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

Ruohong Zhang[*♠] **Liangke Gui**[*] [♠♢]

Zhiqing Sun[♠], **Yihao Feng**[▽], **Keyang Xu**[△], **Yuanhan Zhang**[♡], **Di Fu**[♢], **Chunyuan Li**[♢], **Alexander Hauptmann**[♠], **Yonatan Bisk**[♠], **Yiming Yang**[♠]

[♠]CMU LTI, [♢]Bytedance, [▽]UT Austin, [△]Columbia University, [♡]NTU

Project Page: https://github.com/RifleZhang/LLaVA-Hound-DPO

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 open-ended conversations, remains a significant challenge. While previous studies have explored using large multimodal models (LMMs) as reward models for guiding preference modeling, their ability to accurately assess the quality of generated responses and their alignment with video content has not been conclusively demonstrated. 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 our reward mechanism to DPO algorithm significantly improves model performance on open-ended 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., 2024d; Hosseini et al., 2024), which have been effective in guiding LLMs towards generating more honest, helpful, and harmless content, their effectiveness in video domain remains limited. The critical obstacle lies in developing a robust reward system capable of distinguishing preferred responses from less preferred ones based on video inputs. The challenge is further complicated by the coverage and potential inaccuracies 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 distilling 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 that is both computationally and financially efficient for evaluating the quality of responses generated by video LLMs, serving as a basis for further on-policy preference optimization. We propose the use of detailed video captions as a proxy for video content, enabling a language model analyze the content and assess the quality of an LMM's response to related questions. The language model generates natural language feedback as a chain-of-thought step, and produces a numerical score as the reward, thereby creating an efficient 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 simple prompting technique with the GPT-4V model, comprising 900k captions that encompass a wide range of video content, including temporal dynamics, world knowledge,

[*]Equal contribution.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 694–717
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

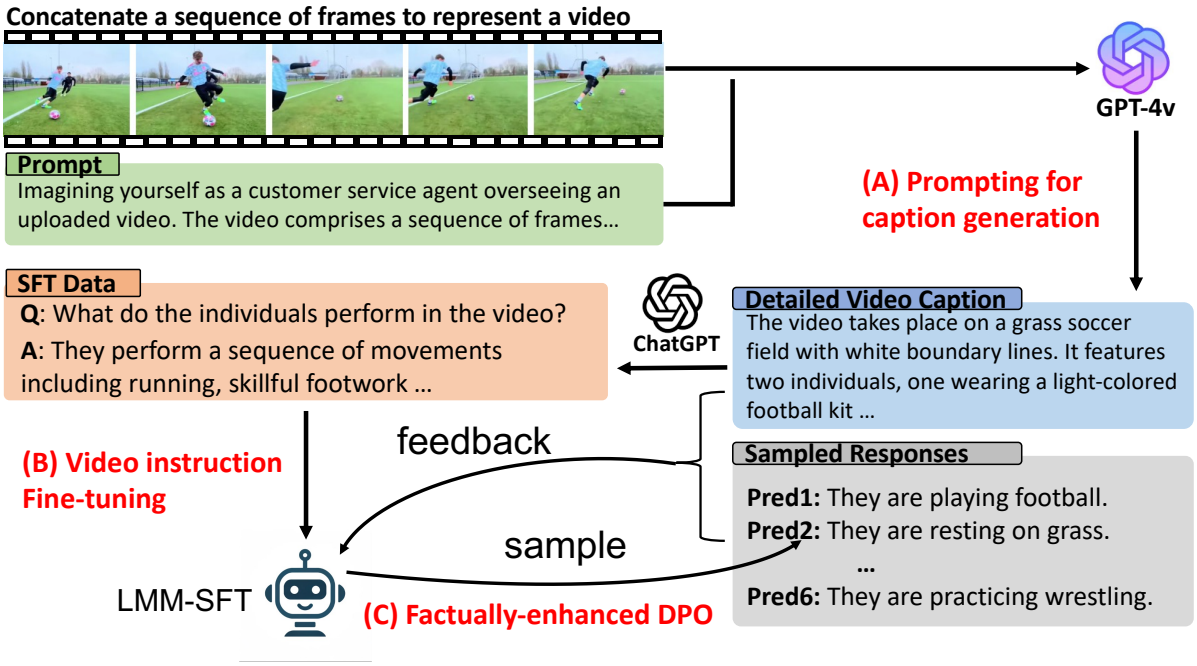


Figure 1: Workflow diagram showing: a) the use of GPT-4V for creating a detailed caption dataset for videos; b) generating video instruction data for SFT; c) integrating captions into a feedback loop for DPO, improving the model's performance on video instruction-following tasks.

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 release a large-scale detailed video caption (900k) and instruction-following (900k) dataset covering a wide range of video content, which facilitates video LMM model training and research.

2. We demonstrate the effective application of DPO to improve model performance by leveraging the language model feedback as reward, which substantially improves model performance on open-ended video QA tasks.

3. We propose an automated *development* benchmark for evaluating video instruction-following capability, serving as a cost-effective way to validate model performance.

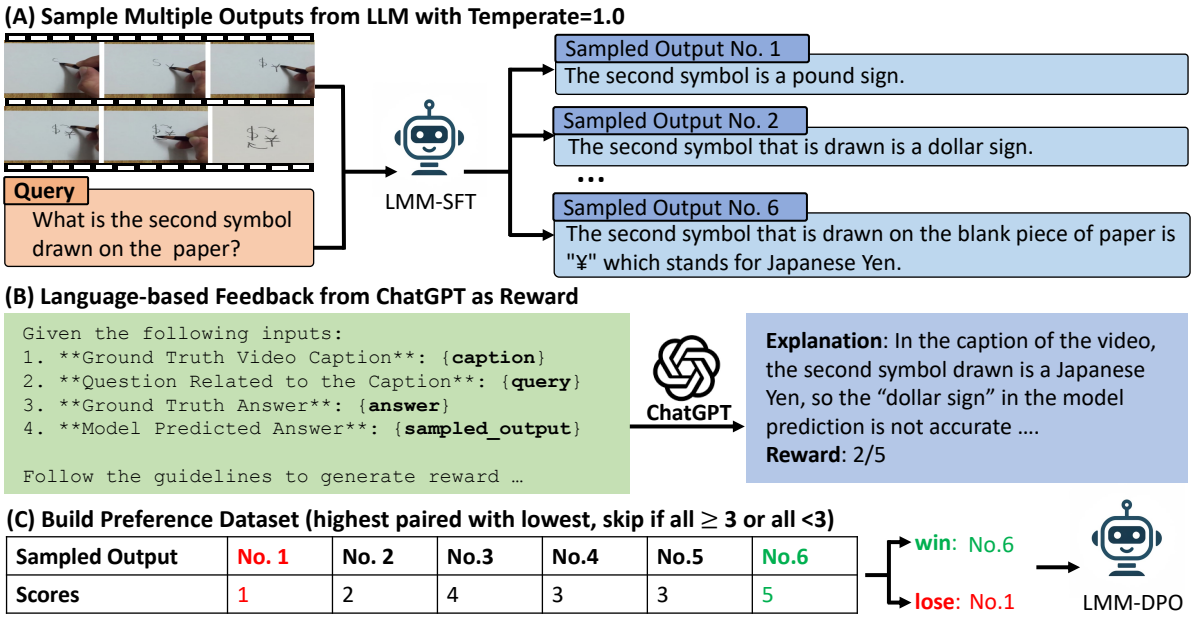
2 Related Work

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 (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; Zhang et al., 2024). Models such as Qwen2-VL (Wang et al., 2024) and InternVL-2.5 (Chen et al., 2024c), which scale in both model size (ranging from 1 billion to 78 billion parameters) and training data volume, have demonstrated highly competitive performance in image and video understanding tasks. However, these studies fall outside the scope of this research. Our work adopts Video-LLaVA (Lin et al., 2023a) 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., 2021a; Wang et al., 2023; Yu et al., 2019; Jang et al., 2017; Xu et al., 2016). (Shvetsova et al., 2023). Video-ChatGPT (Li et al.,



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

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

2023b) employs human effort to create high-quality video instructions, albeit limited to the ActivityNet domain with only 100k instruction pairs. Short2Story (Han et al., 2023) and Vript (Yang et al., 2024) employ GPT-4V for video captioning, with audio details as outputs. Concurrent work (Chen et al., 2024b) leverages GPT-4V to label video captions. Furthermore, previous studies (Xu et al., 2017; Grunde-McLaughlin et al., 2021; Wu et al., 2024) have employed language models or predefined question templates to generate question-answer pairs. While these approaches can produce a substantial volume of questions and answers, they frequently result in low-quality questions that fail to accurately reflect real-world user inquiries. Our work leverages the GPT-4V model to generate detailed video captions and introduces a paradigm for producing question-answer pairs, which can be effectively utilized for training video LLMs. Additionally, we release this resource as a community asset for LMM training.

2.3 Preference Modeling for LMMs

Preference modeling techniques, such as DPO (Deng et al., 2024; Yu et al., 2024; Li et al., 2023d; Gunjal et al., 2023; Sun et al., 2023a) and PPO (Sun et al., 2023a), have been applied to LMM alignment. More recently, (Ahn et al., 2024) employed reinforcement learning based on AI feedback to enhance video LMM performance. Our work extends DPO to video LMM alignment by incorporating detailed captions as factual evidence for reward modeling.

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: WebVid (400k) and VIDAL (450k) ActivityNet (50k) datasets. WebVid and VIDAL videos are in the general domain sourced from YouTube, and ActivityNet videos focus 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. We use GPT-4V to generate a coherent caption for the represented video based

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 3: Example annotation in SHAREGPTVIDEO: Our dataset provides more detailed video caption annotations. The red box demonstrates that textual information in the video is accurately extracted into the caption, while the blue box illustrates the connection to real-world knowledge, such as identifying the spatula's shape as resembling a Stormtrooper helmet from Star Wars.

on the frame sequence. The prompt (fig. 19) adheres to guidelines covering temporal dynamics, world knowledge, object attributes, spatial relationships, aesthetic assessments, etc., with the goal of comprehensively understanding the video contents (examples in fig. 3 and fig. 9).

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 300k video captions and then employ ChatGPT to generate 3 question-answer pairs conditioned on each caption (prompt in fig. 20). We release the 900k instruction-following data to public, but we only use a random subset of 240k for our training. This approach ensures that the instructional data remains factually consistent with the content of the detailed captions.

3.3 DPO with Language Model Reward

Acquiring high-quality on-policy preference data can be costly and labor-intensive. Although GPT-4V can be used for reward distillation, for video data, its high computation cost[1], slow response, and limited accessibility hinder scalability. 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 generates six responses per input at a temperature of 1.0. This procedure results in 120k question-answer pairs. Subsequently, we employ ChatGPT to evaluate the model responses based on the ground truth answer and detailed description (prompt in fig. 22). ChatGPT 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 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:

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

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.

For on-policy reward generation, our method incurs a cost 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.

[1]Video representation is typically encoded with 2048 tokens, while our captions only uses roughly 140 tokens.

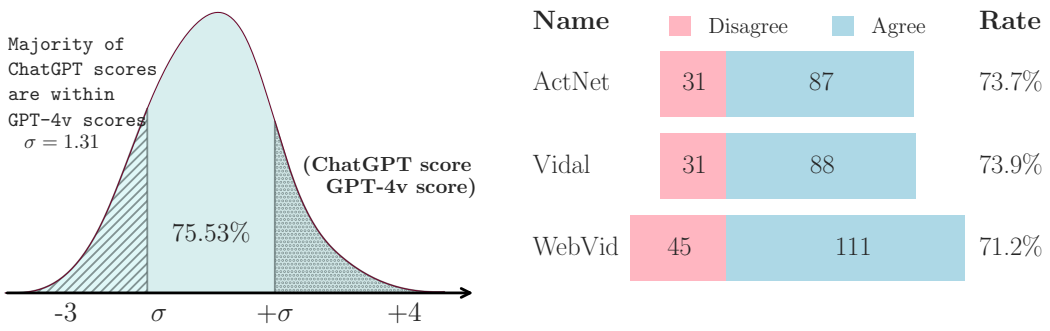


Figure 4: Assessing Evaluator Quality Using Captions in Place of Frames. The left figure shows the distribution of evaluation score differences between ChatGPT (with caption as proxy) and GPT-4V (directly on frames) evaluations. The right figure shows the rate of preference agreement between ChatGPT and GPT-4V as evaluators.

4 Assessment of Evaluator with GPT-4V Caption as Video Content

To assess the effectiveness of our proposed reward assignment method, we conducted a comparative analysis the GPT-4V used as a video QA evaluator. Our method utilizes detailed captions as a proxy of actual video frames, while GPT-4V directly takes in video frames as inputs. Both reward systems follow the same set of guidelines for scoring reward (prompt in fig. 23).

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 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. 4 (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. 4 (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.

Human Annotation of Captions: To evaluate the quality of the distilled captions, we conducted human annotations focusing on two aspects: coverage and accuracy (hallucination). Annotators were asked to assess each caption by identifying the number of missing items and the number of incorrect facts. The assessment was performed on a sample of 75 videos, with 25 from each domain. The results showed that annotators identified a total of 21 inaccurate items across 14 videos (accuracy: 81%) and 12 missing items across 8 videos (accuracy: 89%). Annotated examples are provided in appendix D.

5 Experiments

We adopt Video-LLaVA (Lin et al., 2023a) as the backbone of our video LMM, but our method can be applied to any other architectures as well.

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

SFT Stage (LLaVA-Hound-SFT): We 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. All the experiments are performed on 8 A100 gpus.

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.

No.	Methods	Next-QA	
		Acc.	Score
[1]	Video-ChatGPT (Maaz et al., 2023)	45.23	2.09
[2]	LLaMA-VID-7B (Li et al., 2023e)	49.43	3.24
[4]	Chat-UniVi (Jin et al., 2023)	47.62	3.14
[5]	Video-LLaVA (Lin et al., 2023b)	48.97	3.25
[6]	LLaVA-HOUND-SFT	60.60	3.51
[7]	LLaVA-HOUND-DPO	**74.27**	**3.74**

Table 2: Evaluation on Next-QA benchmark using gpt-3.5-turbo-0611 on official test set.

5.1 Benchmark Evaluation

Dataset and Testing Environment We evaluate model performance on four benchmark datasets: MSVD-QA (Chen and Dolan, 2011), MSRVTT-QA (Xu et al., 2016), TGIF-QA (Jang et al., 2017), and Next-QA (Xiao et al., 2021) 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 We select video LMM models that have demonstrated SOTA performance with with accessible code and checkpoints at the time of paper writing, specifically including Video-LLaVA, which is also our choice of architecture. We replicate results 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 copy 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 of 70.75%, surpassing LLaVA-HOUND-SFT, which has an average accuracy of 62.65%, by 8.1%. Furthermore, LLaVA-HOUND-DPO exhibits superior accuracy compared to other RL methods such as VLM-RLAIF. In table 2, our model demonstrated consistent result on a relative new benchmark Next-QA.

Error Analysis Figure 5 illustrates two examples. In the left example, LLaVA-HOUND-SFT

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

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

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.

5.2 Open-ended QA Analysis

In this section, we conduct analysis on open-ended long-form QA with a proposed development benchmark. Specifically, we select 2,000 videos from each source: WebVid (Bain et al., 2021b), VIDAL (Zhu et al., 2023), ActivityNet (Fabian Caba Heilbron and Niebles, 2015), MSRVTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), TGIF (Jang et al., 2017), and Something-something V2 (SSV2) (Goyal et al., 2017). For each video, ChatGPT was utilized to generate three QA pairs based on the detailed captions, and we evaluate model predictions with our reward mechanism. WebVid, VIDAL, ActivityNet are classified as in-domain, which are involved in the model's training pipeline. MSRVTT, MSVD, TGIF, SSV2 are classified as out-of-domain.

The evaluation reveals insights into (1) the quality of long-form open-ended QA, (2) in-domain and out-of-domain generalization, and (3) Ablations on SFT and DPO experiments. Additionally, we select our best performing model on the development bench before evaluating on public benchmarks, which avoids tuning hyperparameters on

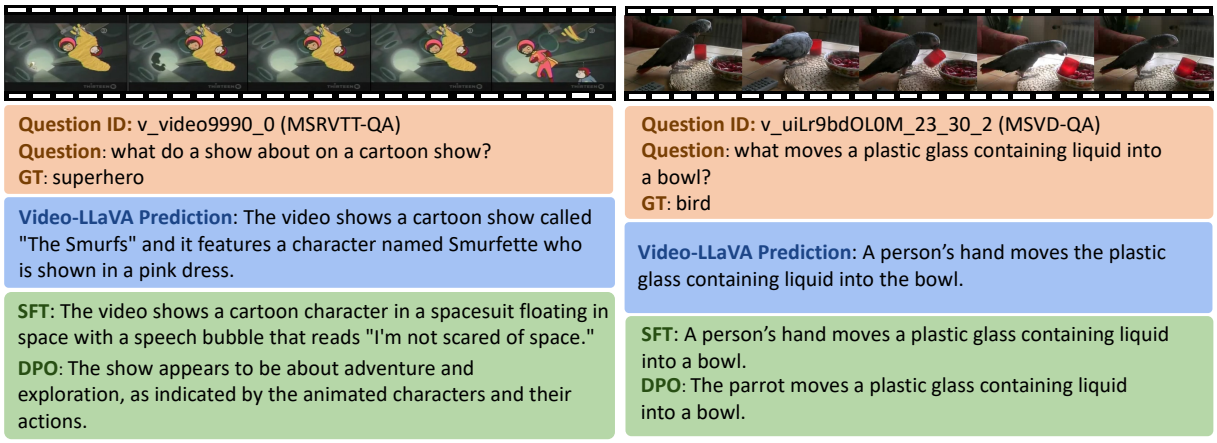


Figure 5: Examples from MSRVTT-QA and MSVD-QA showcase that our LLaVA-Hound-DPO generates better responses, and reveal key limitations of the existing benchmark evaluation.

test data. Comparisons are shown in appendix E.

Domain Generalization: Table 3 and table 4 shows the in-domain and out-of-domain evaluation. SFT with our data tends to perform better both in- and out-of-domain, and DPO further enhances the model performance, showing the effectiveness of preference modeling.

Video LMM without Video Instruction: [8] in table 3 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., 2024d) 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. 6 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.

Test-time Compute. Previous works have demonstrated that leveraging test-time compute by generating and evaluating multiple candidates can enhance model performance (Hosseini et al., 2024). In this study, we assess the effectiveness of test-time compute and compare its performance with direct inference using our DPO model. For ranking, we employ the DPO model as a ranker to score candidate answers produced by the SFT model, operating at a temperature setting of 1.0. As illustrated on the right in fig. 6, we present the test accuracy progression when selecting the best among N candidates using the DPO ranker. Initial observations reveal that the SFT model, when set to a temperature of 1.0, attains a lower accuracy (43.3%) com-

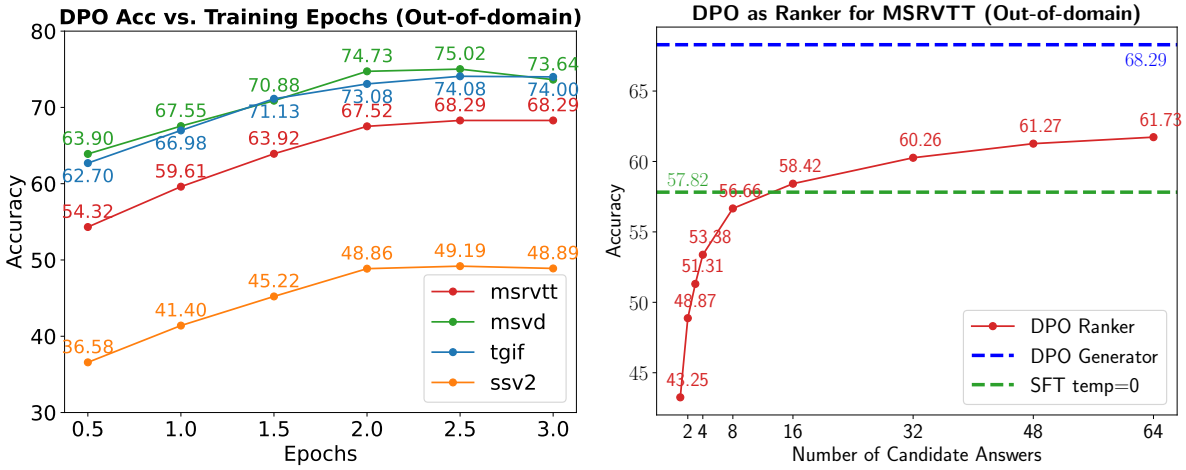


Figure 6: The left figure shows the test set accuracy of the DPO model w.r.t the number of training epochs. The right figure shows a comparison of DPO model performance as generator vs. ranker.

pared to greedy decoding (57.8%). A consistent improvement in performance is observed as the number of candidates increases, eventually plateauing at approximately 62% accuracy with 64 candidates. However, this performance remains inferior to direct answer generation using the DPO model, which achieves an accuracy of 68.29%. This discrepancy suggests that the DPO model exhibits stronger generalization in answer generation, despite being trained with a reward classification loss. The contrasting results compared to (Hosseini et al., 2024) may stem from task differences, specifically Math QA versus Video QA. Refer to appendix F for more results.

6 Conclusion

We study the techniques for effective video LMM alignment. Specifically, we propose an cost-effective reward system that utilizes detailed captions as proxies for video content. We have shown the reward scores is well-aligned with the evaluation metrics of GPT-4V, and DPO training greatly enhances model performance. In addition, we have released 900k detailed video caption, 900k video instruction-following data, and 17k preference data pairs, with a complete code pipeline including pre-training for video captioning, fine-tuning for video instruction following and reinforcement learning with DPO for better LMM alignment.

7 Limitations

Firstly, several evaluation datasets, such as Video-MME (Fu et al., 2024) featuring multiple-choice questions, were not included in our study. These datasets were available at or before the completion of our manuscript. Given our focus on enhancing open-ended question answering, multiple-choice datasets were not incorporated into our training process. Consequently, we did not retrain the model to include this data.

Secondly, the benchmark we developed is fully automated and does not incorporate human corrections for captions and QA. Human annotations indicate that caption accuracy ranges from 80% to 90%, inherently introducing errors. Therefore, we recommend using this benchmark solely for model development and hyperparameter tuning, treating performance metrics as indicative rather than definitive.

References

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

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

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

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021a. 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*, pages 1728–1738.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021b. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*.

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

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

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

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024b. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *ICCV*.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *CVPR*.

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

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. 2023. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*.

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

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

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

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

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

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

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

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2023c. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*.

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

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

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

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

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

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

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

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

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

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

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

Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. 2023. Howtocaption: Prompting llms to transform video annotations at scale. *Preprint*, arXiv:2310.04900.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023a. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

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

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2024. Vript: A video is worth thousands of words. *arXiv preprint arXiv:2406.06040*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

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

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

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

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

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

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

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 5: **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 5, 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 5 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

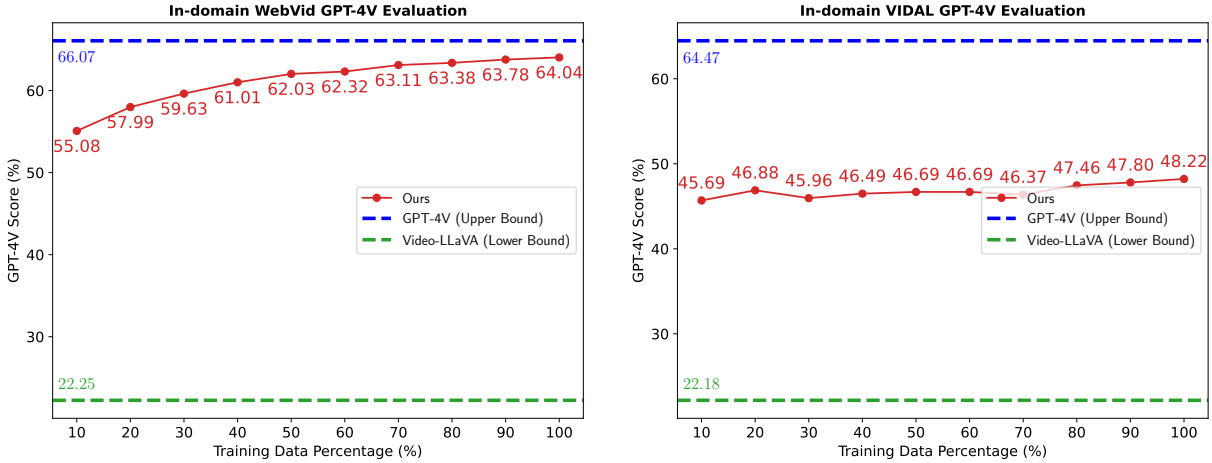


Figure 7: Training subsets exhibit varying levels of generalization difficulty. The WebVid subset (left) requires less data compared to the VIDAL subset (right)

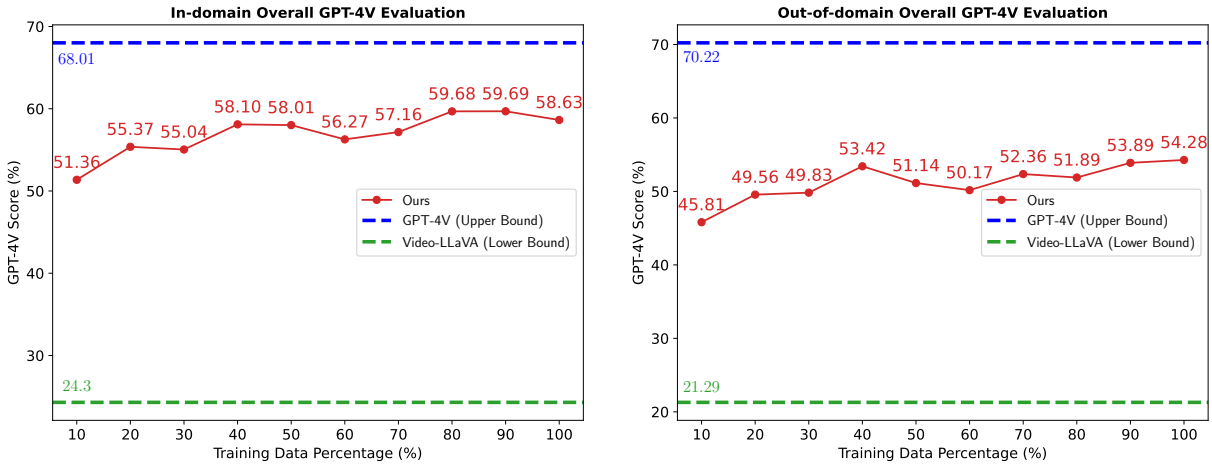


Figure 8: The video caption ability w.r.t number of training data evaluated on both in-domain and out-of-domain test videos using GPT-4V.

In Figure 8, 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. 21), 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.

C GPT-4V Caption Distillation

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.

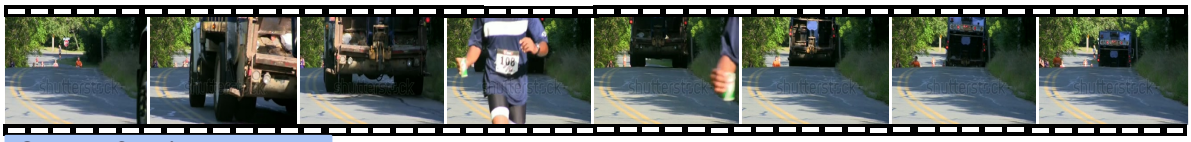
Original Caption: When you flip a pancake (pt.26) #shorts

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 9: A showcase of our dataset for detailed video caption automatically distilled from GPT-4V via our prompt technique in section 3.

D Human Annotated Examples of Distilled Captions

WebVid Video ID: 2994286



GPT-4V Caption

Initially, the video displays a curved road lined with multiple orange traffic cones and greenery on either side. The focus shifts to a large, dark-colored truck approaching, its engine sound becoming increasingly prominent. The camera angle changes to the rear of the truck, revealing a runner, wearing a blue top with a white number bib marked '108', passing the truck on its left side. The video concludes with the truck's rear-end seen driving away with the traffic cones visible in the background.

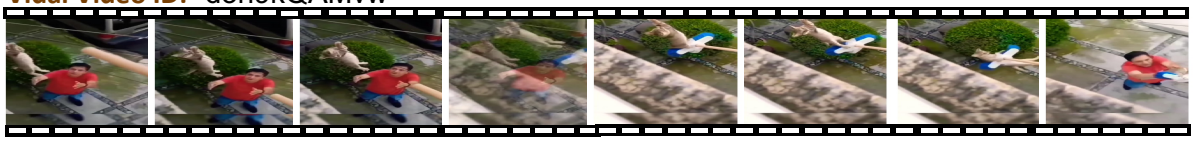
Human Annotation

Inaccuracy (Hallucination):
1. The truck is driving away instead of approaching

Coverage:
1. The runner is holding a cup-like item

Figure 10: Human Annotation Analysis: The video frames depict a truck driving away from the camera, but the distilled caption incorrectly describes the direction, which is identified as a hallucination. Additionally, human annotators noted that the GPT-4V model failed to recognize a man holding a cup.

Vidal Video ID: -d3h9kQAMvw



GPT-4V Caption

The cat is observed suspended by a limb caught in an electrical cable. It is a light-colored feline, possibly tan or cream, with a noticeably fluffy tail. Below, the person stands on what appears to be a blue stool or small ladder, reaching out towards the cat in an attempt to free it. The environment includes a green, neatly trimmed bush directly under the cat, and to the side, there appears a section of a vehicle – likely a car, suggesting the proximity to a driveway or parking spot. The pavement where these activities are occurring has a distinct pattern, likely tiles, surrounded by green grass, which borders a residential building. Above, an electrical pole is partially visible with thick cables extending from it – one of which the cat is caught in.

Human Annotation

Inaccuracy (Hallucination):
1.No blue stool or ladder visible

Coverage:
-

Figure 11: Human Annotation Analysis: GPT-4V incorrectly identified a man wearing blue pants as a man standing on a blue ladder, which is labeled as a hallucination.

ActivityNet Video ID: v__Y4czVoIxvE-Scene-006

GPT-4V Caption

Initially, two people are seen sitting on the floor with crafting materials. They both wear casual clothing: one in a dark blue long-sleeve top with "Johnny's BBQ" written on it and the other in a red and black checkered shirt. They appear to be in a cozy room with a wooden floor and a heater in the background. The next scenes showcase the crafting process in detail, focusing on hands carefully placing square-shaped burgundy materials with patterned textures onto a lighter brown base. The precision of their movements suggests they are experienced in this type of work.

Human Annotation

Inaccuracy (Hallucination):
1. The text on shirt should be "Jimmy's BBQ"

Coverage:
1. Missing text: "Step 3: Adhere Squares Together"

Figure 12: Human Annotation Analysis: The example demonstrates OCR hallucination and a coverage issue. The text on the woman's t-shirt is incorrectly labeled as 'Johnny's BBQ' instead of 'Jimmy's BBQ.' Additionally, the text 'Step 3: Adhere Squares Together' is missing from the description.

ActivityNet Video ID: v__wITx73-BXw-Scene-003

GPT-4V Caption

The video showcases multiple scenes from a soccer match being played on a field surrounded by an athletic track. It appears to be in the evening, and a scoreboard is visible in the background indicating the game is underway. The players, wearing contrasting uniforms, are seen running, maneuvering the ball, and aiming at the goal. The crowd can be faintly seen on the sidelines watching the game.

Human Annotation

Inaccuracy (Hallucination):
-

Coverage:
1. "Gopher" and "Coca Cola" on the board
2. Score board shows time left

Figure 13: Human Annotation Analysis: The caption does not contain any hallucinations, but some text recognized by human annotators is missing, such as 'Coca Cola' and 'Gopher' on the scoreboard, as well as the time of the score match shown.

ActivityNet Video ID: v__Wag6CT_0j8-Scene-003

GPT-4V Caption

Throughout the video, there are multiple sequences featuring women in athletic attire engaged in a field hockey match. The players are distributed across a well-maintained, green artificial turf field. Two teams are visible: one in dark blue and white striped jerseys, and the other in white jerseys with black numerals. The white-jersey team appears to be on defense, while the dark-jersey team is in an offensive posture, advancing the ball. Some of the key visual elements include the athletic gear, such as shin guards, field hockey sticks, and cleats that the players are wearing. The background includes features like team benches, spectator stands, and a scoreboard that shows "MONMOUTH 0 2nd RIDER 2". Throughout the various frames of the video, players can be seen running, likely maneuvering the ball, and repositioning to either defend or continue their offensive push. Ambient shadows suggest sunny weather conditions during the recording of this match.

Human Annotation

Inaccuracy (Hallucination):
-

Coverage:
1. Score transition from "MONMOUTH 0 2nd RIDER 2" to "MONMOUTH 1 2nd RIDER 2".

Figure 14: Human Annotation Analysis: The figures illustrate a scoring transition that is not recognized by the GPT-4V model, highlighting the challenge of using static frames for accurate captioning.

E Video QA Dataset Demonstration

To build the test dataset, we curated a dataset comprising 2,000 videos from each source: WebVid (Bain et al., 2021b), VIDAL (Zhu et al., 2023), ActivityNet (Fabian Caba Heilbron and Niebles, 2015), MSRVTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), TGIF (Jang et al., 2017), and Something-something V2 (SSV2) (Goyal et al., 2017). For each video, ChatGPT was utilized to generate three QA pairs based on the detailed captions. The first three datasets (WebVid, VIDAL, ActivityNet) are classified as in-domain, since the captions and QA pairs derived from these sources are used in the model's training pipeline. Conversely, the remaining datasets (MSRVTT, MSVD, TGIF, SSV2) are classified as out-of-domain, evaluating model's zero-shot QA ability.

Appendix E compares our development benchmark with existing benchmark dataset, we identify several issues with the existing evaluation methods: (1) the auto-generated questions from current benchmarks may be grammatically incorrect or nonsensical, and (2) the answers are limited to a *single* word, which is inadequate for evaluating LMMs in the context of long-form QA. We conduct further analysis on open-ended long-form QA with a proposed development benchmark.

We apply our reward system as described in section 4 and report scores from ChatGPT. A score of ≥ 3 is considered correct for accuracy calculations. The development benchmark reveals insights into (1) the quality of long-form open-ended QA, and (2) in-domain and out-of-domain generalization. Additionally, our development benchmark results correlate with existing benchmarks. We recommend that models be evaluated on the development benchmark first, followed by human evaluation.

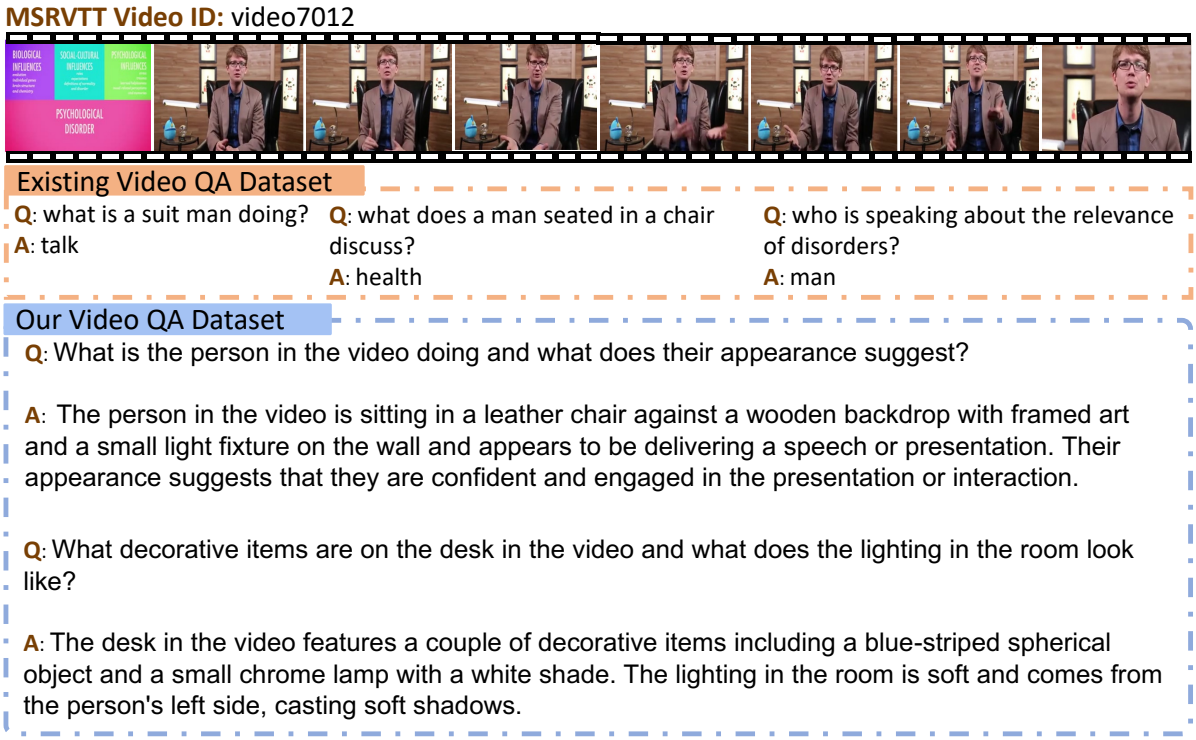


Figure 15: Comparing testing QA in existing benchmark with that in our proposed new benchmark.

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 16: Comparing testing QA in existing benchmark with that in our proposed new benchmark, example 2.

F Additional DPO Results

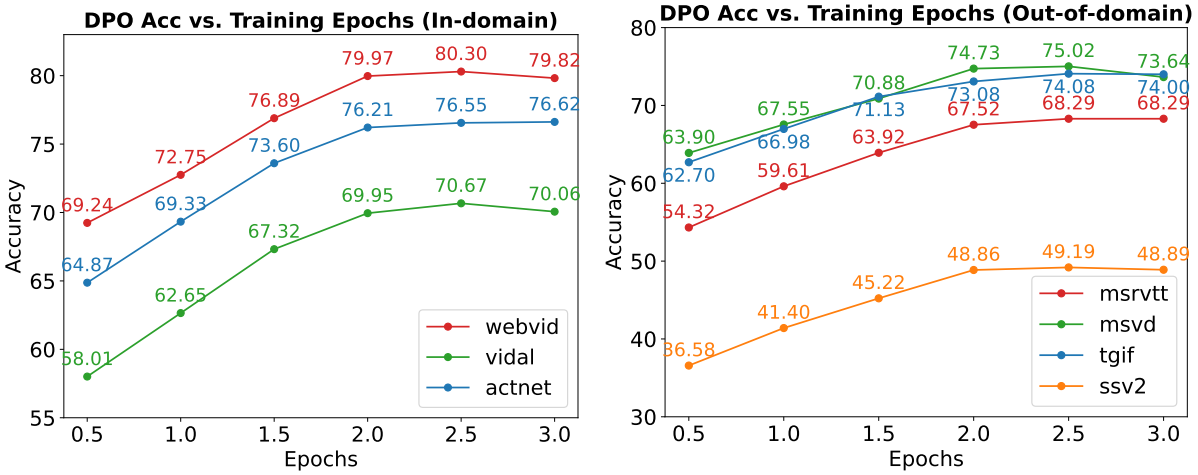


Figure 17: Test Set Accuracy of the DPO Model vs. Training Epochs. The figure illustrates a consistent trend in both in-domain and out-of-domain video QA, with peak performance occurring at approximately epoch 2.5, equivalent to 350 training steps.

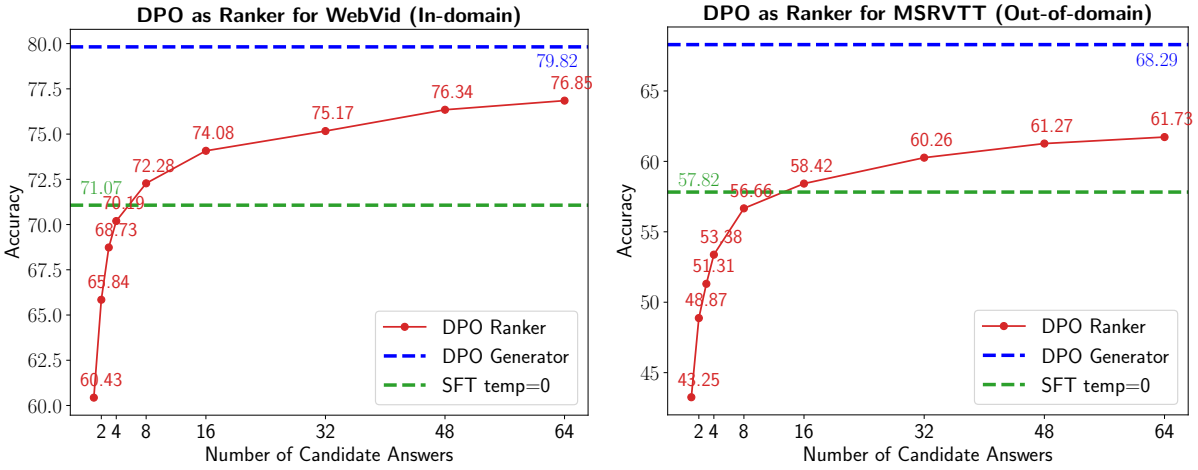


Figure 18: Comparison of DPO Model Performance: Ranker vs. Generator. The DPO model serves as a ranker, assigning reward scores to candidate answers generated by the SFT model with a temperature setting of 1.0. Employing the DPO model directly for answer generation results in superior performance compared to its use as a ranker.

G Prompts for GPT-4V and ChatGPT Queries

```
Picture yourself as a customer service agent managing user-uploaded video. The
uploaded video, captioned with '{}', consists of a seires of images. All the
analysis should be video-level. Your duty is to summarize video content,
highlighting actions and object relationships. Follow this with a detailed
description. The summary briefly covers actions and relationships, while the
detailed description delves into factual, visible details with a logical
structure, considering elements like color, shape, attribute, and count.

Then craft a dialogue between the agent ('A') and the customer ('C') in a manner
suggesting that the agent is actively viewing the video and answering the
customer's questions. Frame questions using 'how many', 'what,' 'how,' 'when,'
'which,' and 'why' to ensure precise and definitive answers, rooted in video
content. Pose varied questions encompassing the visual content, such as object
types, counting objects, object actions, object locations, and relative positions
between objects. Ensure each question has a definite answer, either observed in
the video or confidently determined to be absent. Avoid questions with uncertain
answers.

Ouput format:
Summary: <your summary>
Detail: <your detailed description>
Conversation: <your quesion-answer conversation, clearly labeling the customer
and agent as 'C' and 'A'>
```

Figure 19: GPT-4V prompt for the generation of video summary, detailed caption and conversation generation. We only use detailed caption for experiments.

```
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 20: 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 21: 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 22: 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 23: 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. 22.