ProLongVid: A Simple but Strong Baseline for Long-context Video Instruction Tuning

Rui Wang[1,2], Bohao Li[4], Xiyang Dai[3], Jianwei Yang[3], Yi-Ling Chen[3], Zhen Xing[1,2], Yifan Yang[3], Dongdong Chen[3], Xipeng Qiu[1,2], Zuxuan Wu[1,2] †, Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University,
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing,
[3]Microsoft, [4]CUHK (SZ)
https://github.com/ruiwang2021/ProLongVid

Abstract

Video understanding is essential for multimodal large language models (MLLMs) to interact effectively with users and the real world. However, analyzing long videos remains a major challenge due to the lack of high-quality video instruction data and effective training strategies. In this paper, we introduce a simple yet effective baseline for long-context video understanding, including dataset construction and training recipes. We curate a large-scale video instruction dataset with over 1M samples, encompassing videos from a few seconds to several minutes across diverse sources, without any human annotations. Additionally, we propose a progressive video instruction tuning strategy that incrementally increases input context length, enabling better utilization of videos of varying durations. Comprehensive experiments demonstrate that our dataset significantly outperforms existing video instruction datasets for fine-tuning MLLMs. Furthermore, our training approach establishes a strong video MLLM baseline, surpassing previous open-source models on video benchmarks and outperforming proprietary models like GPT-4V and GPT-4o-mini on VideoMME, even with a compact 7B model.

1 Introduction

Multimodal Large Language Models (MLLMs) have made significant strides in understanding image content and generating meaningful responses based on complex instructions (Liu et al., 2023b; OpenAI, 2023a). However, to effectively interact with users in real-world scenarios, handle real-time production environments, and process vast amounts of internet-sourced data, these models must develop strong capabilities for comprehending visual data within extended contexts (Liu et al., 2023b). Among various types of visual data, video stands out as a crucial source in both online and physical domains due to its inherently long-context nature.

Previous efforts in video-centric MLLMs have explored multiple strategies to adapt image-focused models for video understanding (Jin et al., 2024). Some approaches extend single-image features to multi-frame representations to handle short video clips (Lin et al., 2023), while others focus on feature compression techniques to reduce the number of video tokens required for processing long video sequences (Li et al., 2023b, 2025).

While these studies have laid a strong foundation for integrating video understanding into MLLMs, recent findings (Kim et al., 2024; Zhang et al., 2024c) suggest that treating video frames as image grids or individual frames can outperform earlier video-specific methods in zero-shot video understanding benchmarks. This raises questions about the effectiveness of previous approaches, particularly in the context of video instruction tuning. One major limitation contributing to this gap is the lack of high-quality video instruction datasets. Compared to text and image datasets, video data is information-dense, making dataset construction significantly more expensive as video duration increases. As a result, publicly available video instruction datasets, especially for long-duration videos, remain scarce. This lack of data has, in turn, limited the exploration of scalable training strategies for long-video understanding. To build robust MLLMs with strong long-context video understanding, it is essential to address both dataset construction and the design of efficient training pipelines that can fully utilize long-video data.

In this work, we introduce ProLongVid, a comprehensive framework for training video-centric MLLMs with long-context understanding, encompassing both dataset construction and an optimized training pipeline. To ensure diverse video coverage, our dataset includes videos ranging from a few

† Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30836–30849
November 4-9, 2025

seconds to 20 minutes across diverse sources.

We propose a scalable three-stage automated video instruction annotation pipeline that effectively handles videos of varying lengths. First, we design a novel temporal video segmentation algorithm that clusters frames based on both visual and semantic similarities. Then, we leverage GPT-4o to generate detailed segment-level and video-level descriptions, followed by diverse QA pairs through a multi-granularity approach that captures both local details and global temporal relationships.

Additionally, we introduce a **progressive instruction tuning** to better utilize video data of varying durations. In this strategy, video data is grouped by length and introduced progressively in multiple training stages, with the number of input frames gradually increasing over time. This curriculum-based approach enables MLLMs to adapt incrementally to longer video contexts. To preserve the original visual perception capabilities of MLLMs, we freeze the visual encoder parameters during training. Moreover, we find that extending the LLM's context length before vision-language training further enhances long-video understanding.

Leveraging our instruction dataset and training pipeline, we extend a strong MLLM trained on image-only instruction data into a model capable of understanding long-context videos. Our ProLongVid-7B achieves state-of-the-art performance on multiple video understanding benchmarks, surpassing previous open-source models. Notably, despite its smaller model size, it outperforms proprietary models like GPT-4V and GPT-4o-mini on VideoMME, demonstrating the effectiveness of our approach.

In summary, we introduce **ProLongVid**, a comprehensive framework for training video-centric MLLMs with long-context understanding, integrating both large-scale dataset construction and an efficient training pipeline. The key characteristics of ProLongVid are as follows:

- **Large-Scale Video Instruction Dataset:** ProLongVid includes over one million video-instruction samples, covering a diverse range of topics and durations (from a few seconds to 20 minutes).

- **Automated and Scalable Annotation:** We employ a three-stage annotation framework that combines open-source models for temporal video segmentation and proprietary APIs for refined caption/instruction generation, ensuring cost-efficient, high-quality annotations without human supervision.

- **Progressive Instruction Tuning:** Our curriculum-based training pipeline progressively increases input length across multiple stages, enabling the model to adapt effectively to varying video durations.

- **State-of-the-Art Performance:** ProLongVid establishes a strong video MLLM baseline, achieving superior performance across multiple benchmarks and outperforming previous open-source models, as well as proprietary GPT-4V and GPT-4o-mini on VideoMME.

2 Related Works

MLLMs for Video Understanding. MLLMs have made significant progress in image-text understanding. However, video understanding remains a more challenging task due to constraints in data availability and computational resources. Early video MLLMs primarily relied on either multi-frame features projected through an MLP (Lin et al., 2023; Luo et al., 2023; Ataallah et al., 2024; Xu et al., 2024a) or video features resampled via Q-former architectures (Li et al., 2023a,b, 2024b). Recent training-free approaches (Kim et al., 2024; Zhang et al., 2024c; Xu et al., 2024b) have shown that strong image MLLMs can achieve competitive video understanding performance without explicit video instruction tuning, even surpassing some early video MLLMs. However, most of these methods can only process a limited number of frames (e.g., fewer than 32) through sparse sampling, fundamentally constraining their effectiveness on long video understanding. To address this limitation, several works have explored video token compression techniques, such as memory mechanisms (Song et al., 2024) and token merging (Shen et al., 2024), allowing models to process more frames within constrained context windows. Our work builds upon these efforts by extending the effective visual context length through synthetic long-video instruction data while maintaining a simple and scalable video MLLM architecture.

Datasets for Video MLLMs. Due to the scarcity of human-annotated data, recent multimodal instruction datasets have been synthesized using strong MLLMs. In the image domain, previous works (Chen et al., 2023a, 2024a) have constructed

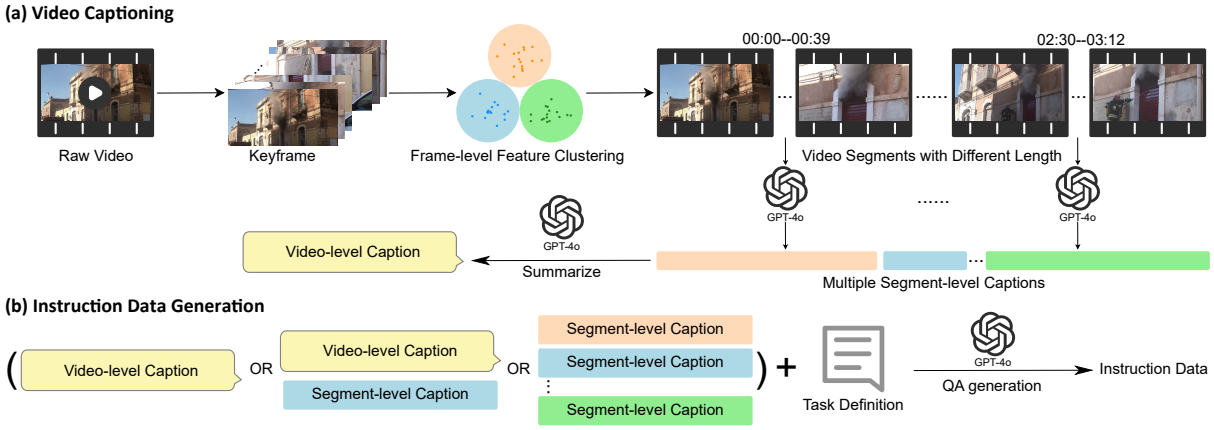


Figure 1: An overview of data generation pipeline in ProLongVid.

high-quality instruction datasets by leveraging the robust image comprehension capabilities of proprietary models such as GPT-4V (OpenAI, 2023b). However, in the video domain, high-quality video instruction data remains limited due to the high cost of video annotation and the inherent difficulty of generating accurate video descriptions. Early video MLLMs, trained on small-scale, low-quality data, have demonstrated limited performance. While some approaches rely on human-annotated video QA data (Yu et al., 2019; Xiao et al., 2021; Patraucean et al., 2023; Mangalam et al., 2024), these methods face scalability challenges. More recent works have explored automatic annotation pipelines using proprietary APIs. LLaVA-Hound (Zhang et al., 2024b) uses GPT-4V to generate video captions from sparsely sampled frames, while ShareGPT4Video (Chen et al., 2024b) improves video captioning by analyzing inter-frame differences. However, these approaches are mostly limited to relatively short videos. To advance long-form video understanding, recent benchmarks (Fu et al., 2024a; Zhou et al., 2024; Wu et al., 2024) have emerged, focusing on QA tasks for videos ranging from several minutes to over an hour in duration. Our work aligns with this direction but offers notable advantages in dataset scale, annotation quality, and video duration coverage in instruction-tuning scenarios.

3 Dataset Construction

Our instruction data generation pipeline consists of three stages: **semantic video segmentation**, **dense video captioning**, and **instruction generation**. We detail each stage in the following sections.

3.1 Data Collection and Processing

To construct a high-quality video instruction dataset, we gather videos of varying lengths, domains, and topics from multiple sources. Unlike some previous instruction datasets, our dataset does not rely on existing video QA training sets, reducing the risk of information leakage and preventing overfitting to the domains of existing benchmarks. Our sources include short videos from PMV (Han et al., 2024), SA-V (proposed in SAM-2 (Ravi et al., 2024)), and VIDAL (Zhu et al., 2023), which originate from short-video platforms, as well as long-form videos from the YouTube-8M (Abu-El-Haija et al., 2016) dataset. To demonstrate the scalability of our pipeline for web videos, we avoid using any pre-existing annotations from these datasets. Instead, we directly process untrimmed videos through our annotation pipeline. Unlike prior video instruction datasets that primarily focus on short videos, we ensure diversity in video lengths by sampling across different duration ranges, enabling research on long video understanding.

Following ShareGPT4Video, our video annotation pipeline aims to automatically generate detailed video captions while we leverage the stronger model, i.e., GPT-4o, as the video captioner. To ensure comprehensive coverage of video content, we sample frames at 1 frame per second (fps). This dense sampling rate makes it impractical to use most existing MLLMs, including GPT-4o, directly. To address this challenge, we introduce a simple yet effective video segmentation algorithm that subdivides long videos into manageable chunks, allowing us to fully utilize GPT-4o's capabilities.

3.2 Semantic Video Segmentation

The first step of our annotation pipeline involves partitioning videos into semantically coherent seg-

ments of manageable size. This not only alleviates the computational burden of subsequent video captioning but also *simplifies* the task, as each segment contains less information. By reducing the likelihood of missing fine-grained details or introducing hallucinations, our segmentation approach enhances the reliability of video captioning models. Additionally, it enables the use of smaller, specialized captioners, improving scalability.

Given N sampled video frames, our segmentation algorithm considers both *visual similarity* and *caption similarity* to group frames into distinct semantic clusters. Specifically, set the sampled frames be $I_1, I_2, \ldots, I_N$. We employ the strong vision-language model Florence-2 (Xiao et al., 2024) to generate frame-level captions, denoted as $C_1, C_2, \ldots, C_N$. This allows us to extract both frame-level image features $f_1^I, f_2^I, \ldots, f_N^I$ and text features $f_1^C, f_2^C, \ldots, f_N^C$ using the Florence (Yuan et al., 2021) model. We then define the pairwise similarity between frames I_i and I_j as follows:

$$s_{i,j} = \frac{f_i^I \cdot f_j^I + f_i^I \cdot f_j^C}{2} + (2 - e^{||\frac{i-j}{N}||}),$$

where i and j denote the frame indices. In this formulation, the first term combines *image-image* and *image-text* similarity, encouraging visually and semantically similar frames to be clustered together. The second term penalizes frames that are temporally distant, ensuring segmentation remains consistent with video structure. Using this similarity metric, we apply a standard hierarchical clustering algorithm, specifically the Agglomerative Clustering implementation from `scikit-learn`. Empirically, we found that *text-text* similarity alone does not produce satisfactory clustering results and therefore exclude it from the total similarity. While we use Florence-2 for captioning, other models, such as BLIP-2 (Li et al., 2023a) or CLIP-like models (Radford et al., 2021), can also be integrated into our segmentation framework.

A key advantage of our approach is its controllable granularity. By adjusting the clustering distance threshold δ, we can fine-tune the segment size or even deliberately *over-segment* videos for specific applications. Based on empirical evaluation, we set $\delta = 0.7$, achieving a balance between segment granularity and semantic coherence.

3.3 Dense Video Captioning

As described earlier, we primarily leverage GPT-4o to automate dense video captioning. Our video segmentation method partitions long videos into manageable chunks, allowing GPT-4o to process them efficiently. As illustrated in Figure 1, we employ GPT-4o to generate detailed segment-level descriptions. These descriptions, along with their corresponding timestamps, are then sequentially fed back into GPT-4o to produce an aggregated video-level dense caption. This process enables GPT-4o to capture contextual relationships between segments, ensuring a cohesive and comprehensive description of the entire video.

For short videos (i.e., those under one minute in duration), segmentation is unnecessary. Instead, we directly use GPT-4o to generate a detailed caption for the entire video in a single pass.

While we primarily use GPT-4o for dense video captioning, our approach is not exclusively dependent on proprietary models. The segmentation algorithm ensures that the maximum segment length remains within the context window limitations of open-source MLLMs. This design enables the use of open-source alternatives such as Qwen2-VL (Wang et al., 2024a) for video captioning.

3.4 Instruction Generation

Building on the video descriptions generated in the previous stage, we use GPT-4o to create diverse video question-answering (QA) data. The prompts for QA generation include not only the video descriptions but also task definitions and example instructions tailored for video understanding. We define four primary QA task types: (1) Video Summarization: Generating concise summaries of the video content. (2) General Video QA: Answering questions about objects, attributes, actions, trajectories, and reasoning based on video details. (3) Creative Writing: Producing imaginative content inspired by the video. (4) Temporal Understanding: Analyzing the sequence of events within the video.

For short videos, we generate QA pairs based on the global video description, incorporating prompts aligned with the predefined task types.

For long videos, we generate both local descriptions for individual segments and a global description for the entire video during the captioning stage. To enhance the ability of MLLMs to understand temporal relationships and long-form video content, we employ three distinct input strategies when generating QA data:

- Global Video Description: Used to generate QA pairs for the first three predefined task

types (summarization, general QA, and creative writing).

- Local Descriptions of Multiple Segments: Used to create QA samples that require understanding temporal relationships and reasoning across segments.

- Local Description of a Single Random Segment: Used to generate general QA focused on local details, with the global description providing background context.

This multi-granularity approach ensures that our instruction data for long videos covers diverse QA tasks across different temporal scales, improving the model's ability to handle both local and global video understanding.

4 Training Recipe

4.1 Base Model

In this work, we adopt a LLaVA-style architecture (Liu et al., 2023b) as the foundation for building a video-centric MLLM. The architecture consists of three main components: a Vision Encoder, an MLP Projector, and an LLM. Specifically, we utilize SigLIP (Zhai et al., 2023) as the Vision Encoder, which processes raw images (384×384) into visual feature maps of size (27×27). These 2D grid features are then downsampled and reshaped into a 1D sequence. A two-layer MLP projector maps these visual features into the embedding space of the LLM. The resulting visual token sequence, combined with the word embeddings of the query text, serves as input to the LLM, which generates responses in an autoregressive manner.

During SFT stage, the visual data primarily comprises high-resolution images and video frames. For high-resolution images, we partition the input into a grid of $a \times b$ crops, preserving its original aspect ratio. Additionally, we create a global view by resizing the entire image to (384×384). Since SigLIP encodes each image into a (27×27) feature map, the total token count for a high-resolution image is $729 \times (1 + a \times b)$. For video inputs, we reduce the number of tokens per frame to accommodate more frames by representing each frame as a global view and downsampling its features to a 12×12 feature map. Thus, for a video with T frames, the total token count is $144 \times T$.

4.2 Training Pipeline

To effectively train our video-centric MLLM, we propose a structured three-phase training pipeline: (1) extending the base LLM's context length, (2) image-text alignment and instruction tuning, and (3) progressive video instruction tuning. This carefully designed strategy enables our model to process long video sequences while maintaining strong image and video understanding capabilities.

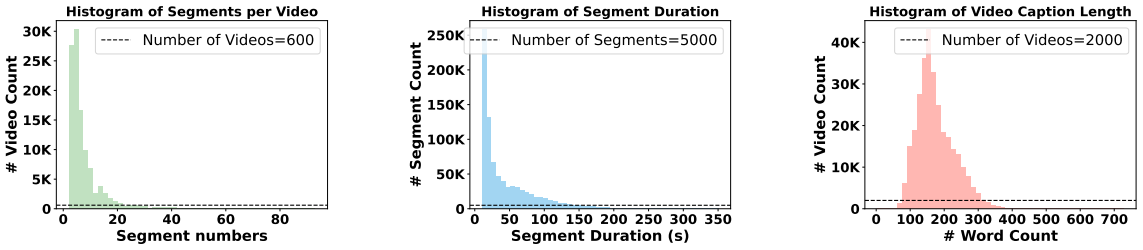
LLM context length extension. The foundation of our approach lies in extending the base LLM's long-context capability. Following (Fu et al., 2024b), we extend the LLM's context length to 256K by continuing pretraining on the SlimPajama (Soboleva et al., 2023) long-text dataset before multimodal training. Additionally, we scale the base frequency of Rotary Position Embeddings (RoPE) (Su et al., 2024) by a factor of 1,000 (from 1M to 1B) to enhance long-range attention. To efficiently support 256K context-length training, we adopt sequence parallelism (Li et al., 2021) based on ring attention (Liu et al., 2023a), following (Zhang et al., 2024a).

Image-text alignment and instruction-tuning. Building on the extended LLM, we establish strong image understanding through a three-stage training process, following (Li et al., 2024a): (1) Image-Text Alignment: We train only the projector, keeping the vision encoder and LLM frozen, using image-text paired data. (2) Knowledge-Driven Instruction Tuning: We finetune all model parameters using high-quality knowledge data. (3) Single-Image Instruction Tuning: We train the model with single-image instruction data to enhance its ability to follow instructions. These phase ensures a solid foundation for visual understanding before introducing video comprehension.

Progressive video instruction-tuning. To develop video understanding, we employ a progressive instruction-tuning strategy that fundamentally differs from previous long-context LLM training approaches (Chen et al., 2023b; Ding et al., 2023; Dubey et al., 2024; Gao et al., 2025). While existing methods focus solely on extending text context length, our approach introduces a dual-scaling paradigm that simultaneously increases both context length and video duration. This adaptation to visual sequences presents unique challenges not addressed by traditional long-context methods, as video understanding requires modeling both temporal dependencies across frames and spatial rela-

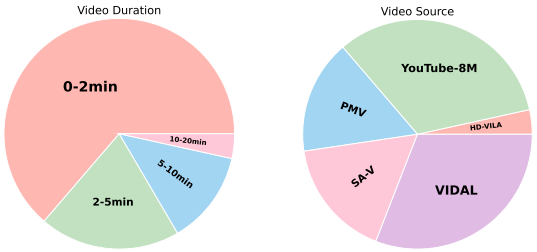
Dataset	Annotation	#Video	Total Video Length	Avg. FPS	#Caption	#QA
VideoInstruct (Maaz et al., 2024)	Human & GPT-3.5	13K	0.4K hr	-	13K	100K
LLaVA-Hound (Zhang et al., 2024b)	GPT-4V	900K	3K hr	0.008	900K	900K
ShareGPT4Video (Chen et al., 2024b)	GPT-4V	40K	0.2K hr	0.15	40K	0
LLaVA-Video (Zhang et al., 2024d)	GPT-4o	178K	2Khr	1	178K	1.2M
ProLongVid (Ours)	GPT-4o	300K	11K hr	1	300K	1.5M

Table 1: Comparison with previous video instruction datasets.



(a) (b) (c)

Figure 2: The statistics of our instruction datasets. We show the (a) histogram of segment numbers, (b) histogram of segment duration and (c) histogram of video caption length.



(a) (b)

Figure 3: The video duration and source of our datasets.

tionships within each frame.

Our progressive strategy gradually increases the model's temporal context length across three stages:

- Stage 1: Training with short videos ($<$ 2 minutes), sampling 32 frames as MLLM input.
- Stage 2: Incorporating medium-length videos (2-5 minutes), sampling 128 frames.
- Stage 3: Extending to long videos (5-20 minutes), sampling 192 frames.

This progressive approach scales the visual context length to 28K tokens while ensuring stable training dynamics. The extended 256K context length from Phase 1 enables temporal extrapolation during inference, allowing the model to process even longer video sequences than those seen during training.

5 Experiments

5.1 Dataset Statistics

Overview. We curate a diverse and well-balanced dataset by selecting data from multiple sources, comprising 300K videos and 1.5M question-answering examples. The dataset distribution is illustrated in Fig. 2 and Fig. 3. Fig. 2(a) visualizes how long videos are divided into smaller segments, with most videos split into 1 to 20 segments, though some extend up to 80 segments. Fig. 2(b) presents the duration distribution of these segments, which range from 0 to 370 seconds, with the majority being under 100 seconds, aligning with typical short video lengths. Fig. 2(c) shows the correlation between video length and caption length, where longer videos tend to have more detailed captions. Additionally, Fig. 3(a) illustrates the overall video duration distribution, with most videos falling within the 0-2 minute range, while a subset exceeds 5 minutes, enhancing the model's ability to process long videos. Fig. 3(b) depicts the dataset's source distribution, with the majority of videos sourced from YouTube-8M and VIDAL, ensuring a broad and diverse range of content categories.

Dataset Comparison. Tab. 1 compares our dataset with existing benchmarks, highlighting key distinctions. We employ GPT-4o for high-quality annotations, ensuring precise and contextually rich labels. Our dataset size of 300K videos is comparable to existing datasets, but we stand out with the highest FPS, enabling finer-grained visual content understanding. The total duration of our dataset is 11K hours, with 5-20min long videos accounting for 6.5Khr. This composition clearly positions our dataset as long-video focused, transcending the previous minute-level video data. Notably, the recent LLaVA-Video only annotates videos within

Dataset	Samples	Frame		VideoMME (w/o subtitles)			
		Train	Test	Short	Medium	Long	Overall
Image SFT	-	-	32	68.8	54.4	49.7	57.6
VideoInstruct	100K	32	32	69.2	57.6	50.0	58.9
LLaVA-Hound	255K	32	32	69.6	56.4	49.3	58.4
LLaVA-Hound	900K	32	32	71.3	58.4	50.6	60.1
LLaVA-Video	1.2M	128	128	74.1	61.0	50.8	62.0
ProLongVid-stage1	800K	32	32	72.8	59.1	48.4	60.1
ProLongVid-stage2	1.2M	128	128	75.0	**64.2**	51.4	63.6
			256	74.3	64.0	54.2	64.2
ProLongVid-stage3	1.5M	192	192	**75.4**	63.0	52.9	**63.8**
			256	75.2	64.0	**54.8**	**64.7**

Table 2: Comparison with previous video instruction datasets on VideoMME benchmark (without using subtitles). "Image SFT" is the start image MLLM for all datasets.

3 minutes. Moreover, our dataset includes 1.5M annotated question-answer pairs, significantly enriching the research community and driving further advancements in video understanding.

5.2 Implementation Details

Architecture and training. We use SigLIP (Zhai et al., 2023) as the vision encoder and Qwen2.5-Instruct (Yang et al., 2024) as the LLM in our experiments. After extending the LLM's context length, we train the entire MLLM using both image-text pairs and single-image instruction data. All image training data are sourced from the open subset of the LLaVA-OneVision (Li et al., 2024a) dataset, without incorporating multi-image data. After this stage, we freeze the vision encoder and perform progressive video instruction tuning to enhance video understanding.

Benchmark and evaluation. We evaluate our model on VideoMME (Fu et al., 2024a), MLVU (Zhou et al., 2024), LongVideoBench (Wu et al., 2024), and TempCompass (Liu et al., 2024b). While TempCompass is a short video benchmark, VideoMME, MLVU, and LongVideoBench focus on long videos. Videos in VideoMME are split into short, medium, and long videos based on duration, and we mainly use VideoMME for ablation study.

5.3 Main Experiments

Progressive training pipeline. We propose a three-stage progressive video instruction tuning approach based on a strong image MLLM baseline. As shown in Tab. 2, while the initial image MLLM already achieves strong zero-shot video understanding performance, our three-stage tuning consistently leads to significant improvements at each stage, particularly for long-video tasks after long-video training. Specifically, compared to the image MLLM baseline, our model achieves an overall improvement of 6.2% on VideoMME after three-stage training, with gains of 6.6%, 8.6%, and 3.2% on short, medium, and long video tasks, respectively. Furthermore, when increasing the number of training frames to 192, our model generalizes effectively to an even larger number of frames during inference. Notably, when using 256 frames at inference, the model's performance on VideoMME improves by 0.9% compared to 192 frames, with a particularly strong 1.9% improvement on long-video tasks.

Comparison with previous datasets. We compare widely used video instruction datasets generated through automated or semi-automated annotation processes. Specifically, for the LLaVA-Hound dataset, we experiment with two settings: one using the same combination of 240K QA samples and 15K captions as in (Zhang et al., 2024b), and another using the full set of 900K QA samples for training. In our experiments, we use the same image MLLM baseline (through the first two training phases) as the starting point for video instruction tuning. Results from video understanding benchmarks indicate that the earlier dataset, VideoInstruct, produces lower-quality data due to weaker visual information extraction techniques and the GPT API used at the time, leading to models with limited instruction-following capabilities. With 255K samples, LLaVA-Hound achieves similar results to VideoInstruct on VideoMME. However, scaling up the dataset size to 900K further

Model	MLVU(M-Avg)	LongVideoBench(val)	Tempcompass(mc)	VideoMME(wo/w sub)
Proprietary Models				
GPT-4V (OpenAI, 2023b)	49.2	59.1	-	59.9/63.3
GPT-4o-mini (OpenAI, 2024)	-	56.5	-	64.8/68.9
GPT-4o (OpenAI, 2024)	64.6	66.7	70.9	71.9/77.2
Open-sourced Models				
Video-LLaVA-7B (Lin et al., 2023)	47.3	39.1	45.6	39.9/41.6
VideoChat2-HD-7B (Li et al., 2024b)	47.9	-	-	45.3/55.7
LongVA-7B (Zhang et al., 2024a)	56.3	-	56.1	52.6/54.3
InternVL2-8B (Chen et al., 2024c)	64.0	54.6	65.6	56.3/59.3
LLaVA-Onevision-7B (Li et al., 2024a)	64.7	56.4	64.8	58.2/61.5
MiniCPM-V-2.6-8B (Yao et al., 2024)	-	54.9	63.0	60.9/63.7
Qwen2-VL-7B (Wang et al., 2024a)	-	55.6	**68.5**	63.3/69.0
ProLongVid-7B	**70.6**	**60.0**	66.3	**64.7/70.7**

Table 3: Comparison with state-of-the-art methods on video understanding benchmarks.

improves performance, highlighting the benefits of larger instruction datasets. In contrast, our model, trained on only a short-video subset of the ProLongVid dataset (videos under 2 minutes), outperforms both VideoInstruct and LLaVA-Hound-255K while achieving comparable performance to LLaVA-Hound-900K. Although LLaVA-Video also uses strong GPT-4o for annotation, our model surpasses it at stage 2 with a similar total amount of training data. Moreover, after progressive training on longer videos, our model not only achieves consistent and significant improvements over the image MLLM baseline but also substantially outperforms video MLLMs trained on previous datasets. These results demonstrate the effectiveness of our dataset and training approach.

Comparison with state-of-the-art. In Tab. 3, we compare our model with previous MLLMs on several video understanding benchmarks. On VideoMME, our model consistently outperforms prior open-source models. Notably, when incorporating video subtitles as input, our 7B model surpasses GPT-4o-mini by 1.8%. On MLVU, ProLongVid achieves an accuracy of 70.6%, significantly outperforming both proprietary models like GPT-4V and GPT-4o, as well as leading open-source models such as LLaVA-OneVision-7B. Similarly, on LongVideoBench, our model achieves 60.0% accuracy, surpassing GPT-4V, GPT-4o-mini, and other open-source competitors. These results highlight the significant impact of scaling up both video length in the instruction dataset and the visual context length during training and inference. We hope this work serves as a strong baseline for future research in video understanding.

Model	Short	Medium	Long	Overall
Image baseline	63.8	49.0	46.2	53.0
Multi-stage	**70.8**	**59.2**	**50.0**	**60.0**
Mixed data	69.7	57.9	48.8	58.8

Table 4: Comparison between progressive training and one-stage data-mixed training. We use Qwen2.5-3B as LLM here.

Extension	Frame	Short	Medium	Long	Overall
None	192	73.1	62.8	52.8	62.9
	256	72.6	62.2	52.2	62.3
Pretrain	192	**75.4**	63.0	52.9	63.8
	256	75.2	**64.0**	**54.8**	**64.7**

Table 5: Ablation on the extension of LLM context length. We start training on Qwen2.5-7B with or without the extension and evaluate MLLMs on VideoMME. "Frame" means the number of frames at inference.

5.4 Ablation Study

Multi-Stage vs. One-Stage Training. Most previous works have used a fixed number of frames for training video MLLMs. In contrast, we introduce a progressive training strategy, where the number of training frames gradually increases over three stages. Each stage follows a complete learning rate schedule, including a warm-up phase. We compare this progressive strategy to a data-mixed training approach, where all three subsets are combined into a single training stage. As shown in Tab. 4, our results demonstrate that the multi-stage progressive training strategy significantly outperforms the data-mixed method on video understanding benchmarks, highlighting the effectiveness of gradual context expansion during training.

Ablation on Context Length Extension. We conduct an ablation study to investigate the impact

of context length extension in LLM pretraining on video understanding capabilities. As shown in Tab. 5, our experiments reveal a key insight: extending the effective visual context length is a fundamental prerequisite for long-video understanding. Specifically, without increasing the LLM's context length, we observe performance degradation when scaling up visual tokens to 36K (256 frames) at inference. This performance gap persists both within and beyond the 32K context window, indicating that the model struggles to maintain coherent understanding over extended temporal spans. These findings empirically validate our hypothesis that extending the LLM's context capacity is essential for effectively processing long videos.

6 Conclusion

In this work, we present ProLongVid, an innovative framework for training video-centric MLLMs with long-context capabilities. Our framework encompasses both dataset construction and a progressive training pipeline, leveraging diverse videos with varying durations sourced from multiple platforms. To generate high-quality training data, we introduce an automated video instruction annotation framework that employs a two-stage process to segment videos and generate video captions and instruction data. Our training pipeline progressively incorporates video data of different lengths across multiple stages, effectively enhancing the ability of existing MLLMs to understand videos across various temporal scales. Finally, our 7B model achieves state-of-the-art performance across multiple video benchmarks, surpassing previous open-source models. We hope this work serves as a simple yet effective baseline for future research on long-context video understanding.

Limitations

Despite the promising results, our work has several limitations. First, due to computational resource constraints and limitations of current training frameworks, it is difficult to further scale up the maximum number of input frames during training (e.g., to 1K frames) without significantly compressing the number of visual tokens per frame. This restriction may limit the model's ability to process extremely long videos. We plan to address this challenge in future work by exploring more efficient training strategies and extending our training context length. Second, since we only utilize open-source data for multimodal image-language training, our base image MLLM's capabilities may be inferior to some proprietary models that have access to larger and more diverse training datasets. This limitation could potentially impact the overall performance of our video MLLMs.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant 62472098). This work was supported by the Science and Technology Commission of Shanghai Municipality (No. 24511103100).

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Anthropic. 2024. Claude-3.5. https://www.anthropic.com/news/claude-3-5-sonnet.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*.

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

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, and 1 others. 2024b. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024b. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. How to train long-context language models (effectively). In *ACL*.

Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, and Heng Wang. 2024. Video recognition in portrait mode. In *CVPR*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. 2024. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*.

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

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

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*.

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. 2021. Sequence parallelism: Long sequence training from system perspective. *arXiv preprint arXiv:2105.13120*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2025. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*.

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

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2025. St-llm: Large language models are effective temporal learners. In *ECCV*.

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

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2024. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*.

OpenAI. 2023a. Gpt-4: Openai's most advanced language model. https://openai.com/research/gpt-4. Accessed: 2024-11-15.

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

OpenAI. 2024. Gpt-4o: Openai's optimized language model for multimodal tasks. https://openai.com/research/gpt-4o. Accessed: 2024-11-15.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, and 1 others. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Yi-Fan Zhang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiawu Zheng, and 1 others. 2025. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuracy. *arXiv preprint arXiv:2502.05177*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. 2024b. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*.

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

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

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, and 4 others. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

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

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. 2024b. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*.

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

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

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

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

A The License and Intended Use For Artifacts

Here we list the license and intended use for the artifacts (datasets and benchmarks) in this paper.

- **PMV** is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License. The data is released for non-commercial research purposes only.
- **SA-V** is intended to be used for computer vision research for the purposes permitted under the CC by 4.0 license.
- **VIDAL** is released under the CC-BY-NC 4.0 license.
- **Youtube-8M** is released under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- **HD-VILA** is released under a Research Use of Data Agreement v1.0.
- **VideoMME** is only used for academic research.
- **MLVU** is under the CC-BY-NC-SA-4.0 license.
- **TempCompass** is under CC BY-NC 4.0 License and is intended for academic research only.
- **LongVideoBench** follows CC-BY-NC-SA 4.0 license and is intended for non-commercial use only.

Our proposed ProLongVid dataset is constructed based on open-source video data, and is intended for non-commercial research purposes only.

B Analysis of Computational Efficiency

Our three-stage progressive training approach demonstrates superior computational efficiency compared to single-stage training when using the same training data. We utilize 32, 128, and 192 frames as model inputs for videos of 0-2, 2-5, and 5-20 minutes, respectively. Our analysis reveals two key advantages of the three-stage method:

Elimination of GPU synchronization overhead: In single-stage training, batches frequently contain samples with varying input lengths, leading to significant computational time differences across GPU processes. Therefore, GPU processes handling short videos must wait for those processing long videos to complete before synchronizing gradients. This synchronization bottleneck introduces computational overhead that is avoided in our decoupled three-stage training method.

Training Method	Stage 1	Stage 2	Stage 3	Overall
Single-stage	-	-	-	3598h
Three-stage (Ours)	272h	661h	767h	1700h

Table 6: GPU hour comparison between single-stage training and three-stage progressive training.

Optimized memory and training configuration: The three-stage method enables stage-specific optimization of training configuration. For the first stage with short videos, we can utilize faster training settings with higher memory consumption (e.g., DeepSpeed Zero-1) due to lower per-sample token counts and memory requirements. For subsequent stages, we use memory-efficient settings (e.g., DeepSpeed Zero-3) to prevent out-of-memory errors. In contrast, single-stage training must adopt the most conservative configuration (Zero-3) based on the maximum input length, resulting in suboptimal overall training efficiency.

We conduct GPU hour comparison between three-stage and single-stage training. All experiments are conducted on 32 MI300X GPUs. Our results show that three-stage progressive training achieves shorter training times while maintaining model performance, as demonstrated in Table 6.

C Comparison with More Works on VideoMME

In Tab 7, we compare our model with previous MLLMs and concurrent works on VideoMME, which includes short, medium, and long video understanding tasks. Our model consistently outperforms prior open-source models in terms of accuracy across all three subtasks, as well as overall accuracy. Notably, our 7B model surpasses several proprietary models, including GPT-4V and Claude 3.5 Sonnet, and even achieves competitive results against the recently released GPT-4o-mini.

D Ablations on Video Segmentation

Our analysis reveals that semantic-based temporal segmentation outperforms fixed-length segmentation for long-video dense captioning. Visual information density varies substantially across tempo-

Model	Frame	VideoMME (w/o subtitles)			
		Short	Medium	Long	Overall
Proprietary Models					
GPT-4V (OpenAI, 2023b)	10	70.5	55.8	53.5	59.9
Claude 3.5 Sonnet (Anthropic, 2024)	20	71.0	57.4	51.2	60.0
GPT-4o-mini (OpenAI, 2024)	250	72.5	63.1	58.6	64.8
Open-sourced Models					
Video-LLaVA-7B (Lin et al., 2023)	8	45.3	38.0	36.2	39.9
ST-LLM (Liu et al., 2025)	64	45.7	36.8	31.3	37.9
ShareGPT4Video (Chen et al., 2024b)	16	48.3	36.3	35.0	39.9
Chat-UniVi-V1.5 (Jin et al., 2024)	64	45.7	40.3	35.8	40.6
LongLLaVA-9B (Wang et al., 2024b)	128/256	52.4	42.2	36.4	43.7
VideoLLaMA2-7B (Cheng et al., 2024)	16	56.0	45.4	42.1	47.9
LLaVA-NeXT-7B (Liu et al., 2024a)	32	58.0	47.0	43.4	49.5
LongVA-7B (Zhang et al., 2024a)	128	61.1	50.4	46.2	52.6
LLaVA-Onevision-7B (Li et al., 2024a)	32	-	-	-	58.2
LongVILA-7B (Xue et al., 2024)	256	69.0	58.3	53.0	60.1
LongVITA-1M-14B (Shen et al., 2025)	256	68.6	59.7	53.8	60.7
ProLongVid-7B	256	**75.2**	**64.0**	**54.8**	**64.7**

Table 7: Comparison with state-of-the-art methods on VideoMME (without using subtitles).

Segmentation	Short	Medium	Long	Overall
Fixed-length	72.2	63.1	50.7	62.0
Semantic-based (Ours)	73.4	63.2	51.9	62.8

Table 8: Comparison of video segmentation methods on VideoMME.

ral dimensions in long videos. When we segment videos into semantically coherent segments, we naturally obtain segments of varying lengths that align with meaningful event boundaries. In contrast, fixed-length segmentation often splits semantically coherent events, creating artificial boundaries that lead to substantial semantic redundancy between adjacent segment captions.

For quantitative evaluation, we conduct ablations using 20K videos sampled from 2-5 min videos. We construct a dataset using fixed-length segmentation with 10-second intervals for comparison. Our results in Tab 8 show that models trained on data generated through our semantic-based segmentation method consistently achieve better performance on VideoMME compared to those trained on fixed-length segmentation data.

E Other Details about Training Data

Our data is released in https://github.com/ruiwang2021/ProLongVid. Due to the limitations of the paper format, we place some long video case studies from our training data in our open-source repository.