Long as a Tale: Unlocking Long-Video Context in VLMs with Hour-long Data

Jinghan Zhang, Caorui Li, Tongyao Zhu, Yitang Gao, Yiyun Deng,
Shiqi Chen, Chang Ma, Junxian He

Abstract

Current Vision Language Models (VLMs) often struggle with long video understanding due to scarcity of annotated long video data. In this work, we propose a multi-step workflow to construct hour-long, high-quality video instruction tuning data in an effective and efficient manner. Specifically, we begin by sourcing diverse, long-duration videos followed by a rigorous filtering step. We then leverage and process the corresponding subtitles to generate detailed textual descriptions for these videos. The resulting dataset, which we name LONGVIDEOSET*, comprises 4,500 videos, each ranging from 20 to 60 minutes in duration. Each video is paired with a rich textual description averaging approximately 2,700 tokens. This contrasts starkly with existing public video instruction tuning datasets, which predominantly consist of much shorter videos—typically only seconds to a minute long—accompanied by 10s of textual tokens. Using* LONGVIDEOSET*, we enhance the long video comprehension capabilities of various SOTA VLMs, enabling them to process and understand hour-long videos. Remarkably, with just 4,500 examples from* LONGVIDEOSET*, our approach achieves substantial performance improvements across multiple long video understanding benchmarks, with gains of up to 6.0 absolute points. We publicly release* LONGVIDEOSET *alongside the data construction pipeline, offering an efficient way to boost current SOTA VLMs in hour-long video understanding. Moreover, this pipeline provides a replicable recipe for expanding the dataset size, paving the way for even greater advancements in this domain.*

1. Introduction

Despite the emergence of strong video-language understanding models [2, 10], processing and comprehending hour-long videos remains a significant challenge for Vision-Language Models (VLMs) [12, 47]. Most current VLMs are constrained to processing a limited number of frames, typically for short-duration videos ranging from a few seconds to about one minute—much shorter than the duration of videos in real-world scenarios. This limitation arises primarily due to two key issues: (1) Processing long videos while maintaining a reasonable frames-per-second (fps) sampling rate necessitates a substantial increase in the number of frames to prevent critical information loss. This, in turn, translates into a large number of input tokens for the models. For instance, sampling at 0.5 fps would result in 600 frames for a 20-minute video, where each frame is encoded into hundreds of visual tokens. Such scaling presents significant challenges for the efficiency and context length of the underlying language models. (2) Although recent works have attempted to extend the context length of language models through text-based methods such as continual pretraining on text data, they often lack an effective video instruction tuning stage to capture the global dependencies inherent in long videos [50, 54]. Typically, these models are fine-tuned on image-text datasets or on short videos of only one or two minutes, which limits their ability to generalize to longer video contexts. These issues are illustrated in Figure 1.

To tackle these challenges, we propose an automatic pipeline to construct large-scale, hour-long video instruction tuning datasets, equipping existing VLMs with the ability to understand long videos. Specifically, we begin by creating a world knowledge taxonomy to ensure diversity during the video data collection process. This taxonomy spans a wide range of domains, including TV shows, documentaries, and vlogs, as shown in Figure 3. Guided by these categories, we curate diverse videos from YouTube. We then apply a fine-grained data cleaning process, which involves multiple filtering steps to enhance data quality and ensure the videos are dynamic and accompanied by accessible subtitles, as depicted in Figure 2. To generate annotations for these videos, we utilize the accompanying subtitles, which are typically transcribed from the audio. However, as these subtitles are often noisy, inconsistent, and lacking in fluency, we employ LLMs to refine them. The LLMs transform the raw subtitles into coherent, detailed, and fluent video descriptions. Unlike common practices that rely on

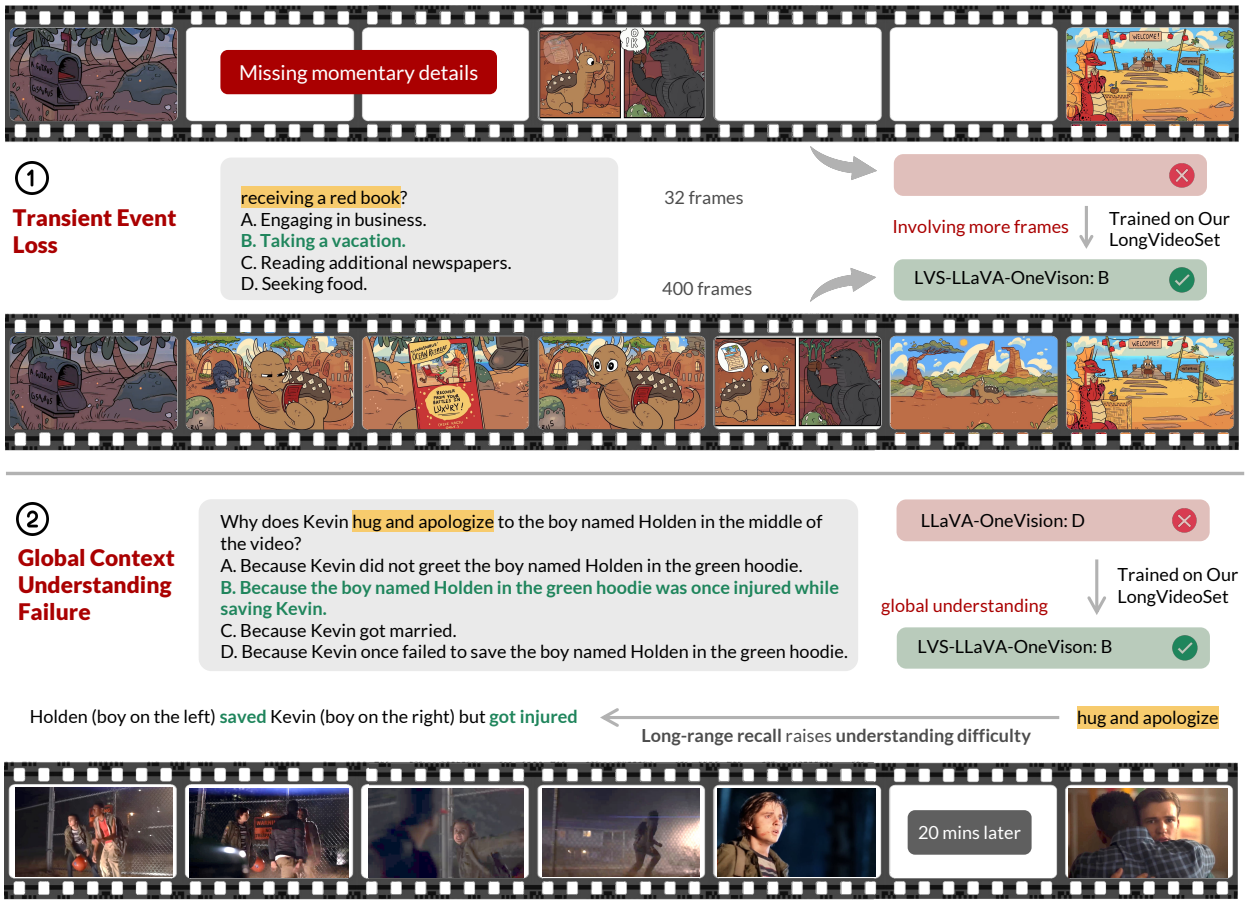


Figure 1. This first case on the top illustrates that the model fails to 'see' the frame referenced by the question due to insufficient sampling. In the second example, the model fails to understand the long-range relationship between events due to its inherent limitations in handling global context.

multimodal LLMs to synthesize text annotations [38, 39], our approach is both cost-effective and scalable, facilitating the construction of large datasets efficiently.

We refer to the resulting dataset as LONGVIDEOSET, which comprises 4,500 videos, each ranging from 20 to 60 minutes in duration, accompanied by detailed video descriptions averaging 2,700 tokens in length. This dataset stands in stark contrast to existing public video instruction datasets, where videos are typically only a few seconds to 1–2 minutes long, with textual annotations averaging around 200 tokens, as summarized in Table 1. While our data construction pipeline is fully automated and can scale up the dataset size at low computational cost, we deliberately limit the dataset to 4,500 examples in this work, aiming to enhance the long-video understanding abilities of existing models in a resource-efficient manner.

In our experiments, we use LONGVIDEOSET to improve the long-video comprehension capabilities of several state-of-the-art VLMs. Our approach involves a two-stage training process. First, we extend the context window of the VLMs by continuing pretraining on long textual data, enabling a text-based expansion of the context length. In the second stage, we use LONGVIDEOSET to fine-tune the models. Experimental results demonstrate that LONGVIDEOSET significantly enhances the performance of LLaVA-OneVision-7B [17] on long-video understanding benchmarks such as Video-MME [12] and MLVU [59], achieving up to a 6-point absolute improvement. Additionally, applying our approach to other models yields consistent performance gains. For instance, on LongVA [54], our method improves the Video-MME performance by 5.6 points and boosts performance on the long set by 6.7 points.

2. Constructing LONGVIDEOSET

Figure 1 demonstrates the two issues of current VLMs when dealing with long videos: they either miss critical details due to low fps sampling rate, or lack proper long video instruction tuning dataset to learn global context understand-

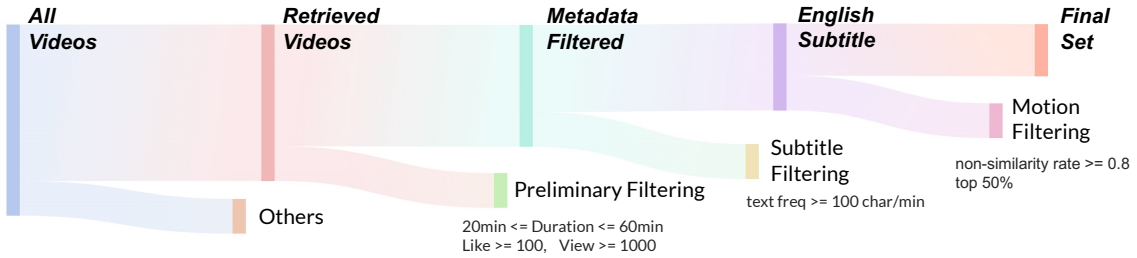


Figure 2. Our Data collection pipeline. Our collection includes three steps: 1). Preliminary Filtering by handcrafted rules: We only keep videos with a duration from 20 to 60 minutes, more than 100 likes, and more than 1000 views. 2). Subtitle Filtering: only keep videos with subtitles of more than 100 characters per minute. 3). Motion Filtering: videos should have at least 50% of frames that are sufficiently different from neighboring frames.

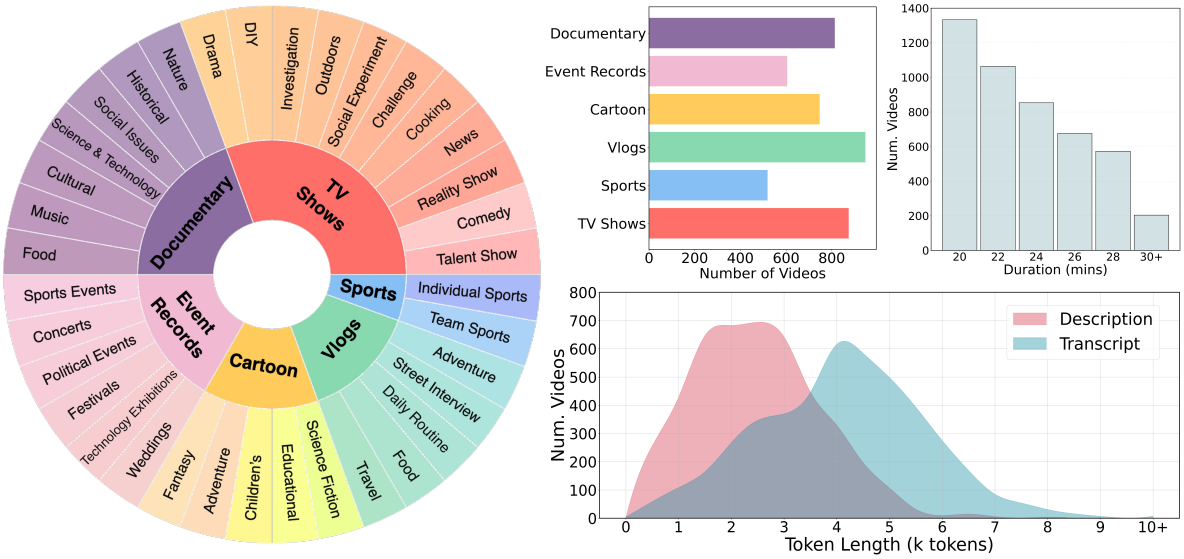


Figure 3. Categories and corresponding subcategories covered in LONGVIDEOSET, together with the lengthy feature of both videos and text descriptions.

ing. While the first issue can be mitigated by extending the context length of the model through text-based pre-training [9, 54], in this section, we focus on tackling the second issue to construct a high-quality, long video instruction tuning dataset automatically, which we name as LONGVIDEOSET. The overall pipeline is illustrated in Figure 2, which we detail next.

2.1. Long Video Collection

Our dataset collection pipeline consists of two main steps: The first step is collecting long videos. We begin by curating a top-down taxonomy based on input from both humans and GPT-4o [28]. This process allows us to manually categorize videos into six main categories and 37 subcategories, maximizing coverage of diverse world knowledge. Ultimately, we leverage **[gpt4o? –JH]** to propose a total of 105 search queries to crawl videos from Youtube, as illustrated in Figure 3 and Table 7. Next, we retrieve videos and their metadata using the YouTube's search API. To ensure uniqueness, we apply de-duplication across all categories, resulting in a collection of approximately 80,000 real-world videos.

The next step is *fine-grained video filtering*, which includes three types of filtering: 1) *Metadata Filtering*. We retain videos with durations ranging from 20 to 60 minutes to keep the long videos. To ensure quality, we retain only those videos with viewer counts and likes exceeding 1000 and 100 respectively. 2) *Subtitle Filtering*. We set a minimum subtitle density threshold of 100 characters per minute to exclude videos with insufficient subtitle information, as they provide limited supervision for captioning. Additionally, we include only videos with English subtitles to avoid potential translation issues. 3) *Motion Filtering*. We aim to filter out videos that lack significant actions or motion

Dataset	Caption	Duration	Text Length	Video Source
VideoChatGPT-100K [25]	Human/BLIP-2/GPT-3.5	123.4	68.0	[5]
LLaVA-Hound [55]	GPT4V Annotated	52.4s	37.6	[3, 5, 62]
ShareGPT4Video [6]	GPT4V Annotated	26.6	273.3	[8, 14, 53], Pexels, Pixabay, Mixkit.
LLaVA-Video-178K [57]	GPT4o Annotated	40.5	-	[5, 13, 14, 16, 31, 35, 44, 52, 61, 62]
VISTA400K [30]	Gemini Rephrased	48.6	-	[8, 11, 15, 27, 44]
LONGVIDEOSET	Transcript Rephrased	**1444.8**	**2756.8**	**New videos sourced from YouTube**

Table 1. The comparison of current Video Training datasets. Here, "Duration" is the average duration of the data samples, measured in seconds. and "Text Length" refers to average token number annotated for each video, measured in tokens. Previous video sources includes: VIDAL [62], HD-VILA-100M [52], Kinetics-700M [16], Ego4D [14], VidOR [31], InternVid [44], YouCook2 [61], ActivityNet [5], Sth-sthv2 [13], Charades [35], Panda-70M [8], WebVid [3], BDD100K [53], MiraData [15], FineVideo [11], OpenVid-1M [27].

changes in this step. Specifically, we encode visual embeddings for each frame using CLIP [29] and calculate cosine similarity scores with neighboring frames. Only videos with at least 50% of frames exceeding the threshold score of 0.8 are retained. After thorough data cleaning, 5.7% of the retrieved videos are retained in LONGVIDEOSET, yielding a curated set of 4,552 high-quality videos with detailed subtitles.

2.2. Video Captioning

The most critical step in VLM training is aligning the visual modality with text, requiring pairwise video-text data with detailed and grounded video captions. Recent approaches often generate captions for shorter clips and merge them together [4, 7, 42, 45]. However, this process is prone to model hallucination, resulting in ungrounded captions. Additionally, annotating long videos—often exceeding 1200 frames at low frame rates (e.g., $fps = 1$)—is costly, as each frame must be processed individually. To address these challenges, we leverage subtitles, obtained during the pipeline process, as natural and grounded video captions to improve descriptive accuracy, as practiced in previous works as well [6, 55]. However, directly using raw subtitles is often inconsistent over long durations [26, 58]. Also, they may introduce distractions, such as irrelevant musical signals or disjointed character dialogues. To overcome these issues, we employ GPT-4o-mini to rephrase subtitles using video metadata and detailed instructions. Given the limitations of LLMs in sustaining long-context generation, we divide subtitles into manageable chunks, generate partial descriptions for each chunk, and merge them back into a cohesive description with a specified word count. Detailed prompts for this process are provided in 8. The resulting subtitles have an average of 2757 tokens, over 10x longer than previous open datasets as shown in Table 1.

3. Training Method

We present our training recipe in this section, which follows a two-stage workflow designed to enhance short-context window VLMs with long-video understanding capabilities.

Warming up with Pure Text Extension Infusing long-context understanding into short-context VLMs requires adapting the model's positional encodings to handle longer inputs. This adaptation can be achieved cost-effectively by using long-context textual data, as demonstrated by Zhang et al. [54]. For extending textual input, we apply the Slimpajama dataset [36] in line with Zhang et al. [54], and fine-tune the language model component of the VLM to extend the maximum sequence length to 224K, encapsulating 1 billion tokens. During training, we set RoPE's θ to $1e9$ to accommodate significantly long inputs [37]. This procedure results in significant gains in extending long video context window, as shown in the Needle-in-Haystack experiment (Figure 9).

However, overtraining on pure text data, while improving the context window, can introduce domain shifts across modalities. Notably, as shown in Figure 10 (in Appendix), our findings indicate that only a minimal amount of data and training is necessary for this phase; specifically, a checkpoint at 200 steps suffices for subsequent video training. Moreover, we observe that text-only extensions may lead models to produce overly lengthy and less fluent responses, as illustrated by examples in Figure 7. Therefore, this training step is most effective as a warm-up phase for models with shorter context windows. For models that already support long-context multimodal inputs, such as `LongVA`, this step can be skipped (see detailed discussion in §4.4). To further enhance multimodal long-context capabilities, we focus on additional cross-modality training to strengthen the grounding ability of VLMs.

Training with Video Descriptions Long video alignment training is essential to strengthen the long-context capabilities of video VLMs. In this step, we train on our curated high-quality description data from LONGVIDEOSET and perform description generation training based on long video inputs, following the widely adopted procedure used by Lin et al. [21]. With our long video data, we are able to extend the number of input frames to above 1200 with $fps = 1$ with 224K context. Bounded by the generating

length, we set the number of frames of both training and evaluation to 400 frames, resulting in nearly 80K length for `LLaVA-OneVision` with 196 tokens per frame, and 60K length for `LongVA` and `VISTA-LongVA` with 144 tokens per frame.

4. Experiments

In our experiments, we evaluate our extended length video training in two distinct settings: (i) **Long Video Retrieval**, aimed at retrieving detailed information (§4.2); and (ii) **Long Video Understanding**, focused on the semantic comprehension of extended video clips (§4.3). Discussion in ablation study are presented in §4.4, while case study are shown in Appendix C.

4.1. Experiment Settings

Benchmarks Our evaluation centers on assessing the ability of video VLMs to comprehend both short videos and extended inputs exceeding 20 minutes in length. Specifically, we tackle the notoriously challenging task of processing long-context inputs, as highlighted by Wu et al. [48]: Long Video Retrieval. Following Wei et al. [46], Zhang et al. [54], we rigorously stress-test long video VLMs using the Needle in a Haystack task, and further examine their robustness with more distractive "needles". Beyond retrieval, our evaluation also delves into long video understanding, which involves grasping the overarching semantics of extended video content. To this end, we leverage three widely used benchmarks: Video-MME [12], MLVU [60], and LVBench [41]. Video-MME evaluates a range of videos, from short clips (2 minutes) to long-form content (up to 60 minutes), while MLVU and LVBench emphasize the assessment of longer video comprehension.

Training Settings We adopt a continual instruction tuning approach to extend the maximum video input length of current video-VLMs. Our exploration involves training three different video-VLMs: `LLaVA-OneVision`[17], `LongVA`[54], and `VISTA-LongVA` [30]. To achieve this, we employ a combination of two distinct training strategies based on our LongVideoSet: Pure Text Extension (**PTE**) and Video Description Alignment (**VD**). PTE utilizes the long-text corpus Slimpajama to address the "cold start" problem in short-context video-VLMs, ensuring smooth adaptation to longer inputs. However, this step is unnecessary for models already adapted to handling extended input contexts. VD, on the other hand, is trained directly using LongVideoSet. Accordingly, we apply the PTE+VD training strategy for `LLaVA-OneVision`, while relying solely on VD training for `LongVA` and `VISTA-LongVA`. The trained models would be refered as `LVS-LLaVA-OneVision`, `LVS-LongVA` and `LVS-VISTA-LongVA`.

Evaluation and Baselines We evaluate our training pipeline and compare it to the pure text extension [54]. Additionally, we explore different combinations of training stages to better understand the optimization of video long-context extension training. We also compare against both proprietary and open-source SOTA long video VLMs.

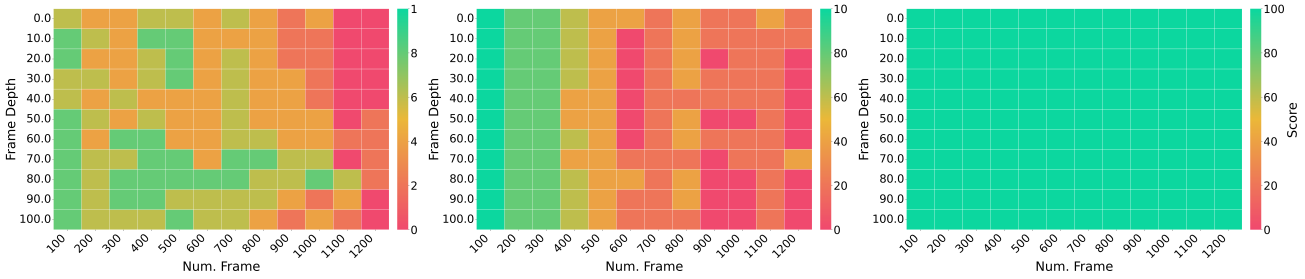
4.2. Long Video Retrieval

Needle in a Haystack The Needle in a Haystack experiment is a key paradigm for evaluating long-context retrieval, where models must locate information about inserted snippets within extended contexts. We benchmark this using V-NIAH [54], where needles are inserted across 0–1200 frames. Comparisons are made between the original `LLaVA-OneVision`, PTE training, and two-stage training (PTE + VD) using LongVideoSet. As shown in Figure 4, `LLaVA-OneVision` struggles to retrieve information beyond 200 frames, with performance deteriorating significantly after 800 frames. PTE training improves retrieval for shorter contexts and maintains consistent performance across varying frame depths. Adding VD training on our curated long video eliminates performance degradation for long contexts. Note that our reproduced results for PTE training is lower than originally reported [54], this is due to the likelihood-like evaluating format, which is more challenging as it penalizes answers that do not rigorously following the instructions. Additional details are provided in Appendix D.1. Further evaluation on the V-NIAH-D [46] benchmark, which includes distracting needles, demonstrates that our training method improves performance and consistently outperforms baselines across all frame lengths (Appendix Figure 12).

4.3. Long Video Understanding

Main Results Table 2 summarizes results from three long video understanding benchmarks. Training with LongVideoSet consistently improves performance across all benchmarks and base models. Notably, even for the state-of-the-art video VLM `LLaVA-OneVision`, our extended training yields improvements of 0.7% on Video-MME, 3.4% on MLVU, and 4.6% on LVBench, while surpassing `GPT-4o` by 4.2% on MLVU. Additionally, our approach enhances performance on longer videos (over 20 minutes) and reliably supports inputs exceeding 400 frames. These results highlight the effectiveness of our training method in improving long video understanding with denser frames.

In contrast, most open-source models (e.g., `LongVA` and `VISTA-LongVA`) show little to no improvement—or even decreased performance—with denser frames. This supports our hypothesis that simply adding more frames does not enhance performance unless the model has sufficient long-context comprehension. Without this capability, the added



(a) LLaVA-OneVision (b) PTE-200 (c) LVS-LLaVA-OneVision

Figure 4. Results of V-NIAH. From left to right, we present the results of LLaVA-OV-7B, which serves as the baseline, followed by the checkpoint after pure text extension, and finally, the checkpoint trained on LONGVIDEOSET from PTE-200.

Models	Size	# Frames	Video-MME†				MLVU§			LVBench	Overall
			Long	Medium	Short	Avg.	Long	Medium	Avg.	Avg.	
Proprietary Models											
GPT-4V [1]	-	10	53.5	55.8	70.5	59.9	-	-	49.2	-	-
GPT-4o [39]	-	384	65.3	70.3	80	71.9	-	-	64.6	34.7	57.1
Gemini 1.5 Flash [38]	-	fps=1	61.1	68.8	78.8	70.3	-	-	-	-	-
Gemini 1.5 Pro [38]	-	fps=1	67.4	74.3	81.7	75.0	-	-	-	33.1	-
Open-source Models											
LongVILA [9]	7B	256	53.0	58.3	69.0	60.1	-	-	-	-	-
LongLLaVA [43]	9B	256	45.2	47.3	60.9	51.1	-	-	-	-	-
Video-XL [34]	7B	256	49.2	53.2	64.0	55.5	-	-	64.9	-	-
LLaVA-Video [56]	7B	64	-	-	-	63.3	-	-	58.6	-	-
LongVU [32]	7B	fps=1	59.5	-	-	60.5	-	-	65.4	-	-
VideoChat-Flash [19]	7B	-	55.4	-	-	65.3	-	-	74.6	48.2	62.0
LongVA	7B	64	45.0	50.9	61.4	52.4	55.8	57.8	58.7	35.9	49.0
		400	46.2	52.8	61.2	53.4	56.7	58.4	59.0	35.1	49.2
PTE	7B	400	47.1	53.2	61.6	54.0	54.8	59.7	60.4	36.2	50.2
LVS-LongVA (Ours)	7B	400	51.7	56.9	63.9	57.5	55.8	61.5	62.0	38.7	52.7
Δ			**+6.7**	**+6.0**	**+2.5**	**+5.1**	**+0.0**	**+3.7**	**+3.3**	**+2.8**	**+3.7**
VISTA-LongVA	7B	64	46.0	53.1	65.6	54.9	56.7	60.9	61.8	39.0	51.9
		400	45.4	53.4	61.8	53.6	56.7	61.0	61.8	36.9	50.8
PTE	7B	400	48.3	55.1	63.8	55.7	61.5	62.7	63.6	36.7	52.0
LVS-VISTA-LongVA (Ours)	7B	400	52.7	58.2	66.0	59.0	58.7	63.1	63.6	40.9	54.5
Δ			**+6.7**	**+5.1**	**+0.4**	**+4.1**	**+2.0**	**+2.1**	**+1.8**	**+1.9**	**+2.6**
LLaVA-OneVision	7B	32	48.9	56.4	70.1	58.5	51.0	64.5	65.3	39.4	54.4
PTE	7B	400	50.8	58.1	70.2	59.7	52.9	65.1	66.5	38.9	55.1
LVS-LLaVA-OneVision (Ours)	7B	400	54.9	62.2	70.8	62.6	53.8	68.4	68.8	44.0	58.5
Δ			**+6.0**	**+5.8**	**+0.7**	**+4.1**	**+2.9**	**+3.9**	**+3.4**	**+4.6**	**+4.1**

Table 2. Our results on three open-source models—LLaVA-OneVision [18], LongVA [54], and VISTA-LongVA [30]—are reported in 10^{-2} accuracy. Results are shown after each training stage. †For Video-MME, we evaluate video understanding without subtitles, using only video input. We report performance on short (0-2 min), medium (4-15 min), and long (over 30 min) videos. §For MLVU, Long (over 20 min) and medium (over 5 min) video understanding are also evaluated.

frames become a burden. Moreover, merely adapting models to longer input sequences using long text is insufficient for understanding long videos. Our findings emphasize the necessity of long video description data to effectively utilize more frames for better long video understanding.

Global Context Understanding As highlighted in Figure 1, global context understanding presents a significant challenge. Here, we further analyze the impact of our training method on improving this aspect. Following the methodology of Zhou et al. [60], we divide the task into three test splits: (1) problems requiring holistic long video

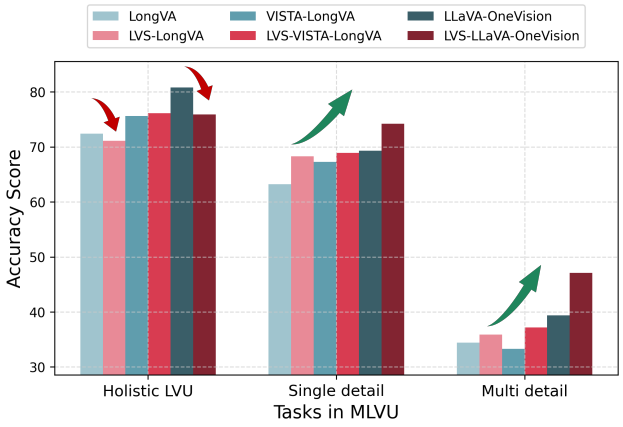


Figure 5. Model performance by detail granularity: holistic questions need general understanding, single-detail questions focus on one detail, and multi-detail questions require combining multiple details from the context.

Model	Length	Video-MME	MLVU	LVBench	Avg.
LLaVA-OneVision	32K	65.3	58.5	39.4	54.4
+ VD	224K	65.7	61.5	38.7	55.3
+ PTE & VD	224K	**68.8**	**62.6**	**44.0**	**58.5**
VISTA-LongVA	224K	54.9	61.8	39.0	51.9
+ VD	224K	**59.7**	**66.5**	38.9	**55.1**
+ PTE & VD	224K	57.9	60.2	**39.5**	52.5

Table 3. Comparison of pure text extension training warm-up on shorter vs. longer context video VLM.

understanding, (2) those that only need understanding of a single detail, and (3) problems requiring the identification and aggregation of multiple details from the video to provide an answer.

As illustrated in Figure 5, across three different base models, our training method demonstrates the most significant improvement in tasks that require aggregating information from multiple details. This improvement stems from the information-rich and extended length descriptions of long videos in LongVideoSet, which align closely with the unique challenges of long video understanding compared to shorter videos. However, our performance shows a slight decline in holistic long video understanding. This is expected, as holistic tasks are often easier with fewer frames and demand stronger modeling capabilities when working with a larger number of frames (increasing from 64 to 400).

4.4. Ablation Study

When is Pure Text Extension warm up necessary? As discussed in §3, pure text extension can adapt shorter-length video VLMs to longer position encodings. In this section, we ablate the effectiveness of pure text extension for shorter vs. longer context VLMs. As shown in Table 3, for the shorter context VLM, `LLaVA-OneVision`, video description training can extend context length and improve performance, but the improvement is insufficient to surpass pure text extension training. In comparison, `VISTA-LongVA` already possesses the ability to process long inputs, and further using PTE warm-up results in decreased performance.

Models	Avg	MLVU	Video-MME	LVBench	LVB
LLaVA-OneVision	54.9	65.3	58.5	39.4	**56.3**
PTE	54.1	66.5	59.7	38.9	51.2
LongVideoSet -128	55.3	67.5	61.6	39.2	52.9
LongVideoSet -256	56.3	68.3	**62.1**	**43.1**	51.6
LongVideoSet -400	**56.5**	**68.6**	62.0	42.3	53.0

Table 4. Performance when trained and tested on a subset of LongVideoSet with different number of frames. LVB is short for LongVideoBench.

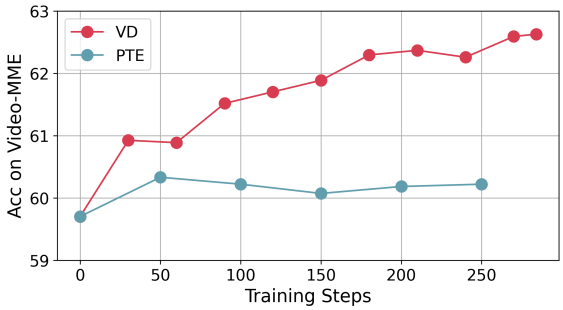


Figure 6. Performance of LLaVA-OneVision in progressive training steps on LongVideoSet during 1 epoch, compared with PTE.

Performance w.r.t Video Frame Density We analyze how the number of video frames used during training impacts performance. Downsampling video frames can reduce detailed information but may simplify learning global patterns. We compare the performance of `LLaVA-OneVision` when trained on 128, 256, and 400 video frames. As shown in Table 4, the model trained on 400 frames achieves the best overall performance. Training with 256 frames results in a slight performance drop but remains much better than training on 128 frames or without extended frames. Notably, the 256-frame model performs comparably to the 400-frame model on several long video benchmarks (e.g., LVBench, VideoMME), highlighting the trade-off between learning difficulty and the richness of information. Similarly, testing performance starts to saturate beyond 400 frames, as shown in Figure 12.

Scaling Behavior of LongVideoSet Training We validate the scaling properties of LongVideoSet video description training. As shown in Figure 6, the performance of `LLaVA-OneVision` on Video-MME consistently improves with the number of training steps and does not satu-

Category	Video-MME			
	AVG.	Long	Medium	Short
LLaVA-OneVision	58.48	48.90	56.40	70.10
+ vlogs	59.63	50.40	58.30	70.10
+ event records	59.59	50.70	58.20	69.90
+ documentary	59.41	50.20	58.30	69.70
+ cartoon	59.74	51.70	58.20	69.30
+ sports	59.30	49.70	58.70	69.60
+ TV shows	**60.15**	**52.00**	58.30	70.10
Combined	**60.15**	50.80	**59.00**	**70.70**

Table 5. Comparison of different video categories for video training. Training done on a uniformly sampled subset.

Category	Video-MME			
	AVG.	Long	Medium	Short
LLaVA-OneVision	58.48	48.90	56.40	70.10
+ video subtitles	60.74	51.60	60.20	**70.40**
+ video description	**62.00**	**53.70**	**62.10**	70.20

Table 6. Comparison of vanilla and rephrased detailed descriptions, conducted on a subset of LONGVIDEOSET.

rate even after 1 full epoch. This demonstrates the scalability of our training method. Notably, our data collection process is highly efficient and can be scaled further; however, the current scale is limited by computational resources. The scaling curve suggests that our proposed long video extension training has potential for further scaling up.

Comparison of Data Domains We conduct an ablation study across different video categories to examine their impact on training results. Specifically, we sample an equal number of videos from each category, as well as a uniformly combined subset of videos from all categories as a reference. The trained checkpoints are then evaluated on the Video-MME benchmark [12].

As shown in Table 5, the "TV shows" category has the most significant improvement, particularly for long videos. This category primarily includes reality shows and TV news reports, which feature abundant actions and frequent scene changes. Notably, TV news reports often present an initial preview of the entire content and sequential event ordering, enhancing the model's ability to handle long-distance dependencies. On the other hand, the combined subset of uniformly sampled video categories performs best overall, particularly for questions about medium and short-length videos. The diversity of video types in this subset helps the model adapt better to in-domain shifts, aligning with the data distribution of real-world scenarios.

Detailed Captioning improves Annotation Quality In this paragraph, we compare training video alignment using simple subtitles versus using rephrased detailed captions generated by `GPT-4o`. As shown in Table 6, training with simple subtitles provides some improvement, but rephrasing them into detailed descriptions yields significantly better results. This demonstrates that video alignment benefits from richer, more descriptive captions, indicating that text caption quality is crucial to the performance of long-context video training.

5. Related Work

Long Video Understanding Models LLMs are a strong foundation for bridging modalities [10, 18, 21, 22, 40, 63]. A major challenge for vision-language models is effectively processing long video inputs, as most are limited to fewer than 64 frames and require aggressive down-sampling for extended videos [10, 21, 24, 40, 51, 57]. However, both prior research [47] and our findings emphasize the importance of maintaining frame density for accurate video understanding. This has motivated recent efforts to extend the maximum input frames of vision-language models [2, 19, 20, 33, 50], often by adapting long-context LLM techniques, such as ring attention [23] or structure-preserving text-based methods [54]. Additionally, there is growing interest in leveraging real-world long videos to push these advancements further [2, 19, 50].

Long Video Instruction Tuning Datasets Video understanding requires specialized instruction-tuning data to enable temporal grounding and alignment [2, 10]. Most existing datasets consist of short videos, averaging under one minute, due to the scarcity of high-quality, long videos [6, 55, 57], and often retain only key frames for training [6]. While prior work has created synthetic long-video datasets by merging shorter clips [23, 30], our work is the first to curate a dataset of real-world, hour-long videos, offering a comprehensive foundation for training video understanding models. Traditional human annotation [26, 49] is difficult to scale, so recent efforts increasingly use multimodal LLMs like GPT-4V [6, 55, 57].

6. Conclusion

In this paper, we propose an efficient method to extend the context window of VLMs and enable it to understand hour-long videos. Our approach utilizes subtitle supervision to provide richer contextual information. To support this, we introduce LONGVIDEOSET, a long video dataset with video duration spanning from 20 minutes to 60 minutes, enriched with fine-grained annotations derived from subtitle supervision. Additionally, we explore multi-stage training strategies to extend the context window of VLMs. Our experiments show that training on LONGVIDEOSETsignificantly enhances the performance of various strong VLMs on long-video comprehension.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 8

[3] 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. 4

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 4

[5] 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*, pages 961–970, 2015. 4

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 4, 8

[7] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13320–13331, 2024. 4

[8] 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*, 2024. 4

[9] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. 2024. 3, 6

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1, 8

[11] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024. 4

[12] Chaoyou Fu, Yuhan Dai, Yongdong 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. 1, 2, 5, 8

[13] 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*, pages 5842–5850, 2017. 4

[14] 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*, pages 18995–19012, 2022. 4

[15] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024. 4

[16] 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. 4

[17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 5, 12

[18] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 6, 8

[19] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 6, 8

[20] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 8

[21] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 4, 8

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 8

[23] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with

blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 8

[24] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 8

[25] 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. 4

[26] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 4, 8

[27] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 4

[28] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 3

[29] 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*, pages 8748–8763. PmLR, 2021. 4

[30] Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. *arXiv preprint arXiv:2412.00927*, 2024. 4, 5, 6, 8

[31] 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*, pages 279–287. ACM, 2019. 4

[32] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 6

[33] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 8

[34] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 6

[35] 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*, pages 510–526. Springer, 2016. 4

[36] Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023. 4

[37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[38] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. 2, 6

[39] OpenAI Team. Gpt-4o system card, 2024. 2, 6

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[41] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 5

[42] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation, 2024. 4

[43] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture, 2024. 6

[44] 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. 4

[45] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024. 4

[46] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025. 5

[47] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2025. 1, 8

[48] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*, 2024. 5

[49] 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*, pages 5288–5296, 2016. 8

[50] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1, 8

[51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 8

[52] 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. 4

[53] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 4

[54] 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*, 2024. 1, 2, 3, 4, 5, 6, 8, 12

[55] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 4, 8

[56] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 6

[57] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 4, 8

[58] Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, et al. Distilling vision-language models on millions of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13116, 2024. 4

[59] 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. 2

[60] 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. 5, 6

[61] Luowei Zhou and Jason J. Corso. Youcookii dataset. 2017. 4

[62] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[63] 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*, 2023. 8