LongVLM: Efficient Long Video Understanding via Large Language Models

Yuetian Weng[1], Mingfei Han[3,4], Haoyu He[1], Xiaojun Chang[2,3], and Bohan Zhuang[1†]

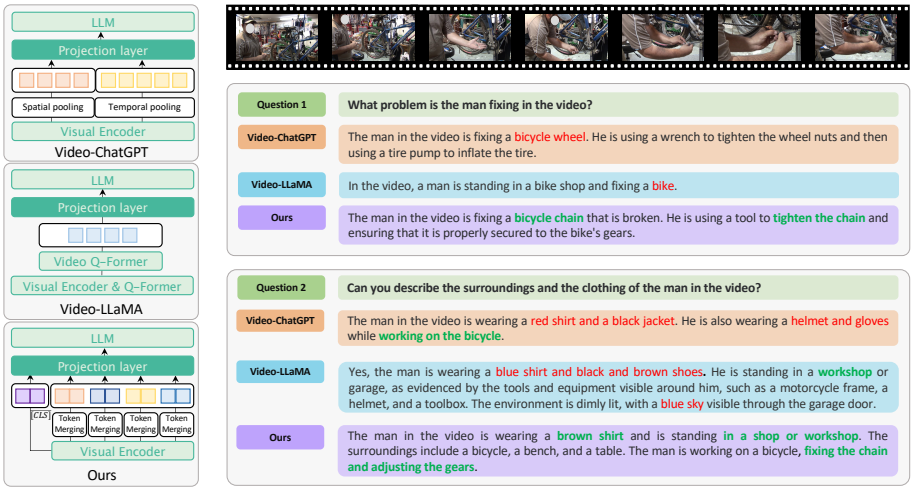
[1]ZIP Lab, Monash University, Australia [2]School of Information Science and Technology, University of Science and Technology of China
[3]Department of Computer Vision, MBZUAI [4]ReLER, AAII, UTS
{yuetian.weng,haoyu.he}@monash.edu, hmf282@gmail.com, xjchang@ustc.edu.cn, bohan.zhuang@gmail.com

Abstract. Empowered by Large Language Models (LLMs), recent advancements in Video-based LLMs (VideoLLMs) have driven progress in various video understanding tasks. These models encode video representations through pooling or query aggregation over a vast number of visual tokens, making computational and memory costs affordable. Despite successfully providing an overall comprehension of video content, existing VideoLLMs still face challenges in achieving detailed understanding due to overlooking local information in long-term videos. To tackle this challenge, we introduce LongVLM, a simple yet powerful VideoLLM for long video understanding, building upon the observation that long videos often consist of sequential key events, complex actions, and camera movements. Our approach proposes to decompose long videos into multiple short-term segments and encode local features for each segment via a hierarchical token merging module. These features are concatenated in temporal order to maintain the storyline across sequential short-term segments. Additionally, we propose to integrate global semantics into each local feature to enhance context understanding. In this way, we encode video representations that incorporate both local and global information, enabling the LLM to generate comprehensive responses for long-term videos. Experimental results on the VideoChatGPT benchmark and zero-shot video question-answering datasets demonstrate the superior capabilities of our model over the previous state-of-the-art methods. Qualitative examples show that our model produces more precise responses for long video understanding. Code is available at https://github.com/ziplab/LongVLM.

1 Introduction

Large language models (LLMs) [1,8,39,45,47,48] have revolutionized natural language understanding tasks and have demonstrated a remarkable capability to follow human instructions and intentions, emerging as a universal agent for general-purpose assistants. Drawing from the development of LLMs, Multi-modal Large

† Corresponding author.



(a) Architectures. **(b)** An example from Video-ChatGPT benchmark [38].

Fig. 1: (a) Comparison of model architectures. (b) Examples generated by different VideoLLMs. Text highlighted in bold green denotes correct content, while text in red indicates errors.

Language Models (MLLMs) [9, 16, 33, 67] have driven advancements in vision-language learning by integrating visual encoders with LLMs and finetuning on vision-language instruction pairs. However, developing Video-based Large Language Models (VideoLLMs) still poses a significant challenge due to the necessity of processing a large number of tokens for jointly modeling spatial-temporal dependencies across consecutive video frames. For instance, employing OpenAI CLIP-ViT-L/14 [42] as a visual encoder for a 100-frame video clip necessitates handling 25.6K visual tokens, leading to impractical computational costs with existing LLMs. To address this issue, recent approaches propose to extract video representation via precompression over visual tokens, utilizing pooling operation [37, 38] or query aggregation [27, 44, 63] over the video token sequence before feeding them into the LLM, as shown in Fig. 1a. While these models showcase impressive capabilities in providing a meaningful understanding of video content, they still face challenges in achieving significant advantages in fine-grained understanding of long-term videos. For example, as shown in Fig. 1b, while all models recognize the overall environment (*workshop*), the object (*bike*), and the action (*fixing*), previous methods may fail to correctly identify details such as the color of the clothes (*brown*), or the specific component being fixed (*bicycle chain*).

The main reason is that long-term videos typically involve numerous key actions, complex activities, and camera movements. Consequently, a long video can be divided into a sequence of short-term segments. For instance, in the example depicted in Fig. 1b, various short-term actions occur, *e.g.* speaking, displaying

spare components, grabbing the bicycle chain, along with the camera moving from the human to the bicycle wheel, and eventually focusing on the broken chain. Similarly, prior methods in video recognition task suggest to decompose complex activities into sequences of sub-activities [15, 22, 49]. These approaches treat the features of each short-range activity as the local information within the videos and emphasize the importance of reasoning over local features to develop a temporal-structural understanding [15, 22, 49–51, 62] within long-term videos for comprehending fine-grained information. From this perspective, existing VideoLLMs treat all visual tokens equally and aggregate them into compact representations through pooling operations [37,38] and query aggregation [27,63]. While they successfully capture the global semantic context spanning the entire long-term videos, they often overlook preserving the local information for the short-term segments and the temporal structure of different short-term components, *e.g.*, the order of events or sub-actions. However, exclusively modeling the temporal structure through the sequence of local features may still lead to inconsistent recognition across different segments and impede the overall understanding of the videos. To comprehend the content in long videos, the human visual system relies on a blend of local and global information [46]. Building on this insight, earlier approaches in video object detection [54,56] suggest integrating global semantics into local localization descriptors, motivating us to include global semantic information into the sequence of local features for enriching the context understanding for each short-term segment.

In this paper, we present LongVLM, a simple yet effective VideoLLM for efficient long video understanding, as illustrated in Fig. 2. We propose to extract video representations as sequences of short-term local features, and integrate global semantics into each short-term segment feature. Specifically, we begin by uniformly sampling a sequence of video frames from long-term videos and utilize a pretrained visual encoder, *e.g.*, CLIP-ViT-L/14 [42], to extract visual features for each individual video frame. These frame-level features include the [CLS] tokens from a range of encoder layers and the patch features from the last second layer of the visual encoder. Then, we divide the sequence of patch features along the temporal dimension, resulting in multiple short-term segments. Each segment is considered as a local unit in the videos and includes patch features of the video frames within that segment. To reduce computational costs and obtain the compact features for each segment, a token merging module is employed to aggregate these patch features for the specific segment into a condensed set of tokens. In this way, we obtain the local features for each segment. We next concatenate these features sequentially to explicitly preserve the temporal order of the short-term segments in long-term videos. Moreover, we average the [CLS] tokens from each video frame along the temporal dimension to represent the global semantic information of the entire video. To integrate the global information, we prepend the averaged [CLS] tokens before the segment-level features, and then feed them into the LLM after passing through a projection layer. Benefiting from the causal attention mechanism in the LLM, we simultaneously achieve temporal structure modeling over the sequence of short-term segments and inject global

semantics into the local features. Finally, the LLM generates responses based on the input sequence, which is composed of the obtained video representation and the designed system command with the specific user queries.

Overall, our main contributions is threefold:

- We propose LongVLM, a simple yet effective VideoLLM for efficient long-term video understanding at a fine-grained level while maintaining affordable computational cost.
- We propose decomposing long videos into short segments and extracting local features for each segment to preserve their temporal order. To compactly represent each segment, we propose a hierarchical token merging module to aggregate visual tokens. Additionally, we integrate global semantics into each segment to enhance context understanding.
- Extensive experiments on VideoChatGPT benchmark [38] and zero-shot video question-answering datasets [57, 61] demonstrate that our LongVLM surpasses the previous state-of-the-art methods by a significant margin while generating more precise and accurate response at fine-grained level for long-term videos.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs) have revolutionized natural language processing in recent years. Pretrained on large text corpora, LLMs like GPT [3], OPT [64], and LLaMA [47, 48] utilize auto-regressive Transformer models to predict subsequent tokens, showcasing remarkable adaptability and generalization. Models such as InstructGPT [40], ChatGPT [39], and GPT-4 [1] benefit from instruction-tuning techeque [53] on instructional datasets, leveraging the knowledge of pretrained LLMs and demonstrating improvements in diverse conversational interaction capabilities. This strategy is widely adopted in open-source models like Alpaca [45] and Vicuna [8], which build upon the advancements made by LLaMA [47] using specially designed instruction pairs. Drawing from the advancement of LLMs, recent Multi-modal Large Language Models (MLLMs), *e.g.*, BLIP-2 [26], Mini-GPT4 [67], LLaVA [33], LLama Adapter v2 [16], have demonstrated the feasibility of enabling visual conversation capabilities of LLMs over input images through instruction tuning on image-text instruction datasets. Our model aims to utilize existing MLLMs to develop efficient video dialogue model for long-term video understanding.

2.2 Video-based Large Language Models

Traditional video-language models [7, 12, 14, 19, 25, 29, 31, 36, 52] have advanced by using large-scale video-text pretraining followed by fine-tuning on specific downstream tasks. With the advent of LLMs, Video-based Large Language Models (VideoLLMs) explore various video-language understanding scenarios through

human-video dialogue interactions. Existing VideoLLMs typically follow a common paradigm, which involves using a pretrained visual encoder to encode visual features, a projection layer to convert visual representations into the text latent space of LLMs, and a pretrained LLM for response generation. VideoChatGPT [38] and Valley [37] rely on pooling over visual tokens to obtain compact visual representations. VideoChat [27] utilizes pretrained video foundation models [28, 52] and Q-Former from BLIP-2 [26] to aggregate video representations. Video-LLaMA [63] proposes a Video Q-Former and an Audio Q-Former, enabling multiple modalities for video comprehension, while Video-ChatCaptioner [6] employs ChatGPT [39] to summarize video descriptions in multiple rounds of interactive question-and-answer conversation. Recently, MovieChat [44] proposes an effective memory management mechanism to enable LLMs to reason over hour-long videos. Multiple video-centric instruction datasets [20, 27, 37, 38] have also been proposed to finetune VideoLLMs for better video understanding capacity. Moreover, BT-Adapter [34] proposes a temporal adapter alongside the visual encoder for post-pretraining, while Video-Teller [32] highlights the importance of modality alignment in pretraining. Overall, these VideoLLMs rely on pooling and query aggregation on the whole long videos to extract visual representation, overlooking local information for fine-grained understanding in long videos. In contrast, we propose a simple yet effective framework that is feasible for aggregating both local and global information in long-term videos and preserves fine-grained content understanding.

2.3 Long-term Video Processing

Long-term video understanding poses several challenges due to the need to exploit complicated spatial-temporal dependencies while removing temporal redundancy over extended time duration. Previous studies propose efficient architectures [10, 23], temporal pooling/aggregation [13, 43, 55, 66], dynamic clip selection [17, 18, 24] to aggregate video representation while removing redundant information in videos. Other methods in video-language understanding tasks suggest to capture event temporality, causality, and dynamics in long-term videos by designing temporal alignment modules [4, 21]. Memory mechanism is also widely adopted in video dense prediction tasks [35, 54, 56, 60, 65] to capture historical information and maintain temporal coherence, which results in more accurate and consistent prediction over time in long-term videos. Differently, we propose to aggregate both local segment-level information and global semantic information, empowering MLLMs enhanced fine-grained understanding for long-term videos.

3 Method

In Sec. 3.1, we introduce the overall architecture and generation pipeline of the proposed LongVLM. In Sec. 3.2, we introduce the process of constructing local representation via short-term feature aggregation. In Sec. 3.3, we discuss the integration of both local segment-level feature and global semantic feature.

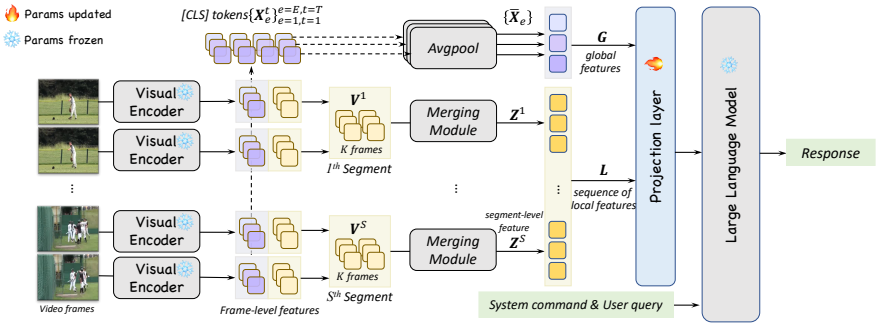


Fig. 2: Overall architecture of the proposed LongVLM. We start by uniformly sampling T frames from a video and employing a visual encoder to extract frame-level features. We divide the input video into S segments, each with K frames. To obtain compact local features, we apply a hierarchical token merging module within each segment. These segment-level features are concatenated sequentially to explicitly preserve the temporal order of multiple short-term segments in long videos. Additionally, we incorporate [CLS] tokens to aggregate global semantic features. The global features and the sequence of local features are concatenated to form the video representations. Finally, the projected visual features are combined with the tokenized system command and user queries and inputted into the LLM to generate the responses.

3.1 Overall Architecture

The overall architecture consists of three components: a visual encoder, a projection layer, and a large language model, as illustrated in Fig. 2.

Given an input video $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$, we employ a visual encoder to extract frame-level features $\{\mathbf{X}^t, \mathbf{P}^t\}_{t=1}^{T}$ for each video frame independently. Following previous methods [33, 38], we utilize the patch feature $\mathbf{P}^t \in \mathbb{R}^{N\times d}$ from the second-to-last encoder layer, where N, d are the number of patch tokens and the channel dimension of the visual encoder, respectively. Additionally, we gather the [CLS] tokens $\mathbf{X}^t \in \mathbb{R}^{E\times d}$ from E selected encoder layers for each individual video frame.

To enable fine-grained understanding in long videos, we propose to divide long videos into a sequence of short-term segments, where each segment corresponds to the local features in the long videos. Without loss of generality, the input video $\mathcal{V}$ is divided into S segments, where each segment includes K frames, *i.e.*, $K = \frac{T}{S}$. We collect patch features within the s^{th} segment, *i.e.*, $\mathbf{V}^s = \{\mathbf{P}^t\}_{t=(s-1)K}^{t=sK}$, and apply a token merging module $\mathcal{G}(\cdot)$ to aggregate $\mathbf{V}^s$ into the compact segment-level feature $\mathbf{Z}^s = \mathcal{G}(\mathbf{V}^s)$. These segment-level features are sequentially concatenated as the sequence of local representation $\mathbf{L}$, explicitly preserving temporal order of short-term segments in long videos. Furthermore, to integrate global semantic information, we propose to collect the [CLS] tokens for each frame from E encoder layers and average them in the time dimension, resulting in our global feature $\mathbf{G}$.

We forward the global features and the sequence of local features into a linear layer to obtain the projected visual features. The projected visual features are concatenated with the tokenized system command and user queries, which are inputted into LLM for response generation.

3.2 Local Feature Aggregation

After obtaining the frame-level patch feature $\{\mathbf{P}^t\}_{t=1}^{T}$ from the last second layer of visual encoder, previous methods either apply factorized spatio-temporal pooling [37,38], or utilize query aggregation [27,63] over all visual tokens, which may miss local information referring to the short-term events or actions. Nevertheless, videos have heavy spatio-temporal redundancy, which results in redundant computational costs by directly considering all the patch features as the local representation for each segment. Therefore, we propose to aggregate compact visual features within each short-term segment. Specifically, we collect the patch feature for the s^{th} segment $\mathbf{V}^s = \{\mathbf{P}^t\}_{t=(s-1)K}^{sK} \in \mathbb{R}^{KN\times d}$ and apply a hierarchical token merging module to aggregate the local feature while reducing the number of visual tokens. Inspired by ToMe [2], we resort to the bipartite soft matching method and gradually merge the visual tokens for each short-term segment. At the i^{th} step, we randomly partition the R_i tokens into two non-overlap token sets $\mathbb{P}_i$ with r_i tokens and $\mathbb{Q}_i$ with $R_i - r_i$ tokens, where initial $R_0 = KN$. Then we calculate the similarity scores between the tokens in set $\mathbb{P}_i$ and $\mathbb{Q}_i$ based on the patch features. To obtain the similarity scores, each visual token is divided into C heads along channel dimension, each with $\frac{d}{C}$ channels. The similarity score for each token pair is obtained by averaging the cosine similarity scores over all heads following Eq. 1:

$$a^{p_i q_i} = \frac{1}{C}[\sum_{c=1}^{C} cos(\mathbf{p}_c^{(p_i)}, \mathbf{p}_c^{(q_i)})], \tag{1}$$

where $p_i \in \{1, ..., r_i\}$ and $q_i \in \{1, ..., (R_i - r_i)\}$ are the indexes of patch feature $\mathbf{p}$ in set $\mathbb{P}_i$ and set $\mathbb{Q}_i$, respectively. We select the top-r_i token pairs with the highest similarity scores and merge the paired tokens by average pooling. Finally, the remaining tokens in the two sets are concatenated back together, resulting in $R_i - r_i$ tokens after the i^{th} merging step. We iteratively merge the tokens within each short-term segment, until the number of visual tokens reaches M, where $M << K \times N$. The compact local feature for the s^{th} segment is denoted as $\mathbf{Z}^s = \{\mathbf{z}_m\}_{m=1}^{M} \in \mathbb{R}^{M\times d}$. These segment-level features are concatenated sequentially as the sequence of local features $\mathbf{L} = \{\mathbf{Z}^s\}_{s=1}^{S} = [\mathbf{z}_1^1, ..., \mathbf{z}_M^1, ..., \mathbf{z}_1^S, ..., \mathbf{z}_M^S] \in \mathbb{R}^{MS\times d}$. Thanks to the positional encoding in the LLM, the sequence of local representation $\mathbf{L}$ explicitly preserves the order of short-term segments in long-term videos, enabling LLMs to be aware of the temporal structure of multiple event occurrences in long videos. By utilizing the token merging module, we efficiently encode compact local features for each segment while eliminating redundancy in the visual token sequence.

3.3 Global Semantics Integration

The local features provide fine-grained information about different events or actions in the generation process of VideoLLMs, enhancing the detailed understanding capabilities of the LongVLM. However, the local features for each segment may be insufficient for the model to reason the relationship between different segments and generate reasonable response over the entire videos. Therefore, we additionally introduce global semantic features to enrich the local features with contextual information. Specifically, we collect the [CLS] tokens of each video frame from E encoder layers, *i.e.*, $\{\mathbf{x}_e^t\}_{e=1,t=1}^{e=E,t=T}$, and then average the [CLS] tokens along temporal dimension, resulting in $\bar{\mathbf{X}}_e = \text{AvgPool}(\{\mathbf{x}_e^t\}_{t=1}^{T}) \in \mathbb{R}^d, e \in [1, ..., E]$. By default, E can be the number of layers in the visual encoder. However, previous studies demonstrate different properties between intermediate features from the shallow layers and deeper layers in ViT models [30,41], and showcase the deeper layers tend to aggregate global semantics. Thus, we concatenate the E-scale features along sequentially, resulting in the global semantic feature for the entire video, *i.e.*, $\mathbf{G} = \{\bar{\mathbf{X}}_e\}_{e=1}^{E} \in \mathbb{R}^{E \times d}$.

Following the previous studies, a projection layer converts the visual features into the language space, and then the visual features are concatenated with the instruction as the input of LLM. By utilizing the attention mechanism in the LLM, we can easily enable each token in the local feature to attend to the global semantic feature, thereby achieving straightforward injection of global semantics into the local feature.

Remark. To address the risk of overlooking detailed understanding in long-term videos, we propose to divide long videos into multiple short-term segments and aggregate local spatial-temporal representation for each segment and preserving the temporal structure over the sequence of local feature vectors. Moreover, we enrich the local features with context information for better response generation by integrating global semantic information into short-term features.

4 Experiments

4.1 Experimental Settings

Datasets and evaluation metrics. We conduct quantitative evaluations of our model using the VideoChatGPT benchmark [38] to assess its performance in generating text from videos. The benchmark comprises 500 videos sampled from ActivityNet-v1.3 dataset [11], with 2000, 2000, 2000, 500, and 1000 questions in terms of five evaluation aspects: Correctness Information(CI), Detail Orientation(DO), Contextual Understanding(CU), Temporal Understanding(TU) and Consistency(C). Additionally, we evaluate the model on the zero-shot question-answering task using the ANET-QA [61] dataset, which contains 8000 QA pairs for 800 videos sampled from ActivityNet-v1.3 dataset [11]. The videos range from several seconds to minutes long and cover a wide range of daily human activities. We also utilize MSRVTT-QA [57,58](72821 QA pairs for 2990 videos)

Table 1: Comparison with state-of-the-art methods on video conversation benchmark [38] in terms of five evaluation aspects and the average scores across all aspects (Mean). We also report the dataset scale used for finetuning the model.

Method	Data Source	CI	DO	CU	TU	C	Mean
VideoChat [27]	10M	2.25	2.50	2.54	1.98	1.84	2.22
LLaMA Adapter v2 [16]	700K	2.03	2.32	2.30	1.98	2.15	2.16
Video LLaMA [63]	10M	1.96	2.18	2.16	1.82	1.79	1.98
Video-ChatGPT [38]	100K	2.50	2.57	2.69	2.16	2.20	2.42
Valley [37]	234k	2.43	2.13	2.86	2.04	2.45	2.38
BT-Adapter [34]	10M	2.16	2.46	2.89	2.13	2.20	2.37
BT-Adapter [34]	10M+100K	2.68	2.69	3.27	2.34	2.46	2.69
Ours	100K	**2.76**	**2.86**	**3.34**	**2.39**	**3.11**	**2.89**

Table 2: Comparison with state-of-the-art methods on three zero-shot question answering datasets. We report the Accuracy (Acc.) and Score for the generated answer for each question, and the dataset scale used for finetuning the model.

Method	Data source	ANET-QA		MSRVTT-QA		MSVD-QA	
		Acc.	Score	Acc.	Score	Acc.	Score
FrozenBiLM [59]	10M	24.7	-	16.8	-	32.2	-
VideoChat [27]	10M	26.5	2.2	45.0	2.5	56.3	2.8
LLaMA Adapter v2 [16]	700K	34.2	2.7	43.8	2.7	54.9	3.1
Video LLaMA [63]	10M	12.4	1.1	29.6	1.8	51.6	2.5
Video-ChatGPT [38]	100K	35.2	2.7	49.3	2.8	64.9	3.3
Valley [37]	234K	45.1	3.2	51.1	2.9	60.5	3.3
BT-Adapter [34]	10M+100K	45.7	3.2	57.0	3.2	67.5	3.7
Ours	100K	**47.6**	**3.3**	**59.8**	**3.3**	**70.0**	**3.8**

and MSVD-QA [5,57](13157 QA pairs for 520 videos) to evaluate the model performance, derived from publicly available video captioning, MSRVTT [58], and MRVDC [5], respectively. Following the evaluation protocol outlined in Video-ChatGPT [38], we employ ChatGPT [39] for response evaluation and report the generation quality scores on VideoChatGPT benchmark and the answer accuracy and quality scores of models on zero-shot video QA tasks.

Implementation details. We employ CLIP-ViT-L/14 [42] as the visual encoder and Vicuna-7B-v1.1 [8] as the LLM. We initialize them with the pretrained weights in LLaVA-7B-v1.1 [33]. We finetune the model on the Video-ChatGPT-100K instruction dataset [38] for 3 epochs, with a learning rate of $2e^{-5}$ and a batch size of 32. We only finetune the linear projection layer to align the visual features into the input space of the LLM, keeping both the visual encoder and LLM frozen. It takes three hours to train three epochs on 4 A100 80GB GPUs. During training and inference, we sample $T = 100$ video frames for each video,

and resize the frames to 224×224 resolutions. We set $C = 16$, the same to the number of heads of CLIP-ViT/L-14. We set $S = 10$ for each video, and the number of tokens in each segment-level feature is $M = 30$. We collect the [CLS] tokens from the last five encoder layers and average them along the temporal axis, resulting in $E = 5$ tokens as the global semantic features. Therefore, the length of visual tokens for a video sequence is $M \times S + E = 305$.

4.2 Main Results

Results on the video-based generation benchmark. In Tab. 1, we present a comprehensive evaluation of our LongVLM against state-of-the-art models on the video-based generation benchmark [38]. Our LongVLM outperforms all other models across all the evaluation aspects. Particularly noteworthy is its significant advantage in Detail Orientation (DO) and Consistency (C), showing improvements of +0.17 and +0.65, respectively, over BT-Adapter [34]. These results underscore superior capability of LongVLM in fine-grained video understanding and robust generation performance.

Results on zero-shot video question-answering. In Tab. 2, we compare the performance of LongVLM against various existing methods on three zero-shot video QA datasets: ANET-QA [11,61], MSRVTT-QA [57,58] and MSVD-QA [5, 57]. Our model achieves the highest accuracy of 47.6%, 59.8%, and 70.0% on the three QA datasets, surpassing the previous SOTA approach BT-Adapter [34] by 1.9%, 2.8% and 2.5%, respectively. Furthermore, we achieve the highest score in terms of generation quality over the three datasets.

4.3 Ablation Study

Effects of local feature aggregation. As discussed in Sec. 1, pooling operations or query aggregation might overlook local information in achieving fine-grained understanding in long-term videos. To this end, we introduce short-term segment-level features to retain local information and temporal structure within long-term videos. The first two rows in Tab. 3 present the effects for the design of using local features as the visual representations for videos. We compare the token merging module with a local pooling operation. Specifically, we apply a 3D average pooling operation within each short-term segment, using a kernel size and stride of (5, 4, 4) to the temporal, height, and width dimensions, respectively. The proposed hierarchical merging module achieves higher scores over spatial-temporal pooling operation. This could be attributed to the dynamic aggregation mechanism via the token similarity in the merging module, while averaging pooling statically aggregates visual tokens within each small 3D window. Additionally, we observe that aggregating local features for short-term segments either improves or maintains comparable performance across all evaluation metrics compared to the SOTA models which extract global semantics only, highlighting the significance of preserving local features for short-term segments in long-term video understanding.

Table 3: Ablation of the local and global aggregation design. Pooling: Using 3D average pooling operation to obtain local features; Merging: Using the proposed hierarchical token merging to obtain local features; [**L**, **G**]: Concatenating local feature then global feature; [**G**, **L**]: Concatenating global feature then local feature.

Variants	Local	Global	CI	DO	CU	TU	C	Mean
Pooling	✓	✗	2.53	2.64	3.13	2.29	2.61	2.64
Merging	✓	✗	2.62	2.74	3.15	2.23	2.86	2.72
[**L**, **G**]	✓	✓	2.69	2.81	3.31	2.31	2.99	2.82
[**G**, **L**]	✓	✓	**2.76**	**2.86**	**3.34**	**2.39**	**3.11**	**2.89**

Table 4: Effect of M. We evaluate the model performance by varying M from $\{10, 20, 30, 40\}$, while keeping the token length of global semantic features at $E = 5$.

M	CI	DO	CU	TU	C	Mean
10	2.61	2.72	3.22	2.26	2.78	2.72
20	2.72	**2.86**	3.34	2.34	2.96	2.84
30	**2.76**	**2.86**	3.34	**2.39**	**3.11**	**2.89**
40	2.74	2.84	**3.39**	2.34	3.06	2.87

(a) Video-ChatGPT Benchmark.

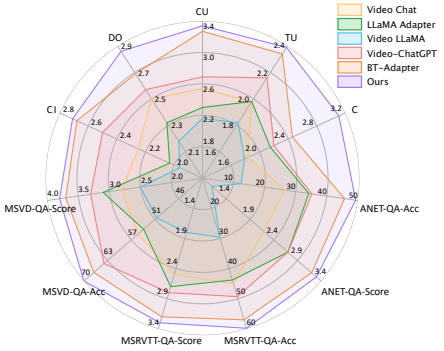
M	Accuracy	Score	Memory(G)
10	44.6	3.2	**14.65**
20	45.7	**3.4**	14.74
30	**47.6**	3.3	14.86
40	46.0	3.3	14.96

(b) ANET-QA.

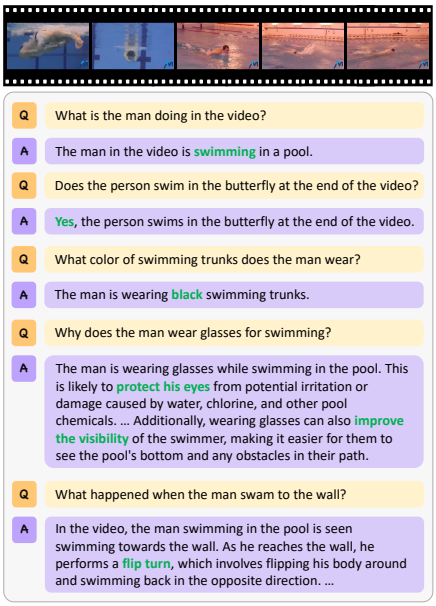
Table 5: Effect of the number of selected encoder layers. We evaluate the performance of model varying E in {1, 5, 10, 15, 20, 24}, keeping $M = 30, K = 10$.

E	CI	DO	CU	TU	C	Mean
1	2.74	2.85	3.23	2.32	3.04	2.83
5	2.76	**2.86**	**3.34**	**2.39**	**3.11**	**2.89**
10	**2.78**	**2.86**	3.24	2.30	3.04	2.84
15	2.72	2.82	3.16	2.16	2.97	2.77
20	2.63	2.77	3.11	2.28	2.93	2.74
24	2.65	2.75	3.08	2.22	2.81	2.70

Effects of global semantics integration. Inspired by the human visual system that using a combination of local and global information for recognizing video content [46], we propose to enhance visual representation by injecting global semantic features into local features. The last two rows in Tab. 3 demonstrate the effects of integrating global semantics. Compared to the first two rows, introducing global semantic features significantly enhances performance compared to models using local features only across all evaluation aspects. The notable improvements in Contextual Understanding (CU) and Consistency (C) underscore the significance of integrating global semantic information with local



(a) Quantitative results on Video-ChatGPT-100K [38] benchmark and the task of zero-shot question answering on ANet-QA [61], MSRVTT-QA [57, 58] and MSVD-QA [5, 57]. Our model delivers the best performance on multiple evaluation aspects, compared with the state-of-the-art video dialogue models: Video Chat [27], LLaMA Adapter [16], Video LLaMA [63], Video-ChatGPT [38], and BT-Adapter [34]. Evaluation metrics and comparison details are given in Section 4.2.



(b) An example of zero-shot question answering. Video duration is 3 minutes and 46 seconds.

Fig. 3: Quantitative results and qualitative examples of our LongVLM.

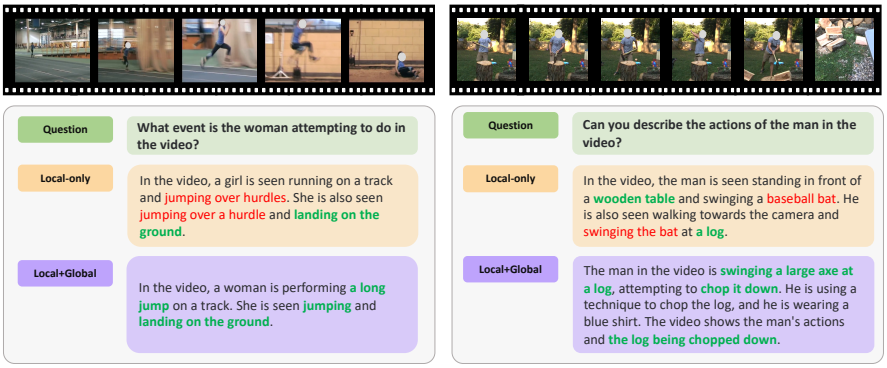


Fig. 4: Two examples from Video-ChatGPT benchmark [38]. Text highlighted in bold green denotes correct content, while text in red indicates errors.

short-term features. Moreover, concatenating global features before local features yields better results than the opposite concatenation order. This arrangement allows each the local feature to access the global semantic information across the entire video by leveraging the causal attention mechanism in the LLM. Consequently, this design enriches the contextual information of the local features and enhances the response consistency of the model.

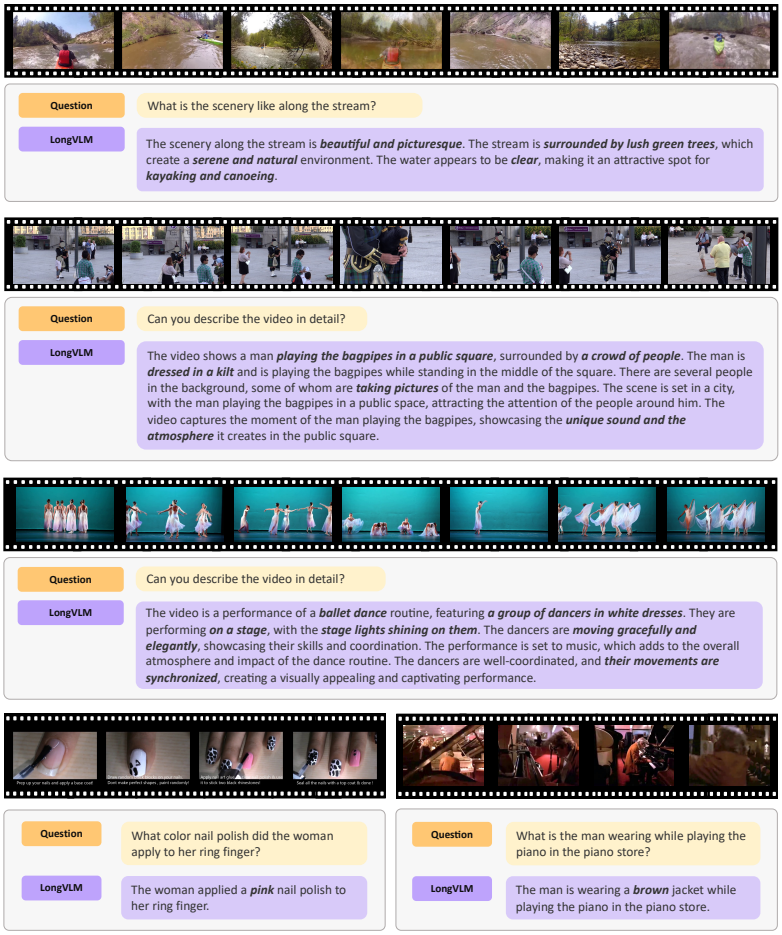


Fig. 5: More generative examples from the Video-ChatGPT benchmark [38] of the proposed LongVLM. Text in bold denotes the correct content. The LongVLM is able to capture the detailed information videos.

Effects of M. We report the model performance on VideoChatGPT benchmark and ANET-QA task on the selection of M, *i.e.*, $M = \{10, 20, 30, 40\}$, keeping the same number of global semantic tokens in Tab. 4. For ANET-QA, we also report the averaging GPU memory usage for generating each answer. In general, the token length involves a trade-off between memory costs and performance. A shorter token sequence reduces computational costs for generating a single new token using LLM, thereby lowering memory costs for generating responses to individual user queries. However, it may also lead to insufficient visual information for generating accurate responses. The performance of our model is beneficial from the suitable length of visual tokens. Increasing M from 10 to 40 results in

a significant improvement in terms of most evaluation aspects, while the setting of $M = 40$ leads to neglecting improvement but requires more memory cost compared to $M = 30$. Therefore, we choose $M = 30$ for our model.

Effects of E. We evaluate the model performance on the VideoChatGPT benchmark with varying E by selecting the [CLS] tokens from the last $1, 5, 10, 15, 20, 24$ visual encoder layers, while maintaining the same M for local features. As depicted in Tab. 5, increasing the number of global semantic tokens from 1 to 5 improves the generation quality scores in terms of all evaluation aspects. However, increasing E from 5 to 24 leads to degraded performance, possibly because the [CLS] tokens from earlier layers carry less semantic information for the model. Therefore, we choose $E = 5$ in our model.

4.4 Qualitative Results

As illustrated in Sec. 1, Fig. 1b demonstrates the advancement of our model in terms of fine-grained understanding in long-term videos. Despite taking the same number of video frames as input, our model excels in capturing detailed information within the videos, discerning nuances like *fixing chain* rather than *fixing wheel*. In comparison, Video-ChatGPT [38] can describe the overall video content but may inaccurately recognize detailed information. For instance, it might identify objects such as *helmets* and *gloves* in the scene but erroneously recognize the location for these objects. This emphasize the importance of decomposing long videos into multiple short-term segments and aggregate local features to achieve fine-grained understanding in videos. The examples depicted in Fig. 4 ablate the effectiveness of integrating global semantic information into local short-term features. With global semantics integration, the model is able to recognize the actions (*long jump*) and objects (*axe*) compared to the variant that using local features only. We provide more generated examples in Fig. 3b and Fig. 5 from ANET-QA and Video-ChatGPT benchmark, respectively, which showcase the precise description of the video content generated by our LongVLM.

5 Conclusion

In this work, we have introduced LongVLM, an effective and efficient VideoLLM designed for long-term video understanding. By extracting local features for short-term segments, we efficiently model local dependencies while preserving the temporal structure of sequential events in long-term video sequences. Through the integration of local and global information, LongVLM captures detailed information and provides consistent and accurate responses for long-term videos.

Limitations and Further work. While we introduce a novel video conversation model for fine-grained long-term video comprehension, our framework is specifically designed for video-to-text generation scenarios. Future work may include extending our framework into video-centric multimodal generation tasks and training the model on large-scale, extended-duration videos for long-context understanding.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: ICLR (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NIPS **33**, 1877–1901 (2020)
4. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the" video" in video-language understanding. In: CVPR. pp. 2917–2927 (2022)
5. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: ACL. Portland, OR (June 2011)
6. Chen, J., Zhu, D., Haydarov, K., Li, X., Elhoseiny, M.: Video chatcaptioner: Towards the enriched spatiotemporal descriptions. arXiv preprint arXiv:2304.04227 (2023)
7. Cheng, F., Wang, X., Lei, J., Crandall, D., Bansal, M., Bertasius, G.: Vindlu: A recipe for effective video-and-language pretraining. In: CVPR. pp. 10739–10750 (2023)
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org (2023)
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
10. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. pp. 2625–2634 (2015)
11. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970 (2015)
12. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
13. Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T.: Rank pooling for action recognition. PAMI **39**(4), 773–787 (2016)
14. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681 (2021)
15. Gaidon, A., Harchaoui, Z., Schmid, C.: Temporal localization of actions with actoms. TPAMI **35**(11), 2782–2795 (2013)
16. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
17. Ghodrati, A., Bejnordi, B.E., Habibian, A.: Frameexit: Conditional early exiting for efficient video recognition. In: CVPR. pp. 15608–15618 (2021)
18. Gowda, S.N., Rohrbach, M., Sevilla-Lara, L.: Smart frame selection for action recognition. In: AAAI. vol. 35, pp. 1451–1459 (2021)
19. Han, M., Wang, Y., Li, Z., Yao, L., Chang, X., Qiao, Y.: Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In: ICCV. pp. 13414–13423 (2023)

20. Han, M., Yang, L., Chang, X., Wang, H.: Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. arXiv preprint arXiv:2311.17043 (2023)
21. Han, T., Xie, W., Zisserman, A.: Temporal alignment networks for long-term video. In: CVPR. pp. 2906–2916 (2022)
22. Hussein, N., Gavves, E., Smeulders, A.W.: Timeception for complex action recognition. In: CVPR. pp. 254–263 (2019)
23. Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., Gong, B.: Movinets: Mobile video networks for efficient video recognition. In: CVPR. pp. 16020–16030 (2021)
24. Korbar, B., Tran, D., Torresani, L.: Scsampler: Sampling salient clips from video for efficient action recognition. In: ICCV. pp. 6232–6242 (2019)
25. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: CVPR. pp. 7331–7341 (2021)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
27. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
28. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Unlocking the potential of image vits for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1632–1643 (2023)
29. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. arXiv preprint arXiv:2303.16058 (2023)
30. Li, K., Wang, Y., Peng, G., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatial-temporal representation learning. In: ICLR (2021)
31. Li, L., Gan, Z., Lin, K., Lin, C.C., Liu, Z., Liu, C., Wang, L.: Lavender: Unifying video-language understanding as masked language modeling. In: CVPR. pp. 23119–23129 (2023)
32. Liu, H., Fan, Q., Liu, T., Yang, L., Tao, Y., Huang, H., He, R., Yang, H.: Video-teller: Enhancing cross-modal generation with fusion and decoupling. arXiv preprint arXiv:2310.04991 (2023)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
34. Liu, R., Li, C., Ge, Y., Shan, Y., Li, T.H., Li, G.: One for all: Video conversation is feasible without video instruction tuning. arXiv preprint arXiv:2309.15785 (2023)
35. Lu, Y., Lu, C., Tang, C.K.: Online video object detection using association lstm. In: ICCV. pp. 2344–2352 (2017)
36. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
37. Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207 (2023)
38. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)

39. OpenAI: Chatgpt. https://openai.com/blog/chatgpt/ (2023)
40. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NIPS **35**, 27730–27744 (2022)
41. Pan, Z., Zhuang, B., He, H., Liu, J., Cai, J.: Less is more: Pay less attention in vision transformers. In: AAAI. vol. 36, pp. 2035–2043 (2022)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
43. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: ECCV. pp. 154–171. Springer (2020)
44. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR. pp. 18221–18232 (2024)
45. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023)
46. Tian, Y., Yang, M., Zhang, L., Zhang, Z., Liu, Y., Xie, X., Que, X., Wang, W.: View while moving: Efficient video recognition in long-untrimmed videos. In: ACMMM. pp. 173–183 (2023)
47. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
48. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
49. Wang, L., Qiao, Y., Tang, X.: Latent hierarchical model of temporal structure for complex activity classification. TIP **23**(2), 810–822 (2013)
50. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36. Springer (2016)
51. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. TPAMI **41**(11), 2740–2755 (2018)
52. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
53. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
54. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: CVPR. pp. 284–293 (2019)
55. Wu, C.Y., Krahenbuhl, P.: Towards long-form video understanding. In: CVPR. pp. 1884–1894 (2021)
56. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR. pp. 13587–13597 (2022)
57. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACMMM. pp. 1645–1653 (2017)
58. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR. pp. 5288–5296 (2016)

59. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. NeurIPS **35**, 124–141 (2022)
60. Yang, T., Chan, A.B.: Learning dynamic memory networks for object tracking. In: ECCV. pp. 152–167 (2018)
61. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. vol. 33, pp. 9127–9134 (2019)
62. Zhang, C., Gupta, A., Zisserman, A.: Temporal query networks for fine-grained video understanding. In: CVPR. pp. 4486–4496 (2021)
63. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
64. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
65. Zhao, Y., Luo, C., Tang, C., Chen, D., Codella, N., Zha, Z.J.: Streaming video model. In: CVPR. pp. 14602–14612 (2023)
66. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV. pp. 803–818 (2018)
67. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)