LLaVA-Video: Video Instruction Tuning With Synthetic Data

Anonymous authors
Paper under double-blind review

Abstract

The development of video large multimodal models (LMMs) has been hindered by the difficulty of curating large amounts of high-quality raw data from the web. To address this, we consider an alternative approach, creating a high-quality synthetic dataset specifically for video instruction-following, namely LLaVA-Video-178K. This dataset includes key tasks such as detailed captioning, open-ended question-answering (QA), and multiple-choice QA. By training on this proposed dataset, in combination with existing visual instruction tuning data, we introduce LLaVA-Video, a new video LMM. Our experiments demonstrate that LLaVA-Video achieves strong performance across various video benchmarks, highlighting the effectiveness of our dataset. We plan to release the dataset, its generation pipeline, and the model checkpoints.

1 Introduction

We are in an era where large-scale computing and data is crucial for multimodal learning (Li et al., 2024d). A significant recent advancement was introduced by visual instruction tuning (Liu et al., 2024a), which laid the foundation for building a general-purpose visual assistant. Notably, it proposed a data generation pipeline to create high-quality image-language instruction-following data. This pipeline has inspired subsequent researches (Li et al., 2024c;b;a; Lin et al., 2024) aimed at generating diverse image-language instruction data across various visual domains, accelerating the development of visual instruction tuning techniques.

Compared to the construction of image-language instruction-following data, obtaining high-quality video-language instruction-following data is challenging (Zhang et al., 2023; Li et al., 2024e). First, sourcing high-quality videos is difficult. We need to find videos with significant temporal changes that provide more knowledge than what image-language data can offer. However, we have found that most videos in current video-language instruction-following datasets (Chen et al., 2024a; Zhang et al., 2024d) are relatively static. Additionally, these videos are mostly trimmed based on scene changes, resulting in simplified plots. Such simplified video-language instruction-tuning data is inadequate for models to understand videos with complex narratives. Furthermore, current video-language instruction-following datasets often use a very sparse sampling rate for frame annotation. For instance, ShareGPT4Video (Chen et al., 2024a) has an average sampling rate of 0.15, sometimes sampling only 2 frames from a 30-second video. This sparse sampling rate is effective in describing overall scenes but fails to capture detailed movements or changes in the video, resulting in hallucination when detailed descriptions of the video are required.

To overcome these shortcomings, we introduce a comprehensive video instruction-tuning dataset named LLaVA-Video-178K, consisting of 178,510 videos ranging from 0 to 3 minutes. This dataset is enriched with detailed annotations, open-ended questions, and multiple-choice questions, developed through a combination of GPT-4o (OpenAI, 2024) and human efforts. It features four favorable properties: (*i*) **Extensive Video Source:** We conduct a comprehensive survey on the video sources of exsiting video understanding datasets, and conclude 10 major video data sources, from which we start our video data collection by building a video pool. Although there are over 40 video-language datasets, their video data are mainly sourced from 10 datasets (Zhou & Corso, 2017; Xue et al., 2022; Goyal et al., 2017; Caba Heilbron et al., 2015; Kay et al., 2017; Sigurdsson et al., 2016; Wang et al., 2023; Shang et al., 2019; Grauman et al., 2022; Zhu et al., 2023a), covering a wide range of video domains, such as activities, cooking, TV shows, and egocentric views. (*ii*) **Dynamic Untrimmed Video Selection:** From these sources, we use several filtering

logic to select the most dynamic videos from the video data pool. Notably, we select original, untrimmed videos to ensure plot completeness. (*iii*) **Recurrent Detailed Caption Generation Pipeline with Dense Frame Sampling:** We propose a detailed video caption pipeline that operates recurrently, enabling us to generate detailed captions for videos of any length. This pipeline has three levels, each level of description represents a different time-range: from 10 seconds to the entire video length. It is recurrent as the historical description from any level serves as the context for generating new descriptions at any level. Additionally, we adopted a dense sampling strategy of one frame per second to ensure the sampled frames are rich enough to represent the videos. (*iv*) **Diverse Tasks:** Based on the detailed video descriptions, we can generate question-answer pairs. To ensure our questions cover a wide range of scenarios, by referring to the video question-answering dataset, we define 16 question types. We prompt GPT-4o to generate question-answer pairs by referring to these question types, covering open-ended and multi-choice questions.

Based upon the LLaVA-Video-178K dataset, we developed LLaVA-Video. Contrary to previous studies suggesting that training with single frames is sufficient for video-language understanding (Lei et al., 2022), our findings reveal a significant impact of frame count on LLaVA-Video's performance, attributable to the detailed features of LLaVA-Video-178K. Observing this, we explored maximizing frame sampling within the constraints of limited GPU memory. We introduce LLaVA-Video `SlowFast`, a video representation technique that optimally distributes visual tokens across different frames. This approach allows for incorporating up to three times more frames than traditional methods, which allocate an equal number of visual tokens to each frame.

Our contributions are as follows:

- *Video-language Instruction-Following Data*: We present a high-quality dataset *LLaVA-Video-178K* tailored for video instruction-following. It consists of 178K video with 1.3M instruction samples, including detailed captions, free-form and multiple-choice question answering.
- *Video Large Multimodal Models*: We develop *LLaVA-Video*, a series of advanced large video-language models that expand the capabilities of open models in understanding video content.
- *Open-Source*: In an effort to support the development of general-purpose visual assistants, we release our multimodal instruction data, codebase, model checkpoints, and a visual chat demo to the public.

2 Related Work

In this work, our goal is to create a high-quality video-language dataset that goes beyond simple video captions. We aim to improve the ability to follow instructions, which includes detailed video descriptions, open-ended video question-answering, and multiple-choice video question-answering data. We discuss related datasets in Table 1. Previous video-language datasets (Miech et al., 2019) include manually annotated data for various tasks, such as video captions (Chen & Dolan, 2011; Xu et al., 2016; Rohrbach et al., 2015; Anne Hendricks et al., 2017; Caba Heilbron et al., 2015; Zhou & Corso, 2017), and video question-answering (Yu et al., 2019; Zadeh et al., 2019; Xiao et al., 2021). However, manual annotation is expensive and limits the size of such datasets. To address the shortage of data, studies like (Miech et al., 2019; Lee et al., 2021; Zellers et al., 2021; Xue et al., 2022) suggest automatically annotating data using subtitles created by ASR. While this method greatly expands the dataset size to 100 million samples, the subtitles often fail to accurately describe the main video content. Additionally, other studies (Xu et al., 2017; Grunde-McLaughlin et al., 2021; Wu et al., 2024a) use language models (Xu et al., 2017) or question templates (Grunde-McLaughlin et al., 2021; Wu et al., 2024a) to generate question-answer pairs. Although this approach can generate a large number of questions and answers, it often produces poor-quality questions that do not reflect real-world user inquiries. More recent research (Chen et al., 2024b) has prompted video-language models such as BLIP-2 (Li et al., 2023), VideoChat (Li et al., 2024e), Video-LLaMA (Zhang et al., 2023), and MiniGPT-4 (Zhu et al., 2023b) to generate video captions. However, these models are limited in their ability to provide detailed descriptions.

The most related works to ours are the recent AI-generated synthetic video instruction tuning data, Islam et al. (2024) introduced Video ReCap, which recursively annotates video captions. Unlike Video ReCap, each clip-wise (level-1) description in our pipeline is generated with historical context. This ensures that connections from previous events in the video timeline are linked to the current event. LLaVA-Hound (Zhang

Source	Filtering Logic
HD-VILA-100M	①②③④
InternVid	①②③④
ActivityNet	③⑤
VIDAL	①②③④
Charades	③⑤
Ego4D	③⑤
VidOR	③⑤
YouCook2	③⑤⑥
Kinetics-700	③⑤⑥
Sth-sthv2	③⑤⑥

Figure 1: **Video sources in the proposed *LLaVA-Video-178K***. (Left) The relationship between 10 video sources we have utilized and other existing video-language datasets. (Right) Filtering logic for video sources. The detail of filtering logic: ① Sorted by Views, ② Number of scenes greater than 2, ③ Video duration between 5 seconds and 180 seconds, ④ Ratio of scenes to video duration less than or equal to 0.5, ⑤ Resolution greater than 480p, ⑥ 50 samples for each category.

et al., 2024d) and ShareGPT4Video (Chen et al., 2024a), where they have used GPT-4 (OpenAI, 2023) to generate video captions and open-ended video question-answering. Although the quality of the captions and question-answer pairs has significantly improved, the video sources they use are too static to produce high-quality data for instruction-following scenarios. They also only use very sparse frames for prompting GPT-4V, which results in annotations that fail to capture nuanced actions and continuous plots in the videos. Additionally, Shot2Story (Han et al., 2023) and Vript (Han et al., 2023) also employ GPT-4V (OpenAI, 2023) for video captioning. Their outputs, however, include audio details, which are outside the scope of this study.

3 Video Instruction-Following Data Synthesis

A high-quality dataset for video instruction-tuning is crucial for developing effective video-language models. We identify a key factor in building such datasets: ensuring richness and diversity in both video content and its language annotations. We perform comprehensive survey on the existing video benchmarks, covering across various public video captioning and question-answering datasets, then identify ten unique video sources that contribute to over 40 video-language benchmarks. From each source, we select videos that exhibit significant temporal dynamics. To maintain diversity in the annotations, we establish a pipeline capable of generating detailed captions for videos of any length. Additionally, we define 16 types of questions that guide GPT-4o in creating question-answer pairs to assess the perceptual and reasoning skills of the video-language models.

3.1 Video source

One important starting point in building a high-quality video instruction-following dataset is to find a sufficiently diverse pool of video data. From this pool, we can select the qualified videos. In our study of public video-language datasets—including video captioning, video question answering, video summarization, and moment-wise captioning—we noticed that although different datasets focus on various video understanding tasks (*e.g.,* , AGQA (Grunde-McLaughlin et al., 2021) for spatial-temporal relations and STAR (Wu et al., 2024a) for situational reasoning), most are sourced from ten main video sources. For instance, both AGQA and STAR use data from Charades (Sigurdsson et al., 2016). Specifically, these ten sources are HD-VILA-100M (Xue et al., 2022), InternVid-10M (Wang et al., 2023), VidOR (Shang et al., 2019), VIDAL

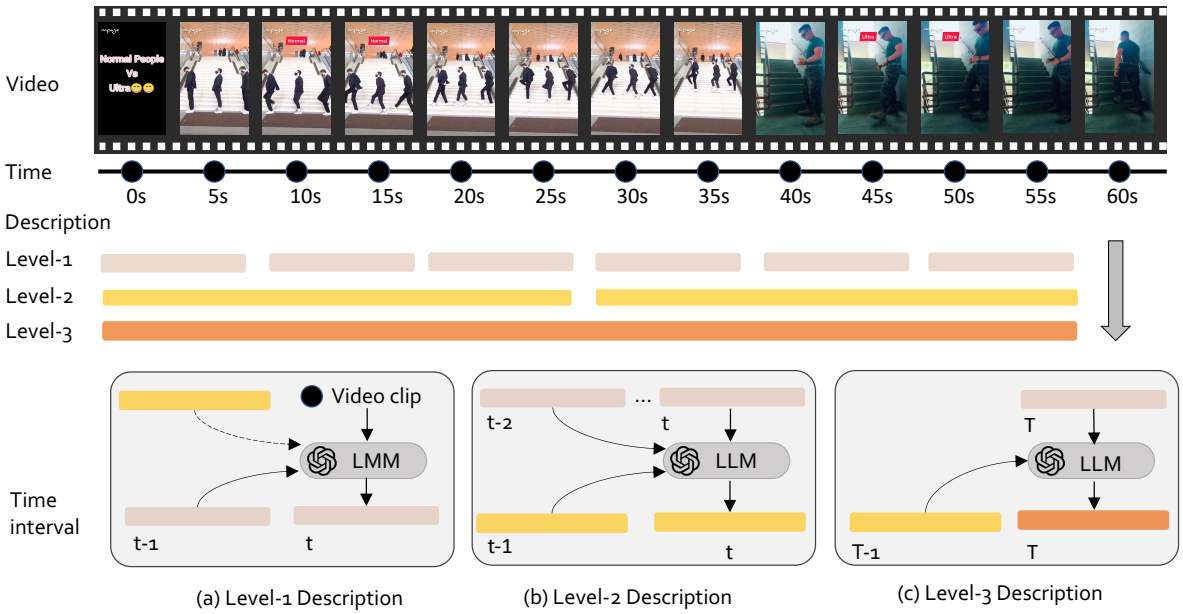


Figure 2: **The video detail description creation pipeline**. A three-level creation pipeline is considered, with each level developed via a recurrent approach. Note that t is the index of time internal at its own level, and T is the last time internal index. (a) To generate the caption for time internal t at level-1, we condition on the current frames in this internal, the caption for time internal $t-1$, and the most recent description summary at level-2 if applicable. (b) To generate caption for time internal t at level-2, we condition on the previous caption at level-2, and captions from three most recent time internals at level-1. (c) To generate the overall caption at the last time internal T at level-3, we condition on the the most recent caption at level-2 and the current caption from level-1.

(YouTube Shorts)(Zhu et al., 2023a), YouCook2(Zhou & Corso, 2017), Charades (Sigurdsson et al., 2016), ActivityNet (Caba Heilbron et al., 2015), Kinetics-700 (Kay et al., 2017), Something-Something v2 (Goyal et al., 2017), and Ego4d (Grauman et al., 2022). These sources offer a wide range of video data from different websites, viewpoints, and domains. The relationship between these ten selected video datasets and others is shown in Fig. 1. The videos from this ten datsets build the video pool for the further video selection. Notably, we use untrimmed videos from each source except for YouCook2 and Kinetics-700. We believe that cutting videos into clips can break the plot continuity, which is essential for understanding the videos.

Based on the video pool, we aim to select dynamic videos. In Figure 1, we outline our criteria for selecting high-quality data. Our main method for identifying dynamic content involves using PySceneDetect, which calculates the number of scenes in a video We found that the number of scenes is a good indicator of video dynamism. Additionally, we have designed a specific approach ④ to exclude videos that mainly contain "slides."

3.2 Video Detail Description

Automated Generation For selected videos, we use GPT-4o (OpenAI, 2024) to systematically describe their content. We start by sampling video frames at one frame per second (fps). However, due to the input size constraints of GPT-4o, we cannot use all sampled frames. Instead, we describe the videos sequentially, as shown in Fig 2. We create descriptions at three distinct levels, detailed below.

- *Level-1 Description*: Every 10 seconds, we provide a level-1 description that outlines the events in that segment. This description considers: frames from the current clip and historical context, which includes all recent level-1 descriptions not yet summarized into a level-2 description and the latest level-2 description.

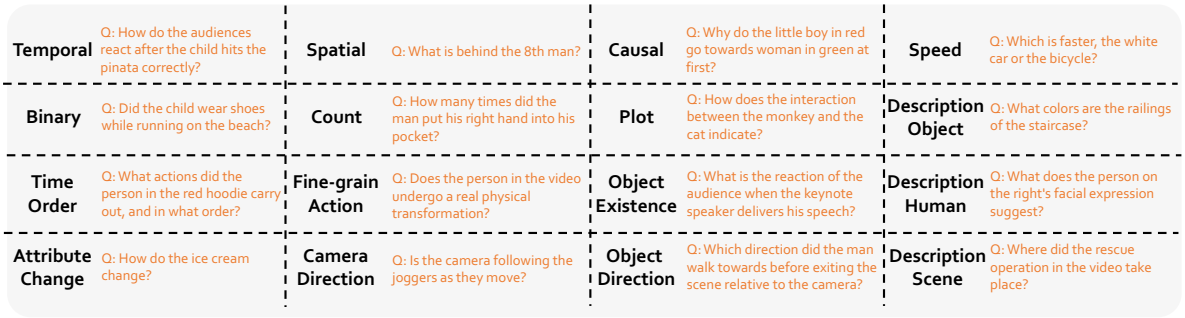


Figure 3: Question types for video question answering in data creation. For each type, we provide its name and an example question.

- *Level-2 Description*: Every 30 seconds, we creat a level-2 summary of the entire video plot up to that point. This is based on the last three level-1 descriptions, covering the most recent 30 seconds; and the latest level-2 description.
- *Level-3 Description*: At the video's end, we generate a level-3 description to encapsulate the entire video. The inputs for this description are the recent level-1 descriptions not yet summarized, covering the last moments of the plot after the recent summary; and the latest level-2 description.

3.3 Video Question Answering

Question Type definition In addition to detailed video descriptions, our dataset includes a variety of question-answer pairs designed for complex interactions. This setup improves the video understanding model's ability to handle real-life queries. We refer to public video question-answering benchmarks (Xiao et al., 2021; Yu et al., 2019; khattak et al., 2024; Liu et al., 2024b) to organize these questions into 16 specific categories, as shown in Fig. 3.

Automated Generation Given a detailed video description, we use GPT-4o to generate at most one question-answer pair for each type of question. The prompts include: (1) The task definition for the current question type. (2) In-context examples for this type, which include three video descriptions and their three question-answer pairs of this specific type. (3) The detailed video description for the current video. We instruct GPT-4o to return *None* if it cannot generate question-answer pairs for a specific question type.

Filtering. To filter out the generated question-answer pairs, we apply the following strategy: (1) remove duplicates using the sentence-transformer (Reimers & Gurevych, 2020), (2) discard answers that begin with phrases like "does not specify," "does not mention," "does not specifically," "does not depict," or "does not show."

3.4 Dataset Statistics

Overview. We carefully select from our collected data sources to form a balanced and comprehensive collection, resulting in a total of 178K videos and 1.3M instruction-following samples. This includes 178K captions, 960K open-ended QAs, and 196K multiple-choice QAs.

Dataset Comparison We provide a comparison of high-quality instruction following video-language datasets, with a focus on synthetic data created with strong AI models, as shown in Table 1. (*i*) *A broad collection of dynamic videos.* In terms of video sources, although LLaVA-Hound (Zhang et al., 2024d) contains the largest number of videos, 44% of its video data are sourced from WebVid (Bain et al., 2021), where most videos are static. ShareGPT4Video (Chen et al., 2024a) includes 30% of its videos from Pexels, Pixabay, and Mixkit, which are aesthetically good but also mostly static. Additionally, the majority of its videos come from Panda-70M, which are short clips from longer videos—suggesting simpler plots. In contrast, we carefully select video sources that offer dynamic, untrimmed videos with complex plots, which are crucial for

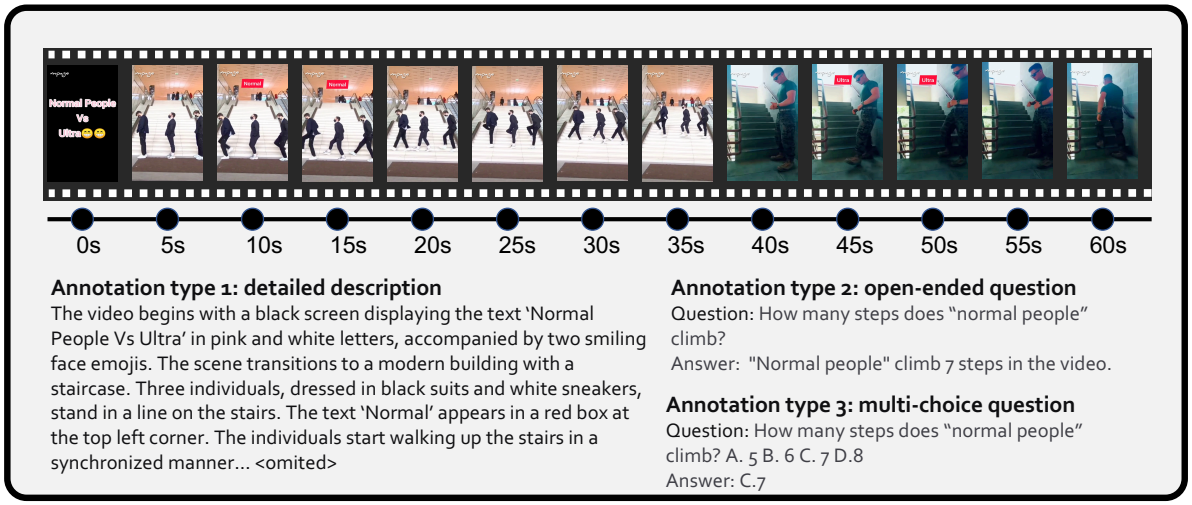
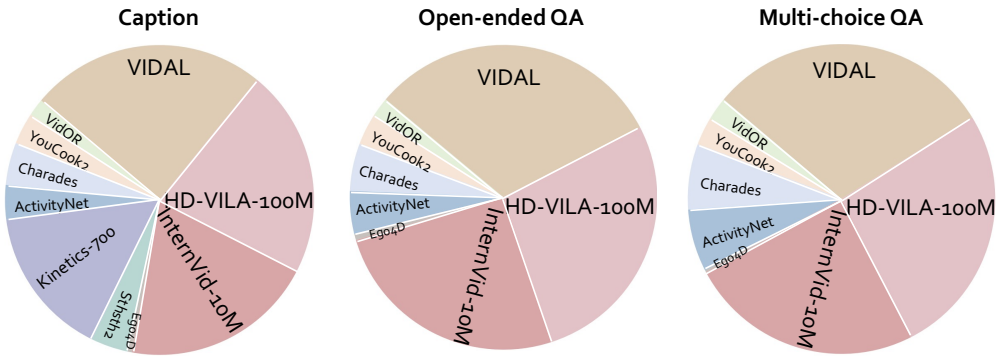


Figure 4: One example to illustrate the video instruction-following data.



Dataset	#Caption	#Open-Ended	#Multi-Choice	Dataset	#Caption	#Open-Ended	#Multi-Choice
VidOR	4,018	19,875	4,773	Sthsth2	8,700	0	0
YouCook2	7,411	32,143	5,776	Ego4D	1,065	5,912	520
Charades	9,803	48,187	13,401	InternVid-10M	45,000	245,840	48,246
ActivityNet	7,953	44,100	12,771	HD-VILA-100M	48,260	263,652	51,743
Kinetics-700	34,998	0	0	VIDAL	55,000	300,472	58,968

Figure 5: Distribution of data across different datasets and question types (Caption, Open-ended, and Multi-Choice).

developing a powerful video understanding model.[1] (*ii*) *High frames per second.* Regarding frame sampling in language annotations, the proposed datasest considers 1 FPS, while other datasets consider much lower FPS. LLaVA-Hound uniformly samples 10 frames from videos of any length. The average FPS is 0.008, which may miss some fine details. ShareGPT4Video picks key frames using CLIP (Radford et al., 2021) based on frame uniqueness. This method might also miss subtle changes in the video because CLIP embeddings do not capture fine-grained dynamics well. Our method samples FPS=1 without using key frame selection algorithms, ensuring the detailed temproal information can be expressed in annotations and high coverage. (*iii*) *Diverse tasks.* The proposed dataset considers three common task types, including caption, free-form and closed-form QA, while existing datasets only consider a subset. Meanwhile, the quality and numbers of samples in our dataset is higher.

[1]Example videos: WebVid,Pixabay,Pexels,Mixkit.

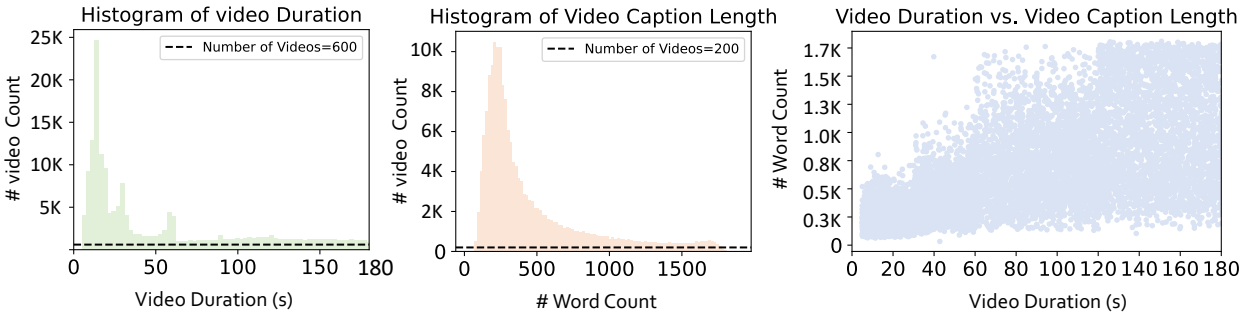


Figure 6: (Left) Visualization of the video duration. (Middle) Visualization of the number of words in the video caption. (Right) Visualization of caption length versus video duration.

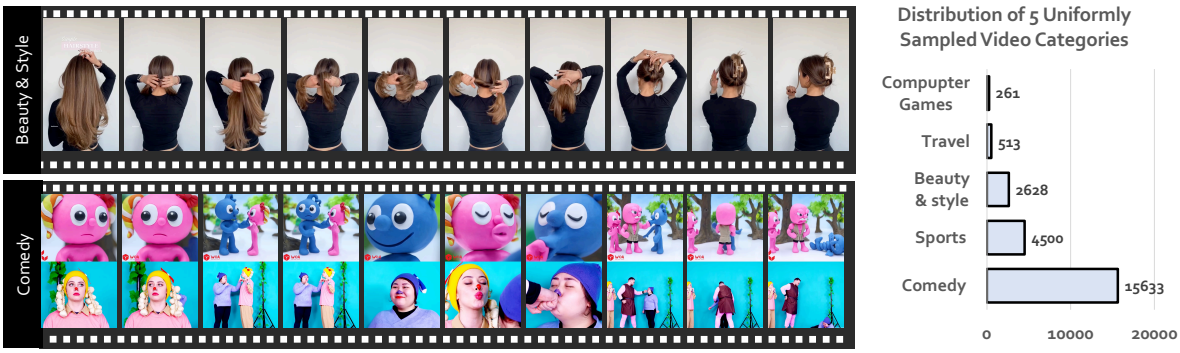


Figure 7: (Left) Display of YouTube Shorts across four video categories. (Right) Distribution of 5 uniformly chosen video categories.

4 Experiments

We conducted evaluations for the LLaVA-Video models across all benchmarks using LMMs-Eval (Zhang et al., 2024a) to ensure standardization and reproducibility. To fairly compare with other leading video LMMs, we primarily used results from original papers. When results were not available, we integrated the models into LMMs-Eval and assessed them under consistent settings. Following LLaVA-OneVision (Li et al., 2024c), we employed SigLIP (Zhai et al., 2023) as our vision encoder, and Qwen2 (Yang et al., 2024) as the LLM. The LLaVA-Video model builds on the single-image (SI) stage checkpoint from the LLaVA-OneVision model (Li et al., 2024c), which was trained using only image data.

Video Representations Following the classic SlowFast idea in video representations (Feichtenhofer et al., 2019; Xu et al., 2024b; Huang et al., 2024), we develop *LLaVA-Video* $_{\texttt{SlowFast}}$ to optimize the balance between the number of frames and the count of visual tokens, within the budget of the limited context window in LLM and GPU memory for video representation. Please refer to Appendix 7 for detailed information. Specifically, we represent each video as a sequence with maximum T frames. Each frame is represented in M tokens. we categorize the frames into two groups, based on the a strike rate s, where the every s frames are uniformly selected to form the *slow* frame group, and the rest of the frames are consdiered as the *fast* frame group. Note that a special case $s = 1$ leads to only one group, reducing the SlowFast representation to the original simple representation. For each group, we apply different pooling rate using Pytorch function pooling `avg_pool2d()`. $p \times p$ pooling and $2p \times 2p$ pooling for slow and fast frames, respectively. To summarize, we paramterize the video representation configuration as $\mathcal{V} = (T, M, s, p)$. The total number of tokens is $\#tokens = \lfloor T/s \rfloor \times \lfloor M/p^2 \rfloor + (T - \lfloor T/s \rfloor) \times \lfloor M/4p^2 \rfloor$

Table 1: **Comparison of LLaVA-Video-178K and other video-language datasets**. Average FPS represents the average number of frames per second that are used to prompt GPT-4o/GPT-4V for annotation. ★ VIDAL, WebVid, ActivityNet. ■ Panda-70M, Pexels, Pixabay, Mixkit, BDD100K, Ego4d. ✪ HD-VILA-100M, Kinetics-700M, Ego4D, VidOR, InternVid, YouCook2, ActivityNet, Sth-sthv2, VIDAL, Charades.

	Text	Video Source	#Video	Total Video Length	Average FPS	#Caption	#OE QA	#MC QA
LLaVA-Hound	GPT-4V	★	900K	3Khr	0.008	900K	900K	0
ShareGPT4Video	GPT-4V	■	40K	0.2Khr	0.15	40K	0	0
LLaVA-Video-178K	GPT-4o	✪	178K	2Khr	1	178K	960K	196K

Evaluation Benchmarks. For full evaluation, we consdier 11 video benchmarks. conducted tests across various video captioning , video open-ended question-answering and video multiple-choice question-answering benchmarks, including ActivityNet-QA (Yu et al., 2019), which features human-annotated action-related QA pairs from the ActivityNet dataset. We also utilized LongVideoBench (Wu et al., 2024b), EgoSchema (Mangalam et al., 2024), and MLVU (Zhou et al., 2024) for long video understanding, PerceptionTest (Pătrăucean et al., 2023) for assessing fine-grained perception skills, and VideoMME (Fu et al., 2024) and NExT-QA (Xiao et al., 2021) for diverse video domains and durations. Additional tests included VideoDetailCaption (LMMs-Lab, 2024), Dream-1K (Wang et al., 2024), Video-ChatGPT (Maaz et al., 2024) for detailed video descriptions, TemporalBench Cai et al. (2024) for fine-grained temporal understanding.

For ablation studies in . 4.2 and Sec. 4.3, we conduct evaluation across 4 datasets. NExT-QA (Xiao et al., 2021) and PerceptionTest (Pătrăucean et al., 2023), which use training data from the LLaVA-Video-178K, are treated as in-domain datasets. Conversely, VideoMME (Fu et al., 2024) and EgoSchema (Mangalam et al., 2024) are consider as zero-shot datasets.

4.1 Overall Results

We fine-tune LLaVA-OneVision (SI) on the joint dataset of video and image data. Specifically, we added video data from the LLaVA-Video-178K dataset and four public datasets: ActivityNet-QA (Yu et al., 2019), NExT-QA (Xiao et al., 2021), PerceptionTest (Pătrăucean et al., 2023), and LLaVA-Hound-255K (Zhang et al., 2024d), focusing on videos shorter than three minutes. These datasets were selected to improve our model's performance, contributing to a total of 1.6 million video-language samples, which include 193,510 video descriptions, 1,241,412 open-ended questions, and 215,625 multiple-choice questions. Remarkably, 92.2% of the video descriptions, 77.4% of the open-ended questions, and 90.9% of the multiple-choice questions were newly annotated. Additionally, we used 1.1 million image-language pairs from the LLaVA-OneVision model (Li et al., 2024c). We consider the same video representation configurations for the training and inference stages. On 128 NVIDIA H100 GPUs, the video representations for LLaVA-Video-7B and LLaVA-Video-72B are $\mathcal{V} = (64, 679, 1, 2)$ and $\mathcal{V} = (64, 679, 3, 2)$, respectively.

In Table 2, we compare the performance of different models on various video benchmarks. The 72B model performs as well as the commercial, closed-source model Gemini-1.5-Flash (Team et al., 2023), highlighting the effectiveness of open-source efforts in achieving comparable results. The LLaVA-Video-7B model outperforms the previous top model, LLaVA-OV-7B, in seven out of ten datasets. Analysis of individual datasets shows some noteworthy trends. For instance, on benchmarks like MLVU, LongVideoBench, and VideoMME, which primarily use video data from YouTube, this improvement may be due to the inclusion of extensive YouTube data in LLaVA-Video-178K, as illustrated in Fig. 5. Additionally, the improvement on ActivityNet-QA is small; this could be because many questions in ActivityNet-QA, such as "What's the color of the ball?" can be answered by viewing a single frame. The visibility of the ball from the beginning to the end of the video means understanding the video sequence is unnecessary, so LLaVA-Video-178K offers little advantage in this context. We find that LLaVA-Video-7B is notably weaker in the specialized task of EgoSchema, an ego-centric dataset. This weakness may be due to a significant reduction in the proportion of ego-centric data in the training dataset of LLaVA-Video. However, this impact is less pronounced in larger models, as demonstrated by the LLaVA-Video-72B model's superior performance over LLaVA-OV-72B in EgoSchema.

Table 2: LLaVA-Video performance on video benchmarks. We report the score out of 5 for VideoDC, VideoChatGPT while other results are reported in accuracy. All results are reported as 0-shot accuracy. *indicates that the training set has been observed in our data mixture.

Model	Caption		Open-Ended Q&A		Multi-Choice Q&A							
	VideoDC	Dream-1K	ActNet-QA	VideoChatGPT	EgoSchema	MLVU	MVBench	NExT-QA	PerceptionTest	LongVideoBench	TemporalBench	VideoMME
	test	test	test	test	test	m-avg	test	mc	val	val	m-acc	wo/w-subs
Proprietary models												
GPT-4o (OpenAI, 2024)	-	39.2	-	-	-	64.6	-	-	-	66.7	35.3	71.9/77.2
Gemini-1.5-Pro (Team et al., 2023)	-	36.2	57.5	-	72.2	-	-	-	-	64.0	25.6	75.0/81.3
Open-source models												
VILA-40B (Lin et al., 2024)	3.37	33.2	58.0	3.36	58.0	-	-	67.9	54.0	-	-	60.1/61.1
PLLaVA-34B (Xu et al., 2024a)	-	28.2	60.9	3.48	-	-	58.1	-	-	53.2	-	-
LongVA-7B (Zhang et al., 2024c)	3.14	-	50.0	3.20	-	56.3	-	68.3	-	-	-	52.6/54.3
IXC-2.5-7B (Zhang et al., 2024b)	-	-	52.8	3.46	-	37.3	69.1	71.0	34.4	-	16.7	55.8/58.8
LLaVA-OV-7B (Li et al., 2024c)	3.75	31.7	56.6	3.51	60.1	64.7	56.7	79.4*	57.1	56.5	18.7	58.2/61.5
VideoLLaMA2-72B (Cheng et al., 2024)	-	27.1	55.2	3.16	63.9	61.2	62.0	-	-	-	-	61.4/63.1
LLaVA-OV-72B (Li et al., 2024c)	3.60	33.2	62.3	3.62	62.0	68.0	59.4	80.2*	66.9	61.3	26.6	66.2/69.5
LLaVA-Video-7B	3.66	32.5	56.5*	3.52	57.3	70.8	58.6	83.2*	67.9*	58.2	22.9	63.3/69.7
LLaVA-Video-72B	3.73	34.0	63.4*	3.62	65.6	74.4	64.1	85.4*	74.3*	61.9	33.7	70.5/76.9

4.2 Dataset Ablation

Note that the training set for LLaVA-Video includes six datasets: LLaVA-Video-178K, LLaVA-Hound (Zhang et al., 2024d), NExT-QA (Xiao et al., 2021), ActivityNet-QA (Yu et al., 2019), PerceptionTest (Pătrăucean et al., 2023), and image data from LLaVA-OneVision (Li et al., 2024c). In this section, we conduct ablation studies to assess the impact of each dataset. We separately fine-tune the LLaVA-OneVision (SI) model for each experimental setting, progressively adding datasets to the baseline.

The results are presented in Table 3. Initially, we used a basic model trained solely on the LLaVA-Hound dataset as our baseline. Compared to this baseline, adding the LLaVA-Video-178K dataset significantly improved performance, enhancing scores in both in-domain and out-of-domain tasks. Specifically, we observed a 31.9-point increase in NExT-QA scores and a 9.1-point rise in VideoMME scores. Furthermore, including the PerceptionTest dataset enhanced its associated task. Additionally, integrating high-quality image data provided modest benefits on EgoSchema.

4.3 Dataset Comparison

We conduct two ablation studies to analyze our dataset and training strategy. In Table 4, we compared three datasets where the language annotations are from GPT-4V/GPT-4o. For each experiment, we fine-tune the LLaVA-OneVision (SI) model separately on each specific dataset setting.

Two group of experiments are considered to assess the data quality of LLaVA-Video-178K compare to LLaVA-Hound and ShareGPT4Video. In the first group, to compare LLaVA-Video-178K with LLaVA-Hound, we randomly selected 900K open-ended questions to match the number in LLaVA-Hound. We included all captions and did not sample the multiple-choice questions. In the second group, comparing LLaVA-Video-178K to ShareGPT4Video, we randomly sampled 40K video captions to align with those in ShareGPT4Video. Since ShareGPT4Video lacks open-ended and multiple-choice questions, we supplemented with annotations from NExT-QA, PerceptionTest, and ActivityNet-QA. In the first group of Table 4, we compare LLaVA-Video-178K

Table 3: Ablation study on the LLaVA-Video model with various configurations of training data. Three Q&A datasets indicate: NExT-QA, ActivityNet-QA and PerceptionTest.

Method	in-domain		out-of-domain	
	NExT-QA	PercepTest	EgoSchema	VideoMME
	mc	val	test	wo
LLaVA-Hound	64.4	51.4	51.0	54.1
+LLaVA-Video-178K	80.1	57.1	56.5	63.2
+Three Q&A datasets	80.1	69.0	55.6	61.9
+LLaVA-OV (images)	83.2	67.9	57.3	63.4

Table 4: Comparison of LLaVA-Video-178K and other video instruction-following datasets.

	#Cap	#OE	#MC	in-domain NExT-QA	in-domain PercepTest	out-of-domain EgoSchema	out-of-domain VideoMME
				mc	val	test	wo
LLaVA-Hound	900K	900k	0	39.8	53.1	25.8	55.2
LLaVA-V-178K	178K	900k	0	73.2	55.9	49.8	59.6
ShareGPT4Video	40K	40K	19K	69.6	55.2	58.9	51.0
LLaVA-V-178K	40K	40K	19K	75.8	55.4	55.8	53.5

with LLaVA-Hound. Although LLaVA-Hound has more captions than LLaVA-Video-178K, our results are still better. The quality of LLaVA-Hound is limited due to two main issues: (1) Static video: Its primary video source is WebVid (Bain et al., 2021), which tends to have relatively static content. (2) Sparse sampling: its sampling rate of 10 frames per video leads to annotations that do not fully capture the complete plot of the video. This underscores that the quality of video instruction-following data is more important than its quantity. Additionally, the second experiment group in Table 4 shows that the model trained with LLaVA-Video-178K outperforms that of ShareGPT4Video, highlighting the superiority of our data's quality.

5 Conclusion

This study introduces the LLaVA-Video-178K dataset, a high-quality synthetic dataset for video-language instruction-following. It is favored for its dense frame sampling rate in longer, untrimmed videos, covering diverse tasks such as captioning, open-ended and multi-choice QA. By training on the joint dataset of LLaVA-Video-178K with existing visual instruction tuning data, we developed a new model family, LLaVA-Video, which considers video representation to effectively use GPU resources. This allows us to include more frames in the training process. The experimental results have demonstrated the effectiveness of the proposed synthetic dataset, and LLaVA-Video models have achieved excellent performance on a wide range of video benchmarks.

6 Limitations

The videos in LLaVA-Video-178K are sourced from various platforms. This diversity introduces potential biases inherent in these sources. Furthermore, there is a concern regarding the potential skew in the question-answer pairs, possibly influenced by the annotators' perspectives.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017. 2

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 5, 10

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015. 1, 2, 4

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 8

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. 2

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024a. 1, 3, 5

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024b. 2

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. URL https://arxiv.org/abs/2406.07476. 9

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019. 7

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 8

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017. 1, 4

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022. 1, 4

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021. 2, 3

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2311.17043*, 2023. 3

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *arXiv preprint arXiv:2403.19046*, 2024. 7

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. *arXiv preprint arXiv:2402.13250*, 2024. 2

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 4

Muhammad Uzair khattak, Muhammad Ferjad Naeem, Jameel Hassan, Naseer Muzzamal, Federcio Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms. *arXiv:2405.03690*, 2024. 5

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10274–10284, 2021. 2

Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024a. URL https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/. 1

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024b. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/. 1

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024c. 1, 7, 8, 9

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 2024d. 1

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597. 2

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024e. URL https://arxiv.org/abs/2305.06355. 1, 2

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024. 1, 9

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a. 1

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024b. 5

LMMs-Lab. Video detail caption, 2024. URL https://huggingface.co/datasets/lmms-lab/VideoDetailCaption. 8

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 8

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 8

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2

OpenAI. Gpt-4v. https://openai.com/index/gpt-4v-system-card/, 2023. 3

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 1, 4, 9

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HYEGXFnPoq. 8, 9

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021. 6

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813. 5

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015. 2

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 279–287. ACM, 2019. 1, 3

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 510–526. Springer, 2016. 1, 3, 4

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 8, 9

Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models, 2024. URL https://arxiv.org/abs/2407.00634. 8

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 multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 3

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024a. 2, 3

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024b. URL https://arxiv.org/abs/2407.15754. 8

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021. 2, 5, 8, 9

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 2

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. 2

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024a. 9

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024b. 7

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 7

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 *AAAI*, pp. 9127–9134, 2019. 2, 5, 8, 9

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019. 2

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651, 2021. 2

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. 7

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL https://arxiv.org/abs/2306.02858. 1, 2

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024a. 7

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024b. 9

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024c. 9

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward, 2024d. 1, 2, 5, 8, 9

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 8

Luowei Zhou and Jason J. Corso. Youcookii dataset. 2017. URL https://api.semanticscholar.org/CorpusID:19774151. 1, 2, 4

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023a. 1, 4

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023b. 2