From Seconds to Hours: Reviewing MultiModal Large Language Models on Comprehensive Long Video Understanding

Anonymous ACL submission

Abstract

The integration of Large Language Models (LLMs) with visual encoders has recently shown promising performance in visual understanding tasks, leveraging their inherent capability to comprehend and generate human-like text for visual reasoning. This paper reviews the advancements in MultiModal Large Language Models (MM-LLMs) for long video understanding. We highlight the unique challenges posed by long videos, including fine-grained spatiotemporal details, dynamic events, and long-term dependencies. We summarize the progress in model design and training methodologies for MM-LLMs understanding long videos and compare their performance on various long video understanding benchmarks. Finally, we discuss future directions for MM-LLMs in long video understanding.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable versatility and capability in understanding and generating human-like text by scaling model size and training data (Raffel et al., 2020; Brown, 2020; Chowdhery et al., 2023; Touvron et al., 2023a). To extend these capabilities to visual understanding tasks, various methods have been proposed to integrate LLMs with specific visual modality encoders, thereby endowing LLMs with visual perception abilities (Alayrac et al., 2022; Li et al., 2023a). Single images or multiple frames are encoded as visual tokens and integrated with textual tokens to help MM-LLMs achieve visual understanding. For long-video understanding, MM-LLMs (Dai et al., 2023; Liu et al., 2024c) are designed to process a larger number of visual frames and diverse events, enabling a wide range of real-world applications such as automatically analyzing highlight reels from sports videos, movies, surveillance footage, and egocentric videos in embodied AI. For example, a robot could learn to make a cup of coffee from a long egocentric video by analyzing key events such as: 1) measuring coffee grounds; 2) adding water to the reservoir; 3) placing coffee grounds in the filter basket; and 4) starting the coffee maker and waiting for it to brew. Modeling long-form videos with complex spatiotemporal details and dependencies remains a challenging problem (Wang et al., 2023a; Mangalam et al., 2024; Xu et al., 2024b; Wu et al., 2024).

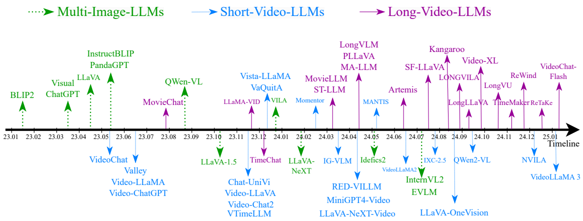


Figure 1: The development of MM-LMMs for multiple images, short videos and long videos.

There are substantial differences between long video understanding (LVU) and other visual understanding tasks. Unlike static image understanding, which focuses solely on spatial content, short video understanding must account for within-event temporal information across sequential frames (Li et al., 2023b; Zhang et al., 2023; Maaz et al., 2023). Long videos, typically exceeding one minute (Wu and Krahenbuhl, 2021; Zhang et al., 2024d), consist of multiple events with varying scenes and visual content, requiring the capture of significant between-event and long-term variations for effective understanding. Balancing spatial and temporal details with a limited number of visual tokens is a considerable challenge for Long-Video-LLMs (LV-LLMs) (Song et al., 2024a; He et al., 2024; Xu et al., 2024b). Additionally, unlike short videos that span only a few seconds and contain tens of frames, long videos often encompass thousands of frames (Ren et al., 2024a; Zhang et al., 2024d). Therefore, LV-LLMs must be capable of memorizing and continually learning long-term correlations

	Image-LLMs	Video-LLMs	Long-Video-LLMs
Task	• Image understanding: ∘ Spatial reasoning: e.g. (Changpinyo et al., 2022; Chen et al., 2024a; Mathew et al., 2021; Peng et al., 2024; Sohoni et al., 2020; Wei et al., 2021).	• Short video understanding: ∘ Spatial reasoning: e.g. (Li et al., 2023b; Ranasinghe et al., 2024). ∘ Within-event reasoning: e.g. (Diba et al., 2023; Huang et al., 2018).	• Long video understanding: ∘ Spatial reasoning: e.g. (Fu et al., 2024a). ∘ Within-event reasoning: e.g. (Cheng et al., 2024) . ∘ **Between-event reasoning:** e.g. (Qian et al., 2024). ∘ **Long-term reasoning:** e.g. (Wu et al., 2024).
Backbone	• Visual encoder: CLIP-ViT (Radford et al., 2021), SigLIP-ViT (Zhai301 et al., 2023), etc. • LLM: LLaMA (Touvron et al., 2023b), etc.	• Visual encoder: CLIP-ViT (Radford et al., 2021), SigLIP-ViT (Zhai301 et al., 2023), etc. • LLM: LLaMA (Touvron et al., 2023b), etc.	• Visual encoder: CLIP-ViT (Radford et al., 2021), SigLIP-ViT (Zhai301 et al., 2023), etc. • **Long-context LLM:** LLaMA3.1 (Dubey et al., 2024), etc.
Connector	• Image-level connector: ∘ Linear-layer-based: e.g. (Liu et al., 2024a; Liu et al., 2024c; Su et al., 2023) ∘ Pooling-based: e.g. (Liu et al., 2024b; Maaz et al., 2023; Xu et al., 2024a) ∘ Transformer-based: e.g. (Dai et al., 2023; Bai et al., 2023b; Jiang et al., 2024))	• Image-level connector: ∘ Image-Q-Former, Spatial-pooling, etc. e.g. (Liu et al., 2024a; Li et al., 2023b; Maaz et al., 2023; Li et al., 2024f) • Video-level connector ∘ Video-Q-Former, Temporal-pooling, etc. e.g. (Zhang et al., 2023; Luo et al., 2023)	• Image-level connector. • Video-level connector. • **Long-video-level connector:** ∘ **Efficient token-compression:** e.g. (Song et al., 2024a; Xu et al., 2024a; Xu et al., 2024b) ∘ **Time-aware design:** e.g. (Huang et al., 2024a; Ma et al., 2023b; Qian et al., 2024; Ren et al., 2024)
Training	• Pre-training: Image-text pairs. e.g. (Chen et al., 2015; Sharma et al., 2018; Chen et al., 2023b). • Instruction-tuning: Image-language instruction data. e.g. (Chen et al., 2023b; Liu et al., 2024c)	• Pre-training: Image-, Short-video-text pairs. e.g. (Chen et al., 2015; Sharma et al., 2018; Chen et al., 2023b; Bain et al., 2021). • Instruction-tuning: Image-, short-video-language instruction data. e.g. (Maaz et al., 2023)	• Pre-training: Image-, video-, **long-video-text pairs**. e.g. (Bain et al., 2021; Zhang et al., 2024d). • Instruction-tuning: Image-, short-video-, **long-video-language instruction data**. e.g. (Li et al., 2023c; Huang et al., 2024a; Ren et al., 2024; Qian et al., 2024)

Figure 2: The comparison of MM-LLMs among Image-, Short-Video-, and Long-Video-LLMs. The **bold content** often highlights special considerations of LV-LLMs for LVU.

in videos that span minutes or even hours.

We summarize the comparison of MM-LLMs among Image-, Short-Video-, and LV-LLMs in Fig. 2. LV-LLMs build upon advancements in multi-image and short-video MM-LLMs, sharing a similar structure of visual encoders, LLM backbones, and cross-modality connectors. To address the challenges in LVU, LV-LLMs incorporate more efficient long-video-level connectors that bridge cross-modal representations and compress visual tokens to a manageable number (Li et al., 2023c; Zhang et al., 2024d). Additionally, time-aware modules enhance the capture of temporal information (Qian et al., 2024). For pre-training and instruction-tuning, video-text pairs and video-instruction data are essential for MM-LLMs to handle both images and videos with shared spatial perception and reasoning capacity (Li et al., 2023b). Long video training datasets are particularly beneficial for temporal cross-modal semantic alignment and capturing long-term correlations, crucial for LV-LLMs (Song et al., 2024b). Our survey provides a comprehensive summary of recent advances in model design and training methods, tracing the evolution from images to long videos.

Recent surveys on visual understanding tasks typically adopt a single perspective, either from a global view of reviewing MM-LLMs (Yin et al., 2023; Zhang et al., 2024a) or from a local view focusing on image- or video-understanding tasks (Zhang et al., 2024b; Nguyen et al., 2024). While these works provide extensive reviews, they often lack discussing the developmental and inheritance relationships between different tasks and methods. Additionally, existing reviews on video understanding (Tang et al., 2023) focus more on general video understanding rather than the more challenging task of LVU. Long videos are prevalent in education, entertainment, and transportation, necessitating comprehensive automatic understanding with powerful models (Apostolidis et al., 2021). Our work is among the earliest to summarize and discuss the LVU task from a developmental perspective.

Our survey is structured as follows: firstly, we find that the LVU task is more complex compared with image and short video understanding tasks (Sec.2.1), and summarize the unique challenges of LVU in Sec.2.2. Next, we provide a detailed summary of the developments in MM-LLMs from the perspectives of model architecture (Sec.3) and training methodologies (Sec.4), with an emphasis on the implementation of LV-LLMs for comprehensive LVU. We then compare the performance of video LLMs on LVU benchmarks (Sec.5), offering insights into the existing results of LV-LLMs. Finally, we discuss future research directions in LVU to advance the research field in Sec.6.

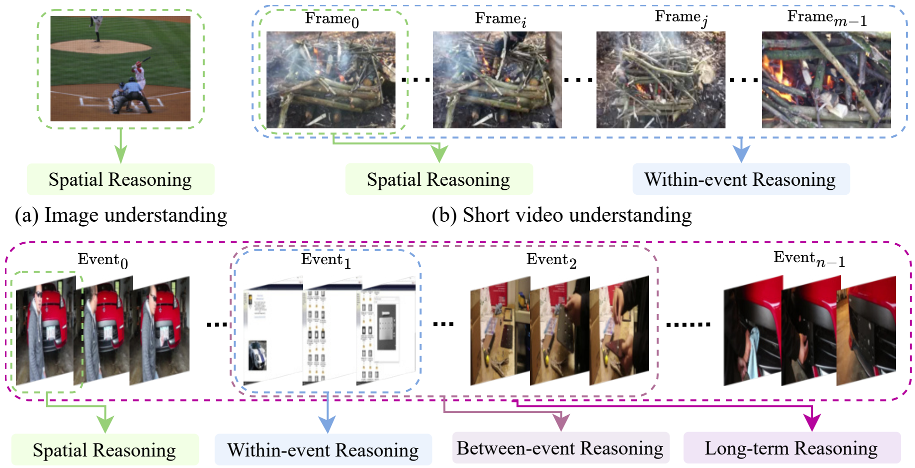
2 Long Video Understanding

In this section, we elaborate on visual understanding tasks among images, short-videos, and long-videos, and further analyze the challenges for LVU.

2.1 Visual Understanding

Visual understanding demands models to interpret visual information, integrating multimodal perception with commonsense reasoning (Johnson et al., 2017; Chen et al., 2024c).

Image understanding. As illustrated in Fig. 3 (a), image understanding tasks involve a single image for various visual reasoning tasks, such as image captioning and image-centered question answering (Sharma et al., 2018; Mathew et al., 2021; Changpinyo et al., 2022; Li et al., 2023a; Chen et al., 2024a). These tasks focus solely on spatial information, encompassing both coarse-grained under-



(a) Image understanding

(b) Short video understanding

(c) Long video understanding

Figure 3: Visual understanding of (a) images, (b) short videos, and (c) long videos.

standing (Ordonez et al., 2011; Sohoni et al., 2020) of global visual context and fine-grained understanding (Wei et al., 2021; Liu et al., 2024b; Peng et al., 2024) of local visual details.

Short video understanding. Unlike image understanding tasks, which involve only static visual data, short video understanding also incorporates temporal information from multiple visual frames (Xu et al., 2016; Bain et al., 2021; Li et al., 2023b, 2024e). In addition to spatial reasoning (Ranasinghe et al., 2024), within-event temporal reasoning and spatiotemporal reasoning across frames play crucial roles for short video understanding (Huang et al., 2018; Lin et al., 2019; Diba et al., 2023).

Long video understanding (LVU). Long videos typically consist of multiple events, encompassing much richer spatial content and temporal variations compared to short videos (Mangalam et al., 2024; Li et al., 2024f; Song et al., 2024a,b). As summarized in Fig. 3 (c), LVU involves not only spatial and within-event temporal reasoning but also between-event reasoning and long-term reasoning from different video events (Wu et al., 2019; Wu and Krahenbuhl, 2021; Wang et al., 2023a; Zhou et al., 2024; Fang et al., 2024).

2.2 Challenges of Long Video Understanding

Compared with images and short videos, long-form videos introduce new challenges to comprehensive visual understanding, as follows:

Rich fine-grained spatiotemporal details. Long videos, which cover a wide range of topics, scenes, and activities, contain varying details such as objects, events, and attributes (Fu et al., 2024a; Wu et al., 2024). These details are much richer compared to static images and short videos with multiple similar frames, making LVU more challenging. For instance, fine-grained spatial question answering can be introduced in any frame, while temporal question answering can be introduced between or among frames for long video reasoning tasks (Song et al., 2024a). MM-LLMs for LVU must capture all relevant fine-grained spatiotemporal details from video frames spanning minutes or even hours, using a limited number of visual tokens.

Dynamic events with scene transitions and content changes. Long videos often contain various dynamic events with significant differences in scenes and content (Wu et al., 2024). These events can be semantically related and temporally coordinated according to their order of appearance (Bao et al., 2021), or they can exhibit significant semantic differences due to plot twists (Papalampidi et al., 2019). Between-event reasoning involving multiple events with diverse visual information is crucial for accurate content understanding (Cheng et al., 2024a; Qian et al., 2024). Distinguishing semantic differences and maintaining semantic coherence across varying events are essential for LVU.

Long-term correlation and dependencies. Long videos often contain actions and events that span extended periods. Capturing long-term dependencies and understanding how different parts of the video relate to each other over the long period is challenging (Wu et al., 2019). Video LLMs designed for images or short videos typically fail to contextualize the present event in relation to past or future events that are far from the current time (Wu and Krahenbuhl, 2021), as well as in long-term decision-making (Wang et al., 2024b).

3 Advances in Model Architecture

In this section, we discuss the advances of MM-LLMs from image-targeted to long-video-targeted models, from the perspective of model architecture.

Model	Year	Backbone		Connector			#Frame	#Token
		Visual Encoder	LLMs	Image-level	Video-level	Long-video-level		
InstructBLIP (2023)	23.05	EVA-CLIP-ViT-G/14	FlanT5, Vicuna-7B/13B	Q-Former	–	–	4	32/128
VideoChat (2023b)	23.05	EVA-CLIP-ViT-G/14	StableVicuna-13B	Q-Former	Global multi-head relation aggregator	–	8	/32
Video-LLaMA (2023)	23.06	EVA-CLIP-ViT-G/14	LLaMA, Vicuna	Q-Former	Q-Former	–	8	/32
Video-ChatGPT (2023)	23.06	CLIP-ViT-L/14	Vicuna1.1-7B	Spatial-pooling	Temporal-pooling	–	100	/356
Valley (2023)	23.06	CLIP-ViT-L/14	StableVicuna-7B/13B	–	Transformer and Mean pooling	–	0.5 fps	/256+T
MovieChat (2024a)	23.07	EVA-CLIP-ViT-G/14	LLama-7B	Q-Former	Frame mergin, Q-Former	Merging adjacent frames	2048	32/32
Qwen-VL (2023b)	23.08	Openclip-ViT-bigG	Qwen-7B	Cross-attention	–	–	4	/256
Chat-UniVi (2024)	23.11	CLIP-ViT-L/14	Vicuna1.5-7B	Token merging	–	–	64	/112
Video-LLaVA (2023)	23.11	LanguageBind-ViT-L/14	Vicuna1.5-7B	–	–	–	8	256/2048
LLaMA-VID (2023c)	23.11	CLIP-ViT-L/14	Vicuna-7B/13B	Context attention and pooling			1 fps	2/
VTimeLLM (2024a)	23.11	CLIP-ViT-L/14	Vicuna1.5-7B/13B	Frame feature	–	–	100	1/100
VideoChat2 (2024e)	23.11	EVA-CLIP-ViT-G/14	Vicuna0-7B	–	Q-Former	–	16	/96
Vista-LLaMA (2023a)	23.12	EVA-CLIP-ViT-G/14	LLaVa-Vicuna-7B	Q-Former	Temporal Q-Former	–	16	32/512
TimeChat (2024a)	23.12	EVA-CLIP-ViT-G/14	LLaMA2-7B	Q-Former	Sliding window Q-Former	Time-aware encoding	96	/96
VaQuitA (2023b)	23.12	CLIP-ViT-L/14	LLaVA1.5-LLaMA-7B	–	Video Perceiver, VQ-Former	–	100	/356
Dolphins (2023b)	23.12	CLIP-ViT-L/14	OpenFlamingo	Perceiver Resamplar, Gated cross-attention		Time embedding	–	–
Momentor (2024)	24.02	CLIP-ViT-L/14	LLaMA-7B	Frame feature, Temporal Perception Module, Grounded Event-Sequence Modeling			300	1/300
MovieLLM (2024b)	24.03	CLIP-ViT-L/14	Vicuna-7B/13B	Context attention and pooling			1 fps	2/
MA-LMM (2024)	24.04	EVA-CLIP-ViT-G/14	Vicuna-7B	Q-Former	Memory Bank Compression	Merging adjacent frames	100	/32
PLLaVA (2024a)	23.04	CLIP-ViT-L/14	LLaVA-Next-LLM	Adaptive Pooling			64	2304
LongVLM (2024)	23.04	CLIP-ViT-L/14	Vicuna1.1-7B	Hierarchical token merging			100	/305
MiniGPT4-Video (2024a)	24.04	EVA-CLIP-ViT-G/14	LLaMA2-7B, Mistral-7B	Merging adjacent tokens	–	–	90	64/5760
RED-VILLM (2024b)	24.04	Openclip-ViT-bigG	Qwen-7B	Spatial pooling	Temporal pooling	–	100	/1124
ST-LLM (2024e)	24.04	BLIP-2	InstructBLIP-Vicuna1.1-7B	Q-Former	Masked video modeling	Global-Local input	16	/512
LLaVA-NeXT-Video (2024e)	24.04	CLIP-ViT-L/14	Vicuna1.5-7B/13B Nous-Hermes-2-Yi-34B	Merging adjacent tokens	–	–	32	4608
Mantis-Idefics2 (2024)	24.05	SigLIP-SO400M	Mistral0.1-7B	Perceiver resampler	–	–	8	64/512
VideoLLaMA 2 (2024b)	24.06	CLIP-ViT-L/14	Mistral-7B-Instruct	Spatial-Temporal Convolution		–	8	/576
LongVA (2024d)	24.06	CLIP-ViT-L/14	Qwen2-7B-224K	Merging adjacent tokens	Expanding tokens	–	384	55,296
Artemis (2024)	24.06	CLIP-ViT-L/14	Vicuna1.5-7B	Average pooling			5	/356
VideoGPT+ (2024)	24.06	CLIP-ViT-L/14 InternVideo-v2	Phi3-Mini-3.8B	Adaptive pooling	Adaptive pooling	–	16	/2560
IXC-2.5 (2024c)	24.07	CLIP-ViT-L/14-490	InternLM2-7B	Merging adjacent tokens	Expanding tokens	Frame index	64	400/25600
EVLM (2024b)	24.07	EVA2-CLIP-E-Plus	Qwen-14B-Chat 1.0	Gated cross attention	–	–	–	/16
SlowFast-LLaVA (2024b)	24.07	CLIP-ViT-L/14	Vicuna1.5-7B	Merging adjacent tokens	Slow and fast pathway		50	3680
LLaVA-Interleave (2024d)	24.07	SigLIP-SO400M	Qwen1.5-0.5B/7B/14B	–	–	–	16	729/11664
Kangaroo (2024d)	24.08	EVA-CLIP-ViT-G/14	LLaMA3-8B	3D Depthwise convolution			–	–
VITA (2024b)	24.08	InternViT-300M-448px	Mixtral 8x7B	MLP	–	–	16	256/4096
LLaVA-OneVision (2024c)	24.08	SigLIP-SO400M	Qwen2-7B	Merging adjacent tokens	–	–	1 fps	729/
LONGVILA (2024)	24.08	SigLIP-SO400M	Qwen2-1.5B/7B	Multi-Modal Sequence Parallelism			1024	256/
LongLLaVA (2024e)	24.09	CLIP-ViT-B/32	LLaVA1.6-13B	Merging adjacent tokens	Mamba Layers	Hybrid architecture	256	144/
Qwen2-VL (2024a)	24.09	CLIP-ViT-L/14	Qwen2-1.5B/7B/72B	Merging adjacent tokens	3D convolutions	–	2 fps	66/
Video-XL (2024)	20.09	CLIP-ViT-L/14	Qwen-2-7B	Merging adjacent tokens	Visual Summarization Token and Dynamic Compression		128	–
Oryx-1.5 (2024g)	24.10	OryxViT	Qwen-2.5-7B/32B	Variable-Length Self-Attention	Dynamic Compressor		64	256/
LongVU (2024)	24.10	SigLIP-SO400M	Qwen2-7B, LLaMA3.2-3B	Selective frame feature reduction	Frame selection and Token Reduction		1 fps	–
TimeMarker (2024d)	24.11	LLaVA-Encoder	LLaVA-LLM	Adaptive Token Merge and Temporal Separator Tokens Integration			128	–
ReWind (2024)	24.11	EVA-CLIP-ViT-G/14	LLaMA2-7B	Pooling, Learnable queries, Cross-attentions and Frame Selection			64	32/256
NVILA (2024f)	24.12	SigLIP-SO400M	Qwen2-7B/14B	Spatial-to-Channel Reshaping	Temporal Averaging		256	/8192
IQViC (2024)	24.12	CLIP-ViT-L/14	d Vicuna-v1-7B	Visual Compressor	Temporal Compressor		–	/640
ReTaKe (2024c)	24.12	CLIP-ViT-L/14	Qwen2-7B	Key frame selection, KV-Cache compression			1024	–
VideoLLaMA 3 (2025a)	25.01	SigLIP-SO400M	Qwen-2.5-7B	Any-resolution Vision Tokenization	Differential Frame Pruner		180	/10240
LLaVA-Mini (2025b)	25.01	CLIP-ViT-L/14	LLaMA-3.1-8B-Instruct	Vision Token Compression	–	–	10000	1/10000
VideoChat-Flash (2025)	25.01	UMT-L@224	Qwen2-7B	Hierarchical Visual Token Compression			1000	16/

Table 1: Comparison of mainstream Video-LLMs across model design choices. The notation A/B in the last column indicates A tokens per frame and a total of B tokens for the entire video.

3.1 Visual Encoder and LLM Backbone

MM-LLMs, encompassing both image-targeted and video-targeted models, typically utilize similar visual encoders for visual information extraction. LLM backbones are also universal in early MM-LLM methods, while existing LV-LLMs tend to use long-context LLMs in the implementation.

Visual encoder. The pretrained visual encoders are responsible for capturing vision knowledge from raw visual data. As summarized in Table 1, image encoders like CLIP-ViT-L/14 (Radford et al., 2021), EVA-CLIP-ViT-G/14 (Sun et al., 2023) and SigLIP-SO400M (Zhai et al., 2023) are widely utilized as visual modality encoders in image- and video-targeted LLMs. Recent work (Li et al., 2024a) shows that the visual representation, including image resolution, the size of visual token, and the pre-training visual resources, play a more important role than the size of the visual encoder.

LLM backbone. The LLM is the core module in visual understanding systems, inheriting properties of reasoning and decision-making. The strength of the LLM typically correlates with superior multimodal capabilities in visual LLMs (Li et al., 2024b,a). For LLMs of equivalent scale, those with superior language capabilities demonstrate enhanced performance, whereas for the same LLMs with varying model sizes, larger models generally achieve better multimodal performance. Additionally, long-context LLMs that extend the context length to hundreds of thousands of tokens support learning with more extensive data (Yang et al., 2024). Recent LV-LLMs effectively transfer the LLM's long-context understanding ability to the vision modality (Zhang et al., 2024d).

3.2 Modality Interface

Connectors between visual encoders and LLMs act as modality interfaces, mapping visual features to the language space. Due to the variability in visual data sources, these connectors are categorized into image-level, video-level, and long-video-level types. (More image- and short-video-level connector designs are summarized in Appendix B.3.)

Image-level connectors. Image-level connectors aim to map image features to the language space for processing raw visual tokens and are widely used in both image- and video-targeted MM-LLMs. These connectors fall into three categories: (1) single linear layers (Liu et al., 2024c) or multi-layer perceptrons (MLPs) (Liu et al., 2024a) for embedding, (2) pooling-based methods, and (3) cross-attention or transformer-based structures like Q-Former (Li et al., 2023a) and Perceiver Resampler (Jaegle et al., 2021) for feature compression.

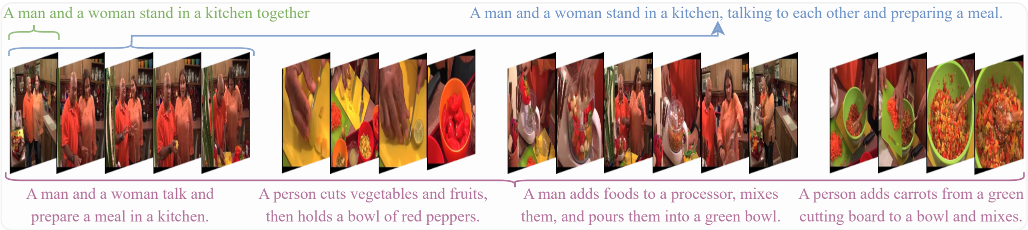
Video-level connectors. Video-level connectors are used for extracting sequential visual data and further compressing visual features. Compared to the solely image-level connectors in image-targeted MM-LLMs, video-level connectors are essential for video-targeted MM-LLMs, including LV-LLMs. Some methods directly concatenate image tokens before inputting them to the LLMs, making them sensitive to the number of frame images (Dai et al., 2023; Lin et al., 2023). Similar structures used for token compression in image-level connectors can be adapted for video-level interfaces, such as pooling-based and transformer-based structures (Maaz et al., 2023; Song et al., 2024a; Zhang et al., 2023; Ma et al., 2023a; Ren et al., 2024a).

Long-video-level connectors. Long-video-level connectors focus more on efficient visual data compression and long-term information preserving.

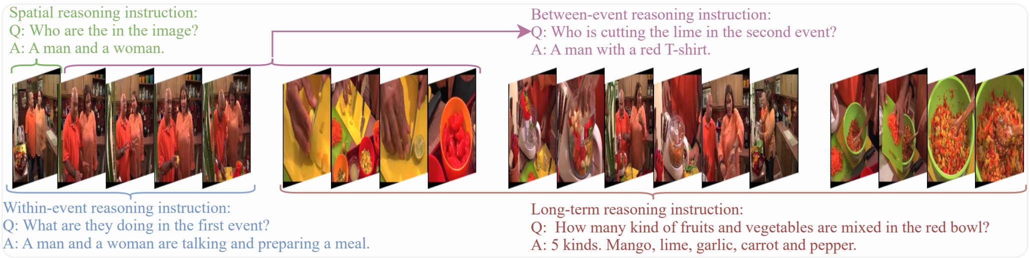
Efficiently compressing visual information requires not only reducing the input visual tokens to an acceptable quantity but also preserving the complete spatiotemporal details contained in long videos. Videos contain two types of data redundancy: spatial data redundancy within frames and spatiotemporal data redundancy across frames (Li et al., 2022; Chen et al., 2023a; Cheng et al., 2024c). On the one hand, spatial data redundancy arises when region-level pixels within frames are the same, leading to inefficiencies when representing the redundant visual frame through full visual tokens. To reduce spatial video data redundancy, the LLaVA-Next-series methods (Zhang et al., 2024e; Li et al., 2024d; Liu et al., 2024b; Li et al., 2024c) merge adjacent frame patch tokens, and Chat-UniVi (Jin et al., 2024) merges similar frame patch tokens. On the other hand, spatiotemporal data redundancy includes both cross-frame pixel redundancy and motion redundancy (Pourreza et al., 2023), where the semantic information is similar among these redundant video frames. To reduce spatiotemporal video redundancy, MovieChat (Song et al., 2024a) and MA-LMM (He et al., 2024) merge frame features with higher frame similarity before inputting them to LLMs. In addition to reducing redundant information, preserving more video spatiotemporal details is crucial for accurate long video reasoning (Diba et al., 2023). To balance global and local visual information and support more frame inputs, SlowFast-LLaVA (Xu et al., 2024b) employs a slow pathway to extract features at a low frame rate while retaining more visual tokens, and a fast pathway at a high frame rate with a larger spatial pooling stride to focus on motion cues.

Additionally, time-involved visual data efficiently manage the temporal and spatial information inherent in long-form videos (Hou et al., 2024). The time-aware design can enhance the temporal-capturing capability of video-related LLMs, which is particularly beneficial for LVU. Both VTimeLLM (Huang et al., 2024a) and InternLM-XComposer-2.5 (IXC-2.5) (Zhang et al., 2024c) use frame indices to enhance temporal relations. The difference lies in their approach: VTimeLLM learns temporal information by training with decoded text that includes frame indices, while IXC-2.5 encodes frame indices along with the frame image context. TimeChat (Ren et al., 2024a) and Momentor (Qian et al., 2024) inject temporal information directly into frame features for fine-grained temporal information capture. Specifically, TimeChat designs a Time-aware Frame Encoder to extract visual features with corresponding timestamp descriptions at the frame level, while Momentor utilizes a Temporal Perception Module for continuous time encoding and decoding, injecting temporal information into frame features.

Retrieval-based LVU. A significant proportion of LVU methods are retrieval-based, addressing challenges like "noise and redundancy" and "memory and computation" constraints. R-VLM selects the most relevant video chunks for question answering (Xu et al., 2023), Goldfish retrieves top-clips to focus on pertinent segments (Ataallah et al., 2024b), and DrVideo transforms videos into text documents to retrieve key frames (Ma et al., 2024). Video-RAG uses visually-aligned auxiliary texts for cross-modality alignment (Luo et al., 2024), while VideoLLaMB employs temporal memory tokens and a SceneTilling algorithm to preserve semantic continuity (Wang et al., 2024f). VideoAgent (Wang et al., 2024d) leverages a LLM to iteratively compile critical information, using vision-language models to enhance LVU.



(a) Pretraining data difference among image, short video and long video descriptions



(b) Long video instruction-tuning data with spatial, within-event, between-event and long-term reasoning instructions.

Figure 4: Long video sample for pretraining and instruction-tuning.

4 Advances in Model Training

Multimodal LLMs for visual understanding consist of two principal stages: pre-training (PT) for vision-language feature alignment and instruction-tuning (IT) for reasoning response (seen in Appendix B.4).

4.1 Pre-training

Vision-language pre-training for MM-LLMs aims to align visual features with the language space using text-paired data. This includes pre-training with image-, short-video-, and long-video-text datasets. Initially introduced for visual LLMs focused on images, **image-text pre-training** is also widely used in video-related understanding tasks. Coarse-grained image-text pair datasets, such as COCO Captions (Chen et al., 2015) and CC-3M (Sharma et al., 2018), are employed for global vision-language alignment. Fine-grained image-text datasets like ShareGPT4V-PT (Chen et al., 2023b) are used for locally spatial semantics alignment. Given the limited changes in semantic content of short videos, short-video-text paired datasets, such as Webvid-2M (Bain et al., 2021), can be used similarly for **short-video-text pre-training**. Similarly, **long-video-text pre-training** is important to capture the temporal semantic alignment of long videos for LVU (Ju et al., 2025). Given the absence of long-term cross-modal correlation in image-text and short-video-text pairs, long-video-text pre-training datasets with pairs of long videos and their corresponding text descriptions are necessary (Argaw et al., 2023). Moreover, as shown in Fig. 4 (a), the scenes and events in long videos vary significantly across frames, necessitating event-level vision-language alignment (Qian et al., 2024) for long-video-text pre-training, which is markedly different from both image-text and short-video-text pre-training (Zhang et al., 2024d).

4.2 Instruction-tuning

Instruction-tuning with vision-language sources enables LLMs to follow instructions and generate human-like text. Multimodal vision-language instruction-following data (Dai et al., 2023; Liu et al., 2024c), including both image-text and video-text pairs, are used to align multimodal LLMs with human intent, thereby enhancing their ability to complete real-world tasks.

Similar to the pre-training stage, **image-text instruction-tuning** is employed in various vision-understanding tasks, including image, short-video, and long-video understanding. Basic image-based instruction-following datasets, such as ShareGPT4V-Instruct (Chen et al., 2023b) and LLaVA-Instruct (Liu et al., 2024c), provide high-quality instruction-tuning data for spatial reasoning and chat capabilities. For video-related LLMs, **short-video-text instruction-tuning** is necessary to enable multimodal LLMs to understand temporal sequences, as seen in models like Video-ChatGPT (Maaz et al., 2023) and VideoChat (Li et al., 2023b). Short-video-LLMs require both spatial and within-event reasoning instructions to understand the satial and small-scale temporal content of short videos. However, the limited content and semantic changes in short videos are insufficient for LVU tasks, where frames are more numerous and exhibit significant variation. **Long-video-text instruction-tuning** is specifically introduced to better capture and understand long videos. In addition to spatial and within-event reasoning instructions,

Model	LLM	Long	VideoVista (131s)	MMBench-Video (165s)	EgoSchema (180s)	LongVideoBench (473s)	MLVU (12mins)	Video-MME (1024s)	LVBench (4101s)
Momentor	LLaMA-7B	✗	–	–	–	–	–	–	
TimeChat	LLaMA2-7B	✓	–	–	33.0†	–	30.9†	–	22.3$^{\spadesuit}$
LLaMA-VID	Vicuna-7B	✓	56.87‡	–	38.5†	–	33.2†	–	23.9$^{\spadesuit}$
LLaVA-NeXT-Video	Vicuna1.5-7B	✗	56.66‡	–	43.9†	43.5$^{\heartsuit}$	–	–	
VideoLLaMA 2 (16)	Mistral-7B-Instruct	✗	60.47‡	–	51.7	–	48.5†	47.9/50.3$^{\clubsuit}$	–
PLLaVA	LLaVA-Next-7B	✓	60.36‡	1.03†	54.4†	39.2$^{\heartsuit}$	–		
LongVA	Qwen2-7B-224K	✓	67.36‡	–	–	–	56.3†	52.6/54.3$^{\clubsuit}$	–
IXC-2.5-7B	InternLM2-7B	✗	68.91‡	1.41	–	–	58.8	55.8/-	
Kangaroo	LLaMA3-8B	✓	69.50‡	1.44	62.7	54.8	61.0	56.0/57.6$^{\clubsuit}$	39.4
Video-XL	QWen2-7B	✓	70.60	–	–	50.7	–	55.5/61.0	–
TimeMarker	LLaVA-7B	✓	78.40	1.53	–	56.3	–	57.3/62.8	41.3
ReTaKe	Qwen2-7B	✓	–	–	–	–	69.8	63.9/68.9	47.8
VideoLLaMA 3	Qwen-2.5-7B	✓	–	–	63.3	59.8	73.0	66.2/70.3	45.3
VideoChat-Flash	Qwen2-7B	✓	–	–	–	64.2	74.5	64.0/69.4	48.2

Table 2: Comparison of Long-Video-LLMs on LVU benchmarks. Results with ‡ are from the VideoVista benchmark (Li et al., 2024f). Results with † are from the Kangaroo (Liu et al., 2024d). Results with $^{\clubsuit}$ are from Video-MME benchmark (Fu et al., 2024a). Results with $^{\spadesuit}$ are from LVBench (Wang et al., 2024b). Results with $^{\heartsuit}$ are from LongVideoBench (Wu et al., 2024).

between-event and long-term reasoning instructions (Ren et al., 2024b; Zeng et al., 2024) are necessary for comprehensive understanding, as shown in Fig. 4 (b). Among the newly introduced long-video instruction-format datasets, Long-VideoQA (Li et al., 2023c), Video-ChatGPT (Maaz et al., 2023) and LongViTU (Wu et al., 2025) are not time-aware. In contrast, VTimeLLM (Huang et al., 2024a), TimeIT (Ren et al., 2024a), and Moment-10M (Qian et al., 2024) are time-aware, incorporating temporal information to enhance reasoning.

5 Evaluation, Performance and Analysis

This section presents a performance comparison across popular evaluation datasets with videos of varying lengths, along with our analysis. Additional comparisons are provided in Appendix C.

To address the unique characteristics of long videos, several long video benchmarks have been introduced in recent years, with video lengths varying from hundreds of seconds to thousands of seconds. EgoSchema (Mangalam et al., 2024) is long-form video understanding datasets designed for multiple-choice question answering, after accessing all frames. VideoVista (Li et al., 2024f), MMBench-Video (Fang et al., 2024), and MLVU (Zhou et al., 2024) cover various topics and are designed for fine-grained capability evaluation. LongVideoBench (Wu et al., 2024) introduces referring reasoning questions to address the longstanding issue of single-frame bias in long videos. Video-MME (Fu et al., 2024a) and LVBench (Wang et al., 2024b) contain numerous hour-level videos. Video-MME further categorizes them into short, medium, and long categories, while LVBench aims to challenge models to demonstrate long-term memory and extended comprehension capabilities.

We further compare and analyze the performance of LVU, specifically summarizing their performance on long video benchmarks with lengths varying from hundreds of seconds to thousands of seconds. As shown in Table 2, LVU-specific methods typically outperform short video understanding methods on LVU tasks. This indicates that specially designed, powerful video-level connectors are essential for LVU. Additionally, the performance on benchmarks with longer video lengths is generally worse than on those with shorter lengths. For example, the performance of methods across VideoVista and MLVU, Video-MME and LVBench, using the same evaluation metric, shows a decline as video length increases. This suggests that LVU remains a challenging research topic.

6 Future Directions

To meet the demands of an AI-driven society with increasingly longer multimodal data, developing more powerful visual LLMs for LVU is crucial.

6.1 More Long Video Training Resources

The two-stage training pipeline, consisting of cross-modal alignment pre-training and visual-instruction tuning, is widely employed for training MM-LLMs (Dai et al., 2023; Liu et al., 2024c). However, there are several challenges for LVU:

- **Hour-long video datasets.** The length of newly introduced long-video training data is limited to minutes, restricting effective reasoning for hour-long LVU (Li et al., 2023c).
- **Necessity of long video pre-training.** Fine-grained long-video-language training pairs are lacking compared to image- and short-video-language pairs during pre-training (Song et al., 2024b; Qiu et al., 2024). Exploring the necessity of long-video-language paired datasets is crucial for evaluating the value of capturing long-term correlations in long videos (Zhang et al., 2024d).
- **Large-scale long video instruction-tuning datasets.** Existing long video datasets, mentioned

in Sec 4.2 are limited in size. Creating large-scale long-video-instruction datasets is essential for comprehensive long-video understanding.

6.2 More Challenging LVU Benchmarks

Recent video understanding benchmarks, such as LongVideoBench (Wu et al., 2024), VideoVista (Li et al., 2024f), and MLVU (Zhou et al., 2024), focus on specific aspects of LVU like long-context interleaved and fine-grained video understanding. However, comprehensive benchmarks that cover frame-level and segment-level reasoning with time and language are necessary but currently unexplored for a thorough evaluation of general LVU methods (Wu et al., 2024). Existing benchmarks, typically at the minute level, fail to adequately test long-term capabilities. LVU methods often suffer from catastrophic forgetting and loss of spatiotemporal details when reasoning with extensive sequential visual information (Wang et al., 2024b), such as hour-level videos. Additionally, most LVU benchmarks focus solely on the visual modality. Incorporating multimodal data, including audio and language, would significantly benefit LVU tasks.

6.3 Powerful and Efficient Frameworks

Visual LLMs for videos need to support more visual frames and preserve more visual details with a fixed number of visual tokens. There are four main considerations when implementing LV-LLMs:

- **Select long-context LLMs as the LLM backbones.** Previous methods have suffered from limited context capacity and required specific fine-tuning to support more tokens (Zhang et al., 2024d). Recent long-context LLMs, such as QWen2 (Yang et al., 2024) and LLaMA-3.1 (Dubey et al., 2024), offer a context window length of 128K and can be utilized in LV-LLM without extensive fine-tuning.
- **Compress visual tokens efficiently with minimal information loss.** Existing methods face issues with insufficient or excessive compression. For example, Chat-UniVi (Jin et al., 2024) uses multi-scale token merging, and LongVA merges adjacent tokens only, while LLaMA-VID (Li et al., 2023c) and MA-LMM (He et al., 2024) compress too much visual information, leading to significant loss of frame details. New frameworks must efficiently compress visual tokens to support more temporal frames and preserve spatiotemporal details. At the image level, adjacent frames can be merged or represented with fewer visual tokens due to their similarity and redundancy (Kim et al., 2024; Xu et al., 2024b). At the video level, relatively independent video events can be compressed into single visual units with corresponding visual tokens, allowing the inputs to cover long-form visual content effectively. Additionally, retrieval-based methods address challenges like "noise and redundancy" and "memory and computation" constraints by leveraging an LLM to iteratively compile critical information(Xu et al., 2023; Ataallah et al., 2024b).
- **Incorporate time-aware designs.** Enhance video reasoning by incorporating temporal information, as seen in designs like TimeIT (Ren et al., 2024a) and Moment-10M (Qian et al., 2024), to improve temporal information extraction in LVU tasks. Temporal information can be injected at various levels: token level, image level, or event level, significantly enhancing the model's ability to understand and reason about long videos.
- **Utilize infrastructure for memory-intensive training.** To handle the increased data load, it is essential to have infrastructure that supports memory-intensive long-context training. Employ infrastructure capable of supporting long-context training with a large number of GPU devices, as demonstrated by LongViLa (Xue et al., 2024), ensuring efficient training on long-form content.

7 Conclusion

In this paper, we summarize the advances of visual LLMs from images to long videos. By analyzing the task differences among image understanding, short video understanding, and long video understanding, we identify key challenges in long video learning. These challenges include capturing fine-grained spatiotemporal details and long-term dependencies within compressed visual information from dynamic sequential events with scene transitions and content changes. We then introduce advances in model architecture and training from Image-LLMs to Long-Video-LLMs, aimed at improving LVU and reasoning. Following this, we review multiple video benchmarks of varying lengths and compare the video understanding performance of various methods, providing insights into future research directions for LVU. Our paper is the first to focus on the development and improvement of Long-Video-LLMs for better LVU. We hope our work will contribute to the advancement of LVU and reasoning with LLMs.

Limitation

We reviewed literature on comprehensive long video understanding, covering methods, training datasets, and benchmarks. Due to space constraints, we omit detailed application scenarios like real-time processing and multimodal tasks. We will maintain an open-source repository and add these contents to complement our survey. The performance comparisons are based on final results from previous papers and official benchmarks, which vary in training resources, strategies, and model architectures, making it difficult to analyze specific models and training differences. We plan to conduct detailed ablation studies on public benchmarks for a more direct analysis of model design, training resources, and methods.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Dawit Mureja Argaw, Joon-Young Lee, Markus Woodson, In So Kweon, and Fabian Caba Heilbron. 2023. Long-range multimodal pretraining for movie understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13392–13403.

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

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. 2024b. Goldfish: Vision-language understanding of arbitrarily long videos. *arXiv preprint arXiv:2407.12679*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738.

Peijun Bao, Qian Zheng, and Yadong Mu. 2021. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 920–928.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. 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.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Jou-An Chen, Wei Niu, Bin Ren, Yanzhi Wang, and Xipeng Shen. 2023a. Survey: Exploiting data redundancy for optimization of deep learning. *ACM Computing Surveys*, 55(10):1–38.

Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, et al. 2024b. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2024c. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36.

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

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024d. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Dingxin Cheng, Mingda Li, Jingyu Liu, Yongxin Guo, Bin Jiang, Qingbin Liu, Xi Chen, and Bo Zhao. 2024a. Enhancing long video understanding via hierarchical event-based memory. *arXiv preprint arXiv:2409.06299*.

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

Zheng Cheng, Rendong Wang, and Zhicheng Wang. 2024c. Focuschat: Text-guided long video understanding via spatiotemporal information filtering. *arXiv preprint arXiv:2412.12833*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Ali Diba, Vivek Sharma, Mohammad Arzani, Luc Van Gool, et al. 2023. Spatio-temporal convolution-attention video network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 859–869.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. 2024. Rewind: Understanding long videos with instructed learnable memory. *arXiv preprint arXiv:2411.15556*.

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

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*.

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

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. 2024b. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Jianlong Hou, Tao Tao, Jibin Wang, Zhuo Chen, Xuelian Ding, and Kai Wang. 2024. Memotichat: A memory-augmented time-sensitive model for ultra-long video understanding. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024a. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. 2018. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375.

Suyuan Huang, Haoxin Zhang, Yan Gao, Yao Hu, and Zengchang Qin. 2024b. From image to video, what do we need in multimodal llms? *arXiv preprint arXiv:2404.11865*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2025. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970.

Hamza Karim, Keval Doshi, and Yasin Yilmaz. 2024. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6848–6856.

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, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: What else influences visual instruction tuning beyond data?

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

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

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

Jiahao Li, Bin Li, and Yan Lu. 2022. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1503–1511.

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 *International conference on machine learning*, pages 19730–19742. PMLR.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024e. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. 2025. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

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

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024f. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

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

Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

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

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024d. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2024e. St-llm: Large language models are effective temporal learners. *arXiv preprint arXiv:2404.00308*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024f. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024g. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.

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*.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2023a. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2023b. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2024. Drvideo: Document retrieval based long video understanding. *arXiv preprint arXiv:2406.12846*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*.

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

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3636–3657, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. *arXiv preprint arXiv:1908.10328*.

Jinlong Peng, Zekun Luo, Liang Liu, and Boshen Zhang. 2024. Frih: Fine-grained region-aware image harmonization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4478–4486.

Reza Pourreza, Hoang Le, Amir Said, Guillaume Sautiere, and Auke Wiggers. 2023. Boosting neural video codecs by exploiting hierarchical redundancy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5355–5364.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*.

Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. 2024. Artemis: Towards referential understanding in complex videos. *arXiv preprint arXiv:2406.00258*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. 2024. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024a. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. 2024b. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. *arXiv preprint arXiv:2412.00927*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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

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

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024a. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. 2024b. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, et al. 2023c. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. 2023a. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397.

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

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2024c. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024d. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

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

Yizhou Wang, Ruiyi Zhang, Haoliang Wang, Uttaran Bhattacharya, Yun Fu, and Gang Wu. 2023b. Vaquita: Enhancing alignment in llm-assisted video understanding. *arXiv preprint arXiv:2312.02310*.

Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. 2024f. Videollamb: Long-context video understanding with recurrent memory bridges. *arXiv preprint arXiv:2409.01071*.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2021. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. Longvlm: Efficient long video understanding via large language models. *arXiv preprint arXiv:2404.03384*.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 284–293.

Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894.

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*.

Rujie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. 2025. Longvitu: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*.

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

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

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. 2023. Retrieval-based video language model for efficient long video question answering. *arXiv preprint arXiv:2312.04931*.

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

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, et al. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Sosuke Yamao, Natsuki Miyahara, Yuki Harazono, and Shun Takeuchi. 2024. Iqvic: In-context, question adaptive vision compressor for long-term video understanding lmms. *arXiv preprint arXiv:2412.09907*.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

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

Rufai Yusuf Zakari, Jim Wilson Owusu, Hailin Wang, Ke Qin, Zaharaddeen Karami Lawal, and Yuezhou Dong. 2022. Vqa and visual reasoning: An overview of recent datasets, methods and challenges. *arXiv preprint arXiv:2212.13296*.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. 2024. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*.

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

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024b. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

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

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025b. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

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

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*.

A Multiple Visual Reasoning

Visual reasoning demands models to comprehend and interpret visual information and integrate multimodal perception with commonsense understanding (Johnson et al., 2017; Chen et al., 2024c). There are three main types of visual reasoning tasks: visual question answering (VQA), visual captioning (VC) or description (VD), and visual dialog (VDia). VQA (Antol et al., 2015; Zakari et al., 2022) involves generating a natural language answer based on the input visual data and accompanying questions. VC and VD systems (Vinyals et al., 2015; Sharma et al., 2018; Li et al., 2019) typically generate a concise, natural language sentence that summarizes the main content of the visual data and a detailed and comprehensive description of the corresponding visual data, respectively. VDia (Das et al., 2017; Qi et al., 2020) involves multi-turn conversations, consisting of a series of question-answer pairs centered around the visual content.

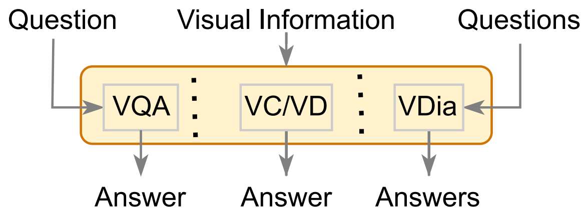


Figure 5: Various visual reasoning tasks.

B Development of MM-LLM Model Architecture

B.1 Multiple MM-LLMs

As illustrated in Fig. 6, MM-LLMs for images, short videos, and long videos share a similar structure comprising a visual encoder, an LLM backbone, and an intermediary connector. Unlike the image-level connector in image-targeted MM-LLMs, the video-level connector is crucial for integrating cross-frame visual information. In LV-LLMs, designing the connector is more challenging, requiring efficient compression of amounts of visual information and incorporating temporal knowledge to manage long-term correlations.

B.2 Mutiple LLM Backbones

Compared to closed-source LLMs like GPT-3/4 (Brown, 2020; Achiam et al., 2023) and Gemini-1.5 (Reid et al., 2024), various open-source LLMs are more commonly used in implementing visual LLMs. These include Flan-T5 (Chung et al., 2024), LLaMA (Touvron et al., 2023b,c; Dubey et al., 2024), Vicuna (Chiang et al., 2023), QWen (Bai et al., 2023a), Mistral (Jiang et al., 2023), Openflamingo (Awadalla et al., 2023), Yi (Young et al., 2024), and InternLM (Team, 2023; Cai et al., 2024).

B.3 Various Connector Designs

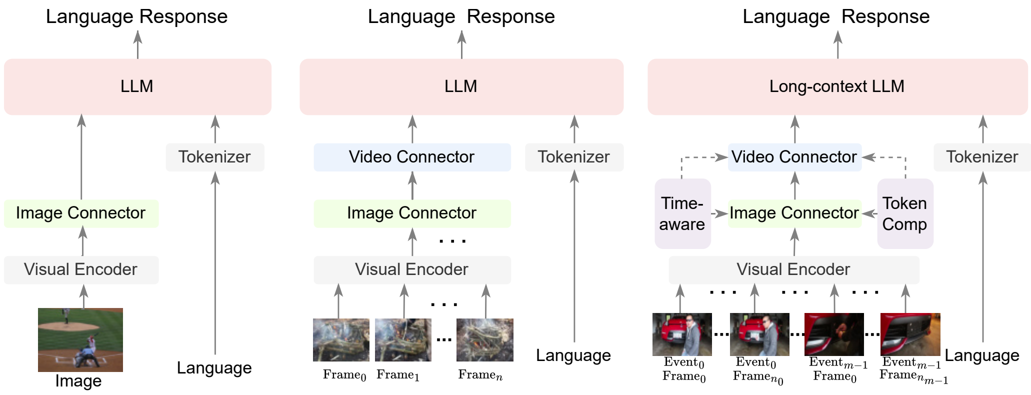
In addition to the detailed discussed long-video-level connectors, the image-level and video-level connectors are also popular.

Image-level connectors. Image-level connectors are used to map image features to the language space for processing raw visual tokens, and they are widely used in both image-targeted and video-targeted MM-LLMs. These connectors can be categorized into three groups: **The first group** directly uses a single linear layer (Liu et al., 2024c) or a multi-layer perceptron (MLP) (Liu et al., 2024a) to map image features into the language embedding space. However, this method, which retains all visual tokens, is not suitable for visual understanding tasks involving multiple images. To address the limitations of retaining all visual tokens, **the second group** employs various pooling-based methods. These include spatial pooling (Maaz et al., 2023), adaptive pooling (Xu et al., 2024a), semantic-similar token merging (Jin et al., 2024), and adjacent token averaging (Zhang et al., 2024e; Li et al., 2024c). **The third group** utilizes cross-attention or transformer-based structures, such as Q-Former (Li et al., 2023a) and Perceiver Resampler (Jaegle et al., 2021), for image feature compression. Q-Former is a lightweight transformer structure that employs a set of learnable query vectors to extract and compress visual features. Many visual LLMs (Dai et al., 2023; Li et al., 2023b; Ma et al., 2023a; Liu et al., 2024e), following BLIP-2, choose the Q-Former-based connector. Other visual LLMs (Ma et al., 2023b; Jiang et al., 2024) opt for the Perceiver Resampler to reduce computational burden by extracting patch features.

Video-level connectors. Video-level connectors are used for extracting sequential visual data and further compressing visual features. Compared to the solely image-level connectors in image-targeted MM-LLMs, video-level connectors are essential for video-targeted MM-LLMs, including LV-LLMs. Some methods directly concatenate image tokens before inputting them to the LLMs, making them sensitive to the number of frame images (Dai et al., 2023; Lin et al., 2023). Similar structures used for token compression in image-level connectors can be adapted for video-level interfaces, such as pooling-based and transformer-based structures. Pooling along the time series dimension is a straightforward way to reduce temporal information redundancy (Maaz et al., 2023; Song et al., 2024a). Transformer-based methods, such as Video Q-Former (Zhang et al., 2023; Ma et al., 2023a; Ren et al., 2024a) and Video Perceiver (Wang et al., 2023b), are effective in extracting video features while reducing data complexity. Additionally, 3D-Convolution-based methods can extract and compress visual data from both the spatial and temporal dimensions (Cheng et al., 2024b; Liu et al., 2024d).

B.4 Training Design for LVU

As shown in Table 3, the training devices and resources used in pre-training and supervised fine-tuning are summarized. Adequate computing power and sufficient training data are essential for developing a robust long video understanding



(a) Image-LLM (b) Short-Video-LLM (c) Long-Video-LLM

Figure 6: MM-LLMs of (b): Image-LLM, (c) Short-Video-LLM and (c) Long-Video-LLM.

model.

C Video Understanding from Seconds to Minutes

As shown in Table 4, we summarize the general video understanding performance of various visual LLMs on open-ended video question answering benchmarks, including TGIF-QA (Jang et al., 2017), MSVD-QA, MSRVTT-QA (Xu et al., 2017), NEXT-QA (Xiao et al., 2021), and ActivityNet-QA (Yu et al., 2019). Additionally, we consider the VideoChatGPT-introduced video-based generative performance benchmark (Maaz et al., 2023), which evaluates five aspects of video-based text generation: Correctness of Information (CI), Detail Orientation (DO), Context Understanding (CU), Temporal Understanding (TU), and Consistency (CO). The video benchmarks with lengths shorter than 1 minute, such as TGIF-QA, MSVD-QA, MSRVTT-QA, and NEXT-QA, are commonly used for short video understanding. In contrast, benchmarks exceeding one minute, such as ActivityNet-QA and the ActivityNet-200-based (Caba Heilbron et al., 2015) generative performance benchmark, are used for long video understanding.
By comparing the performance in Table 4, we can conclude that long video understanding is challenging, with the following findings: **(1)** Video reasoning with more frames introduces more complex visual information and is more challenging. Methods designed to support long videos, such as LongVA (Zhang et al., 2024d), show better performance compared to being fed with fewer frames on the same video dataset. However, performance decreases when being fed with more frames from the same video dataset for methods without special designs for long videos, like VideoLLaMA2 (Cheng et al., 2024b). **(2)** Short video understanding methods that perform well on seconds-level video understanding often do not perform well on minutes-level moderately long video understanding, such as RED-VILLM (Huang et al., 2024b) and MiniGPT4-Video (Ataallah et al., 2024a). Long video understanding methods tend to share consistently good performance on both short and moderately long video benchmarks, such as ST-LLM (Liu et al., 2024e), SlowFast-LLaVA (Xu et al., 2024b), PLLaVA (Xu et al., 2024a), and MovieChat (Song et al., 2024a). This improvement likely stems from better-captured spatiotemporal information in specially designed long video understanding methods.

Analyzing the trade-offs between design choices for modeling short and long videos is crucial. Current long video understanding methods are still suboptimal. Short video understanding methods can outperform long video methods on long video benchmarks if trained with more high-quality data or equipped with larger LLM backbones, such as LLaVA-NeXT-Video (Zhang et al., 2024e) and LLaVA-OneVision (Li et al., 2024c). Conversely, long-video-specific models do not perform well on short video benchmarks. As noted in LLaV-Next (Zhang et al., 2024e), combining different training resources has proven more effective. However, model design must balance these trade-offs: long video models require more frames but fewer visual details compared to short video models. Supporting hour-long videos necessitates powerful visual token compression, which may reduce short video understanding performance. Future versions will include a detailed examination of key aspects such as token compression strategies, handling temporal dynamics, and architectural choices for long video model design.

Model	Year	Hardware	Training	
			PT	IT
InstructBLIP (2023)	23.05	16 A100-40G	Y-N-N	Y-N-N
VideoChat (2023b)	23.05	1 A10	Y-Y-N	Y-Y-N
Video-LLaMA (2023)	23.06	–	Y-Y-N	Y-Y-N
Video-ChatGPT (2023)	23.06	8 A100-40G	N-N-N	N-Y-N
Valley (2023)	23.06	8 A100 80G	Y-Y-N	Y-Y-N
MovieChat (2024a)	23.07	–	E2E	E2E
Qwen-VL (2023b)	23.08	–	Y-N-N	Y-N-N
Chat-UniVi (2024)	23.11	–	Y-N-N	Y-Y-N
Video-LLaVA (2023)	23.11	4 A100-80G	Y-Y-N	Y-Y-N
LLaMA-VID (2023c)	23.11	8 A100	Y-Y-N	Y-Y-Y
VTimeLLM (2024a)	23.11	1 RTX-4090	Y-Y-N	N-Y-N
VideoChat2 (2024e)	23.11	–	Y-Y-N	Y-Y-N
Vista-LLaMA (2023a)	23.12	8 A100-80GB	E2E	E2E
TimeChat (2024a)	23.12	8 V100-32G	Y-Y-N	N-N-Y
VaQuitA (2023b)	23.12	8 A100-80GB	E2E	E2E
Dolphins (2023b)	23.12	4 A100	N-Y-N	Y-Y-N
Momentor (2024)	24.02	8 A100	Y-Y-N	N-Y-N
MovieLLM (2024b)	24.03	4 A100	Y-Y-N	Y-Y-Y
MA-LMM (2024)	24.04	4 A100	E2E	E2E
PLLaVA (2024a)	23.04	–	Y-N-N	Y-Y-N
LongVLM (2024)	23.04	4 A100 80G	Y-N-N	Y-Y-N
MiniGPT4-Video (2024a)	24.04	–	Y-Y-N	N-Y-N
RED-VILLM (2024b)	24.04	–	Y-N-N	Y-Y-N
ST-LLM (2024e)	24.04	8 A100	E2E	E2E
LLaVA-NeXT-Video (2024e)	24.04	–	Y-Y-N	Y-Y-N
Mantis-Idefics2 (2024)	24.05	16 A100-40G	Y-N-N	N-Y-N
VideoLLaMA 2 (2024b)	24.06	–	Y-Y-N	Y-Y-N
LongVA (2024d)	24.06	8× A100-80G	–	Y-N-N
Artemis (2024)	24.06	8 × A800	Y-Y-N	N-Y-N
VideoGPT+ (2024)	24.06	8 × A100 40G	Y-Y-N	N-Y-N
IXC-2.5 (2024c)	24.07	–	Y-Y-N	Y-Y-N
EVLM (2024b)	24.07	–	Y-Y-N	Y-Y-N
SlowFast-LLaVA (2024b)	24.07	A100-80G	–	–
LLaVA-NeXT-Interleave (2024d)	24.07	–	Y-N-N	Y-Y-N
Kangaroo (2024d)	24.08	–	Y-Y-N	Y-Y-Y
VITA (2024b)	24.08	–	Y-Y-N	Y-Y-N
LLaVA-OneVision (2024c)	24.08	–	Y-N-N	Y-Y-N
LONGVILA (2024)	24.08	256 A100 80G	Y-Y-N	Y-Y-Y
LongLLaVA (2024e)	24.09	24 A800 80G	Y-N-N	Y-Y-N
Qwen2-VL (2024a)	24.09	–	Y-N-N	Y-Y-N
Video-XL (2024)	20.09	8 A800-80G	Y-N-N	Y-Y-N
Oryx-1.5 (2024g)	24.10	64 A800-80G	Y-Y-N	Y-Y-Y
TimeMarker (2024d)	24.11	–	Y-Y-Y	Y-Y-Y
NVILA (2024f)	24.12	128 H100-80G	Y-Y-N	Y-Y-Y

Table 3: Comparison of mainstream Video-LLMs on training design. "PT" and "IT" denote the two stages of pre-training and instruction-tuning during model training. The letters "Y" (Yes) and "N" (No) indicate whether image, short-video, and long-video language datasets are used in these stages. "E2E" stands for an end-to-end training pipeline.

D More Application Scenarios on Long Video Understanding

Long video understanding with large models faces several key challenges for more long video applications. Contextual understanding is critical, as long videos require models to maintain temporal coherence and contextual awareness over extended periods (He et al., 2024). Real-time processing (Karim et al., 2024) is essential for applications like surveillance, live event analysis, and embodied AI, necessitating the development of low-latency models capable of processing video streams in real-time. Multi-modal integration is another frontier, as long videos often contain audio, text, and visual information (Zhang et al., 2023; Cheng et al., 2024b). Future models should better integrate these modalities to enhance understanding and provide a more holistic analysis of video content.

Model	LLM	Long	TGIF-QA (2-5s)	MSVD-QA (10-15s)	MSRVTT-QA (10-15s)	NeXT-QA (42.9s)	ActivityNet-QA (2mins)	GPT-based Evaluation(2mins)					
								CI	DO	CU	TU	CO	Average
InstructBLIP	Vicuna-7B	✗	–	41.8/	22.1/	–	–	–	–	–	–	–	–
Video-ChatGPT	Vicuna1.1-7B	✗	51.4/3.0	64.9/3.3	49.3/2.8	–	35.2/2.8	2.40	2.52	2.62	1.98	2.37	2.38
MA-LMM	Vicuna-7B	✓	–	60.6/	48.5/-	–	49.8/	–	–	–	–	–	–
Valley	StableVicuna-7B	✗	–	60.5/3.3	51.1/2.9	–	45.1/3.2	2.43	2.13	2.86	2.04	2.45	2.38
MovieLLM	Vicuna-7B	✓	–	63.2/3.5	52.1/3.1	–	43.3/3.3	2.64	2.61	2.92	2.03	2.43	2.53
Vista-LLaMA	Vicuna-7B	✗	–	65.3/3.6	60.5/3.3	60.7/3.4	48.3/3.3	2.44	2.31	2.64	3.18	2.26	2.57
RED-VILLM	LLaVA-7B	✗	55.9/3.1	68.9/2.8	52.4/2.9	–	39.2/3.0	2.57	2.64	3.13	2.21	2.39	2.59
Momentor	LLaMA-7B	✗	–	68.9/3.6	55.6/3.0	–	40.8/3.2	–	–	–	–	–	–
Video-LLaVA	Vicuna1.5-7B	✗	70.0/4.0	70.7/3.9	59.2/3.5	–	45.3/3.3	–	–	–	–	–	–
Artemis	Vicuna1.5-7B	✓	–	72.1/3.9	56.7/3.2	–	39.3/2.9	2.69	2.55	3.04	2.24	2.70	2.64
MovieChat	LLaMA-7B	✓	–	75.2/3.8	52.7/2.6	–	45.7/3.4	2.76	2.93	3.01	2.24	2.42	2.67
VaQuitA	LLaMA-7B	✗	–	74.6/3.7	68.6/3.3	–	48.8/3.3	–	–	–	–	–	–
RED-VILLM	QWen-VL-7B	✗	62.3/3.3	71.2/3.7	53.9/3.1	–	44.2/3.2	2.69	2.72	3.32	2.32	2.47	2.70
MiniGPT4-Video	Mistral-7B	✗	72.2/4.1	73.9/4.1	58.3/3.5	–	44.3/3.4	2.97	2.58	3.17	2.38	2.44	2.71
VTimeLLM	Vicuna-7B	✗	–	–	–	–	–	2.49	2.78	3.10	3.40	2.47	2.85
MiniGPT4-Video	LLaMA2-7B	✗	67.9/3.7	72.9/3.8	58.8/3.3	–	45.9/3.2	2.93	2.97	3.45	2.47	2.60	2.88
Chat-UniVi	Vicuna1.5-7B	✗	69.0/3.8	69.3/3.7	55.0/3.1	–	46.1/3.3	2.89	2.91	3.46	2.40	2.81	2.89
LLaMA-VID	Vicuna-7B	✓	–	69.7/3.7	57.7/3.2	–	47.4/3.3	2.96	3.00	3.53	2.46	2.51	2.89
LongVLM	Vicuna1.1-7B	✓	–	70.0/3.8	59.8/3.3	–	47.6/3.3	2.76	2.86	3.34	2.39	3.11	2.89
VideoChat2	Vicuna0-7B	✗	–	70.0/3.9	54.1/3.3	–	49.1/3.3	3.02	2.88	3.51	2.66	2.81	2.98
SlowFast-LLaVA	Vicuna1.5-7B	✓	78.7/4.2	79.1/4.1	65.8/3.6	64.2/	56.3/3.4	3.09	2.70	3.57	2.52	3.35	3.04
PLLaVA	LLaVA-Next-7B	✓	77.5/4.1	76.6/4.1	62.0/3.5	–	56.3/3.5	3.21	2.86	3.62	2.33	2.93	3.12
VideoLLaMA2-16	Mistral-7B-Instruct	✗	–	70.9/3.8	–	–	50.2/3.3	3.16	3.08	3.69	2.56	3.14	3.13
VideoLLaMA2-8	Mistral-7B-Instruct	✗	–	71.7/3.9	–	–	49.9/3.3	3.09	3.09	3.68	2.63	3.25	3.15
ST-LLM	Vicuna-7B	✓	–	74.6 /3.9	63.2/3.4	–	50.9/3.3	3.23	3.05	3.74	2.93	2.81	3.15
LongVA-32	Qwen2-7B-224K	✓	–	–	–	67.1/	/2.8	3.65	3.08	3.10	3.74	2.28	3.17
LongVA-64	Qwen2-7B-224K	✓	–	–	–	68.3/	/2.8	3.64	3.05	3.09	3.77	2.44	3.20
LLaVA-NeXT-Video	Vicuna1.5-7B	✗	–	–	–	–	53.5/3.2	3.39	3.29	3.92	2.60	3.12	3.26
LLaVA-NeXT-Interleave	Qwen1.5-7B	✗	–	–	–	78.2	55.3/3.13	3.51	3.28	3.89	2.77	3.68	3.43
LLaVA-OneVision	Qwen2-7B	✗	–	–	–	–	56.6/	–	–	–	–	–	3.49
LongVA-32-DPO	Qwen2-7B-224K	✓	–	–	–	69.3/	/2.8	4.07	3.55	3.32	4.09	2.86	3.58
LLaVA-NeXT-Video-DPO	Vicuna1.5-7B	✗	–	–	–	–	60.2/3.5	3.64	3.45	4.17	2.95	4.08	3.66
InstructBLIP	Vicuna-13B	✗	–	41.2/	24.8/	–	–	–	–	–	–	–	–
LLaMA-VID	Vicuna-13B	✓	–	70.0/3.7	58.9/3.3	–	47.5/3.3	3.07	3.05	3.60	2.58	2.63	2.99
PLLaVA	LLaVA-Next-13B	✓	77.8/4.2	75.7/4.1	63.2/3.6	–	56.3/3.6	3.27	2.99	3.66	2.47	3.09	3.27
LLaVA-NeXT-Interleave	Qwen1.5-14B	✗	–	–	–	79.1	56.2/3.19	3.65	3.37	3.98	2.74	3.67	3.48
LLaVA-NeXT-Interleave-DPO	Qwen1.5-14B	✗	–	–	–	77.9	55.0/3.13	3.99	3.61	4.24	3.19	4.12	3.83
SlowFast-LLaVA	Nous-Hermes-2-Yi-34B	✓	80.6/4.3	79.9/4.1	67.4/3.7	–	59.2/3.5	3.48	2.96	3.84	2.77	3.57	3.32
LLaVA-NeXT-Video	Nous-Hermes-2-Yi-34B	✗	–	–	–	–	58.8/3.4	3.48	3.37	3.95	2.64	3.28	3.34
PLLaVA	LLaVA-Next-34B	✓	80.6/4.3	79.9/4.2	68.7/3.8	–	60.9/3.7	3.60	3.20	3.90	2.67	3.25	3.48
LLaVA-NeXT-Video-DPO	Nous-Hermes-2-Yi-34B	✗	–	–	–	–	64.4/3.6	3.81	3.55	4.24	3.14	4.12	3.77

Table 4: Comparison of mainstream Video-LLMs on video understanding benchmarks of different lengths. Methods with ✓ in the "Long" column are designed for long videos.