Video-ChatGPT (Maaz et al., 2023)	7B	68.6	3.8	58.9	3.4	47.8	3.2
Chat-UniVi (Jin et al., 2023)	7B	70.0	3.8	53.1	3.1	46.1	3.1
VideoChat2 Li et al. (2023c)	7B	70.0	3.9	54.1	3.3	-	-
Video-LLaVA Lin et al. (2023b)	7B	71.8	3.9	59.0	3.4	48.4	3.2
LLaMA-VID (Li et al., 2023e)	7B	72.6	3.9	58.7	3.4	49.2	3.3
LLaMA-VID Li et al. (2023e)	13B	74.3	4.0	59.8	3.4	50.8	3.3
VLM-RLAIF (Ahn et al., 2024)∗	7B	76.4	4.0	63.0	3.4	-	-
LLAVA-HOUND-SFT	7B	75.7	3.9	58.7	3.3	53.5	3.3
LLAVA-HOUND-DPO	7B	**80.7**	**4.1**	**70.2**	**3.7**	**61.4**	**3.5**

Table 1: **Evaluation of Model Performance on Zero-Shot Video Question Answering Benchmarks Using gpt-3.5-turbo-0613.** Models denoted with ∗ have their results directly sourced from their original publications. Caution is advised when interpreting these results; see Appendix A for an in-depth analysis of evaluation challenges. All other baseline models were reproduced by our team.

SFT Stage (LLAVA-HOUND-SFT): We use instructional data from both image and video domain to fine-tune the model for instruction-following ability. Our SFT model use 600k image instruction data from ALLaVA and our generated 240k video instruction data, with learning rate 5e-6 and batch size 128.

DPO training Stage (LLAVA-HOUND-DPO): We use the 17k preference data introduced in section 3.3 for DPO training. Following Ivison et al. (2023), we train our policy model with full model training for 3 epochs with learning rate 5e-7, and a batch size of 128, resulting in roughly 420 training steps. All the experiments are performed on 8 A100 gpus.

5.2 Existing Benchmark Evaluation

Dataset and Testing Environment We evaluate model performance on three benchmark datasets: MSVD-QA Chen & Dolan (2011), MSRVTT-QA Xu et al. (2016), and TGIF-QA Jang et al. (2017), using ChatGPT with version gpt-3.5-turbo-0611 to assess model predictions. The evaluation prompts follow Maaz et al. (2023). In our experiment, we found that different ChatGPT versions have high impact on absolute score of metric, but the overall ranking of models is relatively stable. We select gpt-3.5-turbo-0613 due to its closeness to the reported score in Video-LLaVA paper. Further details on the selection rationale and evaluation pitfalls are discussed in Appendix A.

Baseline Selection Our selection criteria include video LMM models that have demonstrated SOTA performance, specifically including Video-LLaVA, which is also our choice of architecture. We consider other contemporaneous SOTA models with similar reported performance levels to Video-LLaVA, yet have not been directly compared in prior studies. A key consideration in our selection is the availability of models with accessible code and checkpoints, which is crucial for ensuring reproducibility of our findings. To this end, we replicate models including Video-ChatGPT (Maaz et al., 2023), LLaMA-VID (Li et al., 2023e) (7B and 13B), Chat-UniVi (Jin et al., 2023), and Video-LLaVA Lin et al. (2023b). We adopt the results from additional baselines including FrozenBiLM (Yang et al., 2022), VideoChat (Li et al., 2023b) and VideoLLaMA (Zhang et al., 2023a), sourced from their original publication.

Results In table 1, our analysis shows that within the SFT models, LLaMA-VID-7B and Video-LLaVA exhibit comparable performance, with LLaMA-VID-13B performing the best. Our LLAVA-HOUND-SFT model achieves comparable performance to LLaMA-VID-13B. Incorporating preference modeling, LLAVA-HOUND-DPO achieves an average accuracy
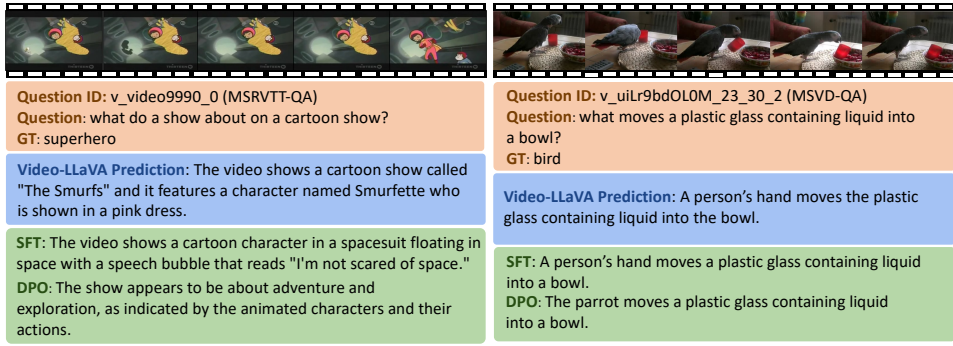
Figure 4: Examples from MSRVTT-QA and MSVD-QA showcase that our LLAVA-HOUND-DPO generates better responses, and reveal key limitations of the existing benchmark evaluation.

| No. | Methods | Proposed Video QA Benchmark (In-domain) | | | | | |
| | | ActivityNet-QA | | VIDAL-QA | | WebVid-QA | |
		Acc.	Score	Acc.	Score	Acc.	Score
[1]	Video-ChatGPT (Maaz et al., 2023)	34.17	2.19	29.35	2.10	38.88	2.27
[2]	LLaMA-VID-7B (Li et al., 2023e)	36.54	2.27	30.58	2.15	36.99	2.24
[3]	LLaMA-VID-13B (Li et al., 2023e)	37.33	2.29	32.50	2.18	39.73	2.30
[4]	Chat-UniVi (Jin et al., 2023)	39.35	2.32	31.40	2.16	40.05	2.31
[5]	Video-LLaVA Lin et al. (2023b)	41.35	2.38	34.30	2.24	42.47	2.39
[6]	LLAVA-HOUND-SFT	66.62	3.05	60.50	2.88	71.07	3.17
[7]	LLAVA-HOUND-DPO	**76.62**	**3.18**	**70.06**	**3.04**	**79.82**	**3.29**
[8]	LLAVA-HOUND-PT + Image Inst.	69.31	3.09	60.57	2.85	68.03	3.02
[9]	LLAVA-HOUND-PT + VChat	67.34	3.02	62.33	2.89	68.98	3.00
[10]	LLAVA-HOUND-DPO + training MLP	71.89	3.10	65.57	2.95	75.37	3.21
[11]	LLAVA-HOUND-SFT + Self-play	64.11	2.85	56.28	2.68	67.89	2.95
[12]	LLAVA-HOUND-DPO w/ lr3e-7	71.13	3.08	64.90	2.92	73.25	3.17

Table 2: Our proposed video QA benchmark evaluation on in-domain dataset using gpt-3.5-turbo-0301, with detailed captions as supporting evidence.

of 70.75%, surpassing LLAVA-HOUND-SFT, which has an average accuracy of 62.65%, by 8.1%. Furthermore, LLAVA-HOUND-DPO, enhanced by DPO, exhibits superior accuracy compared to VLM-RLAIF's performance achieved through reinforcement learning.

Error Analysis Figure 4 illustrates two examples. In the left example, LLAVA-HOUND-SFT provides an accurate description of the video's first half but introduces a hallucination with the phrase "I'm not scared of space," absent in the video content. LLAVA-HOUND-DPO yields a more accurate inference. In the right example, both LLAVA-HOUND-SFT and Video-LLaVA models produce incorrect inferences, whereas LLAVA-HOUND-DPO successfully correctly identifies the subject in the video. More critically, these examples unveil two significant issues within the current benchmark: (1) the auto-generated questions from existing benchmark may be grammatically incorrect or even nonsensical, and (2) the answers are limited to a *single* word, which is insufficient for evaluating LMMs with long-form text generation. Such constraints in the ground truth answers hinder the evaluation of crucial aspects like helpfulness and hallucination detection.

5.3 Proposed Benchmark Evaluation with GPT-4V Caption as Supporting Evidence

As a solution to the above limitations in existing benchmark evaluation, we propose a new set of test questions for same videos in the benchmark datasets with generated QA from detailed captions, illustrated in appendix D. Applying the our reward system in section 4, we report the score from ChatGPT, and a score value ≥ 3 will be considered correct for accuracy calculation. This new long-form QA evaluation potentially support diverse aspects in responses including relevance, accuracy, clarity and completeness in prompt 16.

7

| Methods | Proposed Video QA Benchmark (Out-of-domain) | | | | | | | |
| | MSVD-QA | | MSRVTT-QA | | TGIF-QA | | SSV2-QA | |
	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
Video-ChatGPT (Maaz et al., 2023)	34.06	2.20	25.65	1.98	31.35	2.09	19.36	1.75
LLaMA-VID-7B (Li et al., 2023e)	34.14	2.21	25.02	1.99	27.18	2.00	22.16	1.84
LLaMA-VID-13B (Li et al., 2023e)	35.81	2.25	26.34	2.02	27.58	2.01	21.98	1.83
Chat-UniVi (Jin et al., 2023)	35.61	2.23	25.89	2.01	33.23	2.13	20.59	1.79
Video-LLaVA Lin et al. (2023b)	39.46	2.37	30.78	2.15	32.95	2.18	24.31	1.90
LLAVA-HOUND-SFT	66.99	3.09	57.82	2.85	66.13	3.07	35.07	2.23
LLAVA-HOUND-DPO	**73.64**	**3.12**	**68.29**	**2.98**	**74.00**	**3.12**	**48.89**	**2.53**
LLAVA-HOUND-PT + Image Inst.	65.19	2.96	48.66	2.52	53.83	2.62	29.60	2.04

Table 3: Our proposed video QA benchmark evaluation on out-of-domain dataset using gpt-3.5-turbo-0301, with detailed captions as supporting evidence.

Table 2 and table 3 shows the in-domain and out-of-domain evaluation. We use "gpt-3.5-turbo-0301" for evaluation as it is the same version for constructing DPO dataset. The model performance is more distinguishable from our evaluation with Video-LLaVA performing the best among the other baseline models.

Video LMM without Video Instruction: [8] in table 2 is baseline trained with only image instruction fine-tuned on LLAVA-HOUND-PT, which achieves an average accuracy of 65.97%, comparable to the LLAVA-HOUND-SFT model's 66.06% in in-domain QA scenarios. However, its performance significantly drops in out-of-domain QA contexts (49.32% vs. 56.50%), suggesting that Video QA training could potentially enhance generalization capabilities.

Quality of Generated SFT: [9] substitutes our generated video QA with the Video-ChatGPT dataset for Video-LLaVA fine-tuning. A comparison between the findings of [9] and [6] reveals a marginal performance disparity of 0.2% in average accuracy, indicating that the quality of our generated QA closely parallels that of the existing video QA datasets. Given the similar quality in SFT data, the large gain of [6] over [5] can be reasonably concluded from large-scale pre-training on video captions.

Unfreeze MLP: The comparison between [10] and [7] reveals a significant decrease in performance when the MLP is unfrozen during DPO training. Despite this drop, however, the performance remains superior to that of the SFT baseline.

Smaller Learning Rate: The comparison between [12] and [7] reveals that using a smaller learning rate of 3e-7 (vs. 5e-7) results in a decreasing of model performance. This highlights the future improvements by finding better hyperparameters.

Self-Play vs. DPO: Chen et al. (2024b) introduced a self-play methodology for DPO training, which designates ground truth answers as preferred and model-generated responses as dispreferred. When comparing the results of [11] with those in [6], a notable decrease in accuracy by 3% from the SFT model is observed, suggesting that self-play may be less effective for video LMM alignment, and introducing reward model is helpful.

DPO Accuracy vs. Training Epochs. The left of fig. 5 depicts the generalization performance of the model on out-of-domain video QA tasks with respect to the number of training epochs. We observe a consistent enhancement in model performance among datasets during the initial 0 to 2 epochs, with peak performance materializing at around 2.5 epochs, which corresponds to 350 training steps.

DPO as Ranker vs. Generator. Following Hosseini et al. (2024), we compare the performance of employing the DPO model as a ranker for candidate answers produced by the SFT model, operating at a temperature setting of 1.0. As depicted on the right in fig. 5, we illustrate the test accuracy progression through the selection of the best among N candidates by the DPO ranker. Initial observations indicate that the SFT model, when set to a temperature of 1.0, demonstrates a reduced accuracy (43.3%) compared to that achieved through greedy decoding (57.8%). A steady enhancement in performance is noted as the number of candidates increases, plateauing at an accuracy of approximately 62% with 64 candidates. This
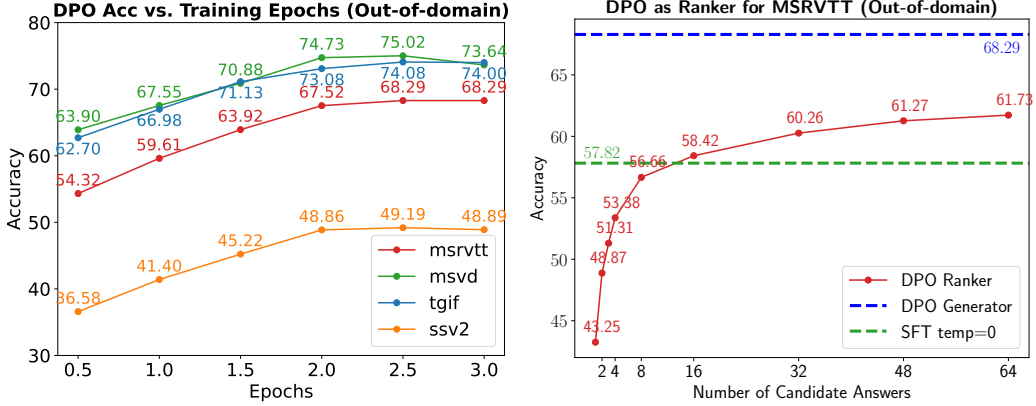
Figure 5: The left figure shows the test set accuracy of the DPO model w.r.t the number of training epochs. The right figure shows a comparison of DPO model performance as generator vs. ranker.

performance, however, falls short when compared with the direct application of the DPO model for answer generation, which yields an accuracy of 68.29%. This difference suggests the stronger generalization of DPO model in answer generation, despite it is trained on a reward classification loss. The contradictory results to Hosseini et al. (2024) may be due to the difference of tasks, i.e. Math vs. Video QA. Refer to appendix E for more results.

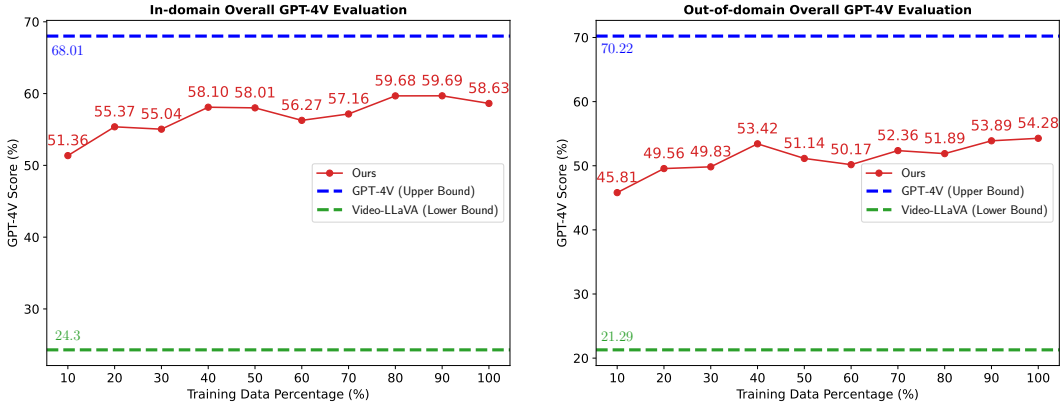## 5.4 Analysis on Video Captioning Ability from Pre-training

Figure 6: The video caption ability w.r.t number of training data evaluated on both in-domain and out-of-domain test videos using GPT-4V.

In Figure 6, we present the video captioning ability of models across various datasets, with a total of 900k distilled data instances. GPT-4V is employed for self-evaluation (fig. 14), serving as the upper-bound performance, while the Video-LLaVA serves for comparative analysis, establishing a baseline. Notably, Video-LLaVA is trained on 54k video QA data instances. However, our first checkpoint, utilizing only 10% of the data, is trained on 90k high-quality caption data instances, likely accounting for the observed performance disparity in the video captioning task. Our results demonstrate that incorporating more distilled data contributes to improved model performance across both in-domain and out-of-domain datasets. Despite these improvements, a performance discrepancy with the GPT-4V model remains. Further, we evaluate the generalization potential in specific data subsets, as shown in fig. 7 in the Appendix. These subsets reveal varying degrees of generalization challenges for different types of dataset. For example, the WebVid subset, which concentrates on relatively static scenes, necessitates less data for effective training compared to the VIDAL subset, which is marked by dynamic scene transitions and a diversity of video themes.

6   Conclusion

In this study, we propose an cost-effective reward system that utilizes detailed captions as proxies for video content. Our findings demonstrate that the reward scores is well-aligned with the evaluation metrics of GPT-4V, and the incorporation of this reward mechanism enhances DPO training, resulting in SOTA performance on video QA tasks.

7   Reproducibility Statement

The ensure reproducibility of our work, we plan to release the following items:

1. Distilled video captions with corresponding frames.
2. The model weights including the pre-trained, SFT, and DPO models.
3. Code for training and testing using existing and our proposed benchmark.

References

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024b.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

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. *arXiv preprint arXiv:2311.17005*, 2023c.

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*, 2023d.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023e.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2023a.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*, 2023c.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. Howtocaption: Prompting llms to transform video annotations at scale, 2023.

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. *arXiv preprint arXiv:2309.14525*, 2023a.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Salmon: Self-alignment with principle-following reward models. *arXiv preprint arXiv:2310.05910*, 2023b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multi-modal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pp. 5288–5296, 2016.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *NeurIPS*, 2022.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023b.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

A   Effect of ChatGPT Version on Official Benchmark Evaluation

Methods	LLM Size	MSVD-QA		MSRVTT-QA		TGIF-QA		Summary	
		Acc.	Score	Acc.	Score	Acc.	Score	Avg Acc.	Rank
gpt-3.5-turbo-0301 evaluation									
Video-ChatGPT (Maaz et al., 2023)	7B	78.62	4.00	71.67	3.63	56.31	3.45	68.87	6
LLaMA-VID (Li et al., 2023e)	7B	82.57	4.12	71.94	3.65	59.00	3.63	71.17	4
LLaMA-VID (Li et al., 2023e)	13B	83.72	4.16	73.63	3.68	59.72	3.66	72.36	3
Chat-UniVi (Jin et al., 2023)	7B	80.52	4.02	66.92	3.41	57.73	3.49	68.39	7
Video-LLaVA (Lin et al., 2023b)	7B	81.44	4.08	73.29	3.65	58.34	3.61	71.02	5
LLaVA-Hound-SFT	7B	85.65	4.10	73.85	3.62	64.98	3.65	74.83	2
LLaVA-Hound-DPO	7B	**88.50**	**4.20**	**82.10**	**3.84**	**75.48**	**3.81**	**82.03**	1
gpt-3.5-turbo-0613 evaluation									
Video-ChatGPT (Maaz et al., 2023)	7B	68.55	3.80	58.90	3.36	47.83	3.21	58.43	6
LLaMA-VID (Li et al., 2023e)	7B	72.62	3.92	58.73	3.38	49.21	3.28	60.19	4
LLaMA-VID Li et al. (2023e)	13B	74.29	3.96	59.82	3.41	50.83	3.33	61.65	3
Chat-UniVi (Jin et al., 2023)	7B	70.01	3.79	53.08	3.14	46.09	3.12	56.39	7
Video-LLaVA (Lin et al., 2023b)	7B	71.75	3.88	58.97	3.39	48.39	3.24	59.70	5
LLaVA-Hound-SFT	7B	75.70	3.86	58.73	3.31	53.51	3.30	62.65	2
LLaVA-Hound-DPO	7B	**80.73**	**4.07**	**70.15**	**3.66**	**61.38**	**3.46**	**70.75**	1
gpt-3.5-turbo-1106 evaluation									
Video-ChatGPT (Maaz et al., 2023)	7B	73.02	4.01	62.09	3.61	47.76	3.36	60.96	6
LLaMA-VID (Li et al., 2023e)	7B	75.49	4.08	62.09	3.61	51.72	3.47	63.10	4
LLaMA-VID (Li et al., 2023e)	13B	76.97	4.10	63.16	3.61	52.53	3.50	64.22	3
Chat-UniVi (Jin et al., 2023)	7B	72.22	3.92	55.02	3.35	48.16	3.31	58.47	7
Video-LLaVA (Lin et al., 2023b)	7B	74.76	4.04	62.70	3.60	51.21	3.45	62.89	5
LLaVA-Hound-SFT	7B	81.09	4.08	64.13	3.57	58.05	3.53	67.76	2
LLaVA-Hound-DPO	7B	**86.05**	**4.23**	**76.75**	**3.85**	**70.02**	**3.71**	**77.61**	1

Table 4: **Performance Evaluation Across ChatGPT Versions on Zero-Shot Video Question Answering Benchmarks.** This table compares the performance of state-of-the-art video LMMs evaluated under different ChatGPT versions. The absolute performance metrics scored by ChatGPT vary by versions. However, the comparative ranking of models under the same ChatGPT version is relatively stable.

In Table 4, we show impact of using different ChatGPT versions on metric scores within zero-shot video question answering benchmarks. Our analysis reveals significant variations in the absolute scores across ChatGPT versions, but based on the average accuracy metric, the relative ranking of models under the same ChatGPT version shows consistency.

This comparison underscores a critical issue: many prior studies neglect to specify the ChatGPT version used, potentially leading to inaccurate conclusions during evaluation. We advocate for the explicit designation of the ChatGPT version in future evaluations. Analysis from Table 4 indicates that the version gpt-3.5-turbo-0613 aligns most closely with the performance of the Video-LLaVA (Lin et al., 2023a) model, serving as the benchmark for model performance comparison in our study.

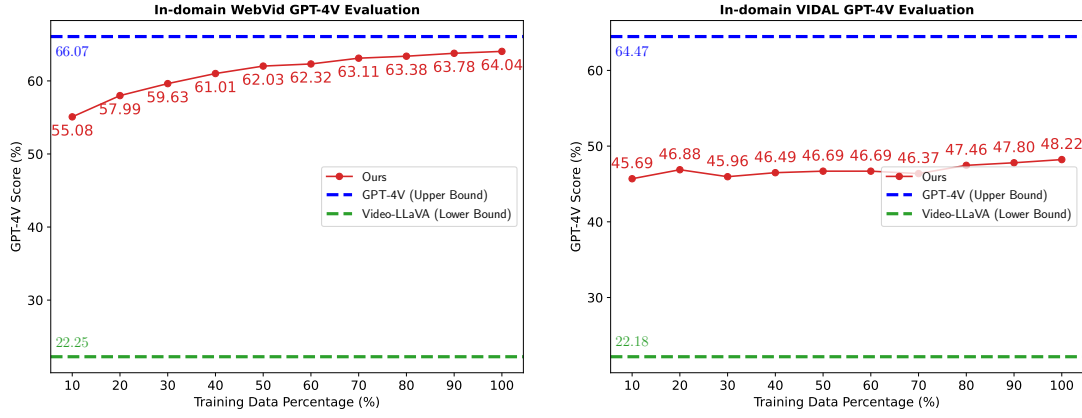# B   Evaluation of Captioning Ability from pre-training

Figure 7: Training subsets exhibit varying levels of generalization difficulty. The WebVid subset (left) requires less data compared to the VIDAL subset (right)

C Distilled Caption Demonstration

Original Caption: Time-lapse of purple iris flowers blooming.

GPT-4V Annotation: The video captures a close-up time-lapse of purple irises with bright yellow accents against a black background. Initially, the buds are tightly closed, with green sepals protecting the petals. As the frames advance, the buds begin to open slowly, revealing the velvety texture and deep purple hue of the petals. The blooming is sequential, with some buds opening earlier than others, offering a dynamic contrast between the stages of bloom. At final bloom, the yellow stamen stands out against the purple petals. Throughout the sequence, the number of visible flowers changes.

Original Caption: When you flip a pancake (pt.26) #shorts

GPT-4V Annotation: The video begins with a close-up view of a pancake on a griddle or pan, which has not been flipped yet, showing its uncooked, batter side. The pancake is pale in color with bubbles forming on the surface, typical of a pancake as it cooks. There is a spatula with a novelty design. It resembles a Stormtrooper helmet from the Star Wars franchise placed beside the pancake. Text urging the viewer to "Send this to your bestie!" along with a Snapchat handle are overlaid on top of the video. The spatula is then slid underneath the pancake, which is then flipped over to reveal a cooked side featuring the drawing of a cartoon crab and the phrase "U R CLAWSOME!" written on it. The cooked side of the pancake has a golden-brown color, and the cartoon and text appear to be created with some dark-colored substance, possibly chocolate or some edible ink.
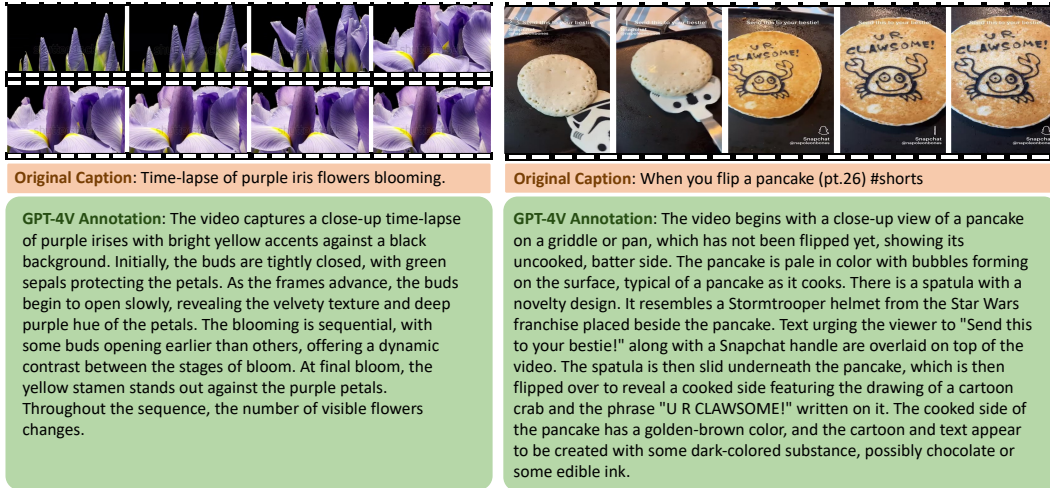
Figure 8: A showcase of our dataset for detailed video caption automatically distilled from GPT-4V via our prompt technique in section 3.

D Video QA Dataset Demonstration

MSRVTT Video ID: video7012

Existing Video QA Dataset

Q: what is a suit man doing?
A: talk

Q: what does a man seated in a chair discuss?
A: health

Q: who is speaking about the relevance of disorders?
A: man

Our Video QA Dataset

Q: What is the person in the video doing and what does their appearance suggest?

A: The person in the video is sitting in a leather chair against a wooden backdrop with framed art and a small light fixture on the wall and appears to be delivering a speech or presentation. Their appearance suggests that they are confident and engaged in the presentation or interaction.

Q: What decorative items are on the desk in the video and what does the lighting in the room look like?

A: The desk in the video features a couple of decorative items including a blue-striped spherical object and a small chrome lamp with a white shade. The lighting in the room is soft and comes from the person's left side, casting soft shadows.

Figure 9: Comparing testing QA in existing benchmark with that in our proposed new benchmark.

MSRVTT Video ID: video7016

Existing Video QA Dataset

Q: what are girls and their mom doing?
A: marble

Q: who does a woman teach how to paint eggs?
A: girl

Q: what is kids doing?
A: color

Our Video QA Dataset

Q: What is the purpose of the video?
A: The purpose of the video is to showcase the process of dyeing Easter eggs.

Q: What are the tools and materials used in the video?
A: The tools used in the video are a wire egg holder, and the materials are a variety of colorful dyes, eggs, and a tablecloth.

Q: What is the significance of the tablecloth in the video?
A: The tablecloth is adorned with whimsical fish designs and protects the workspace where the dyeing process takes place.

Figure 10: Comparing testing QA in existing benchmark with that in our proposed new benchmark, example 2.
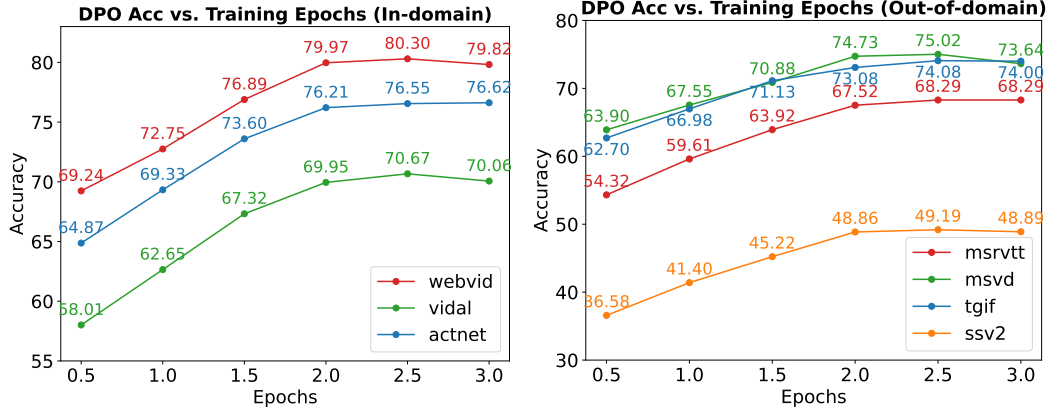
E   Additional DPO Results

Figure 11: Test Set Accuracy of the DPO Model vs. Training Epochs. The figure illustrates a consistent trend in both in-domain and out-of-domain video QA, with peak performance occurring at approximately epoch 2.5, equivalent to 350 training steps.
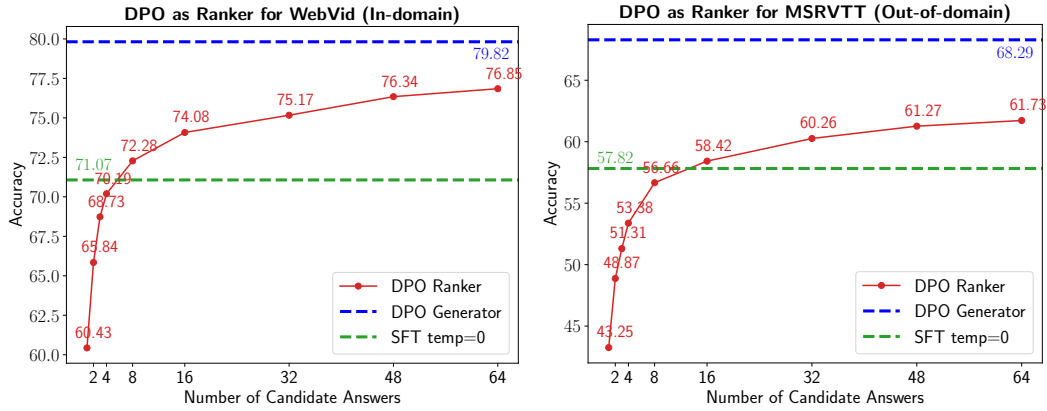

Figure 12: Comparison of DPO Model Performance: Ranker vs. Generator. The DPO model serves as a ranker, assigning reward scores to candidate answers generated by the SFT model with a temperature setting of 1.0. Employing the DPO model directly for answer generation results in superior performance compared to its use as a ranker.

18

F Prompts for GPT-4V and ChatGPT Queries

```
Task Instructions:

Given a caption that summarizes the content of a video, generate three question-answer
pairs that relate directly to the information and context provided in the caption. The
questions should be grounded to the understanding of the video content.

Guidelines for QA Generation:

1. Helpfulness: Answers should provide sufficient detail and depth to fully address
the question. They should include relevant explanations, or context where appropriate,
to enhance understanding.

2. Faithfulness: The answers must accurately reflect the information presented in the
video caption. Avoid speculation or the inclusion of information not contained or
implied by the caption to maintain the integrity of the content.

3. Diversity: Craft questions that cover different aspects of the video caption to
provide a comprehensive understanding of the content. This includes factual inquiries,
inferential questions, and those that may elicit explanatory responses.

Input Video Caption:
{caption}

Output format:
Q1: <question1>
A1: <answer1>
Q2: <question2>
A2: <answer2>
Q3: <question3>
A3: <answer3>
```

Figure 13: ChatGPT for instruction generation.

```
Your role is to serve as an impartial and objective evaluator of a video caption
provided by a Large Multimodal Model (LMM). Based on the input frames of a video,
assess primarily on two criteria: the coverage of video elements in the caption and
the absence of hallucinations in the response. In this context, 'hallucination' refers
to the model generating content not present or implied in the video, such as incorrect
details about objects, actions, counts, or other aspects not evidenced in the video
frames.

To evaluate the LMM's response:

Start with a brief explanation of your evaluation process.
Then, assign a rating from the following scale:

Rating 6: Very informative with good coverage, no hallucination
Rating 5: Very informative, no hallucination
Rating 4: Somewhat informative with some missing details, no hallucination
Rating 3: Not informative, no hallucination
Rating 2: Very informative, with hallucination
Rating 1: Somewhat informative, with hallucination
Rating 0: Not informative, with hallucination

LMM Response to Evaluate
{LLM_response}

Output format:
Judgment: <your judgment>
Score: <integer value rating>
```

Figure 14: GPT-4V evaluation prompt for video captioning.

```
Given the following inputs:

1. **Ground Truth Video Caption**: {caption}
2. **Question Related to the Caption**: {question}
3. **Ground Truth Answer**: {answer}
4. **Model Predicted Answer**: {prediction}

Your task is to evaluate the model's predicted answer against the ground truth answer,
based on the context provided by the video caption and the question. Consider the
following criteria for evaluation:

- **Relevance**: Does the predicted answer directly address the question posed,
considering the information provided in the video caption?
- **Accuracy**: Compare the predicted answer to the ground truth answer. Does the
prediction accurately reflect the information given in the ground truth answer without
introducing factual inaccuracies?
- **Clarity**: Assess the clarity of the predicted answer. Look for issues such as
repetition, unclear descriptions, or any grammatical errors that could hinder
understanding.
- **Completeness**: Determine if the predicted answer fully covers the scope of the
ground truth answer. Does it leave out critical information or does it include all
necessary details?

**Output Format**:
Explanation: <brief judgement of prediction>
Score: <a integer score of quality from 1-5>
```

Figure 15: ChatGPT-Evaluation Prompt for Video Question Answering. This prompt takes in a detailed caption, question, ground truth answer, and model prediction, subsequently generating an assessment of the prediction's quality alongside a corresponding score based on predefined criteria. A score value ≥ 3 will be considered correct for accuracy calculation.

```
Your task is to act as an impartial and objective assessor of answers generated by a
Large Multimodal Model (LMM) for video-based questions. Utilizing video frames, a
posed question, and the model's provided answer, your evaluation should focus on the
following aspects:

- **Relevance**: Does the predicted answer directly address the question posed,
considering the information provided in the video caption?
- **Accuracy**: Compare the predicted answer to the ground truth answer. Does the
prediction accurately reflect the information given in the ground truth answer without
introducing factual inaccuracies?
- **Clarity**: Assess the clarity of the predicted answer. Look for issues such as
repetition, unclear descriptions, or any grammatical errors that could hinder
understanding.
- **Completeness**: Determine if the predicted answer fully covers the scope of the
ground truth answer. Does it leave out critical information or does it include all
necessary details?

**Input**:
Question: {question}
Model Predicted Answer: {prediction}

**Output Format**:
Explanation: <brief judgement of prediction>
Score: <an integer score of quality from 1-5>
```

Figure 16: GPT-4V Evaluation Prompt for Video Question Answering. Together with video frames input in GPT-4V API, this prompt takes in a question, and model prediction, subsequently generating an assessment of the prediction's quality alongside a corresponding score based on predefined criteria. A score value ≥ 3 will be considered correct for accuracy calculation. This is used to assess the quality of ChatGPT evaluation in fig. 15.