MovieChat: From Dense Token to Sparse Memory for Long Video Understanding

Enxin Song[1♠*] Wenhao Chai[2♠♡] Guanhong Wang[1♠]
Yucheng Zhang[1♢] Haoyang Zhou[1♢] Feiyang Wu[1♢] Haozhe Chi[1]
Xun Guo[3] Tian Ye[4] Yanting Zhang[5] Yan Lu[3] Jenq-Neng Hwang[2] Gaoang Wang[1,6,✉]

[1] Zhejiang University [2] University of Washington [3] Microsoft Research Asia
[4] Hong Kong University of Science and Technology (GZ) [5] Donghua University
[6] Shanghai Artificial Intelligence Laboratory

Abstract

Recently, integrating video foundation models and large language models to build a video understanding system can overcome the limitations of specific pre-defined vision tasks. Yet, existing systems can only handle videos with very few frames. For long videos, the computation complexity, memory cost, and long-term temporal connection impose additional challenges. Taking advantage of the Atkinson-Shiffrin memory model, with tokens in Transformers being employed as the carriers of memory in combination with our specially designed memory mechanism, we propose the MovieChat to overcome these challenges. MovieChat achieves state-of-the-art performance in long video understanding, along with the released MovieChat-1K benchmark with 1K long video and 14K manual annotations for validation of the effectiveness of our method. The code, models and data can be found in https://rese1f.github.io/MovieChat.

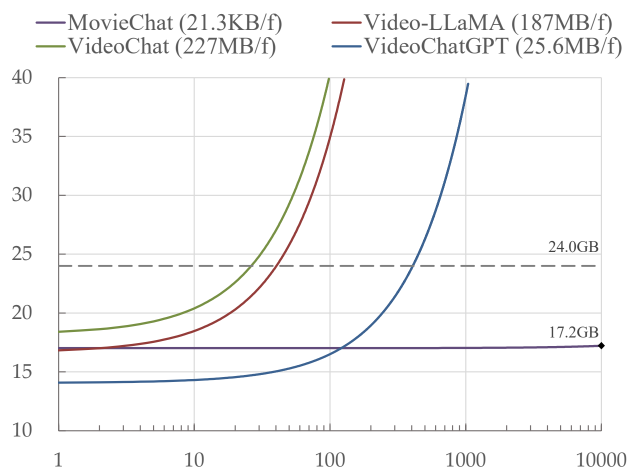


Figure 1. VRAM cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison. We test the visual-only inference of all methods at a resolution of 224×224 without frame sampling. While the previous method can only support around 100 frames of inference, MovieChat can handle videos with >10K frames on a 24GB graphics card. MovieChat has a 10000× advantage over other methods in terms of the average increase in VRAM cost per frame (21.3KB to $\sim$ 200MB per frame).

1. Introduction

Recent advances in Large Language Models (LLMs) [12, 18, 44, 59, 61] acheive great success in Natural Language Processing (NLP). It is a natural progression to introduce multi-modality [15] into LLMs and turn it into Multi-modal Large Language Models (MLLMs), which is able to conduct multimodal rationalization and understanding. MLLMs have shown incredible emergent capabilities in various multimodal tasks such as perception (*e.g.*, count, OCR) [1, 30, 31, 40, 65, 81], commonsense reasoning [23, 25, 30, 31, 33, 40, 58, 81], and code reasoning [19, 22, 23, 36, 38, 74], resulting in a potential path to Artificial General Intelligence (AGI). Compared to LLMs and other task-specific models, MLLMs provide a more human-like interpretation of the scenarios, a user-friendly interface for interaction, and a broader range of capabilities.

Existing vision-centric MLLMs follow the paradigm that utilizing pre-trained LLMs and visual encoder with additional learnable modules (Q-former [19, 31, 33, 78] or simple projection layer [20, 36, 40, 58]). In video field, some previous works [40, 78] follow this paradigm to build video MLLMs, while works in the other paradigm [34, 63] combine existing visual perception tools and LLMs through Ap-

*♠ Equal contribution, ♡ Project lead, ♢ Data collection, ✉ Corresponding Author.

plication Programming Interface (API) to build a system without training. Yet, previously, there is no exploration of a model or system based on long videos (over one minute), and there is also a lack of a standardized benchmark to evaluate the capabilities of these systems.

In this paper, we present MovieChat, a novel framework that integrates vision models and LLMs to conduct long video understanding tasks. We claim that the computation complexity, memory cost, and long-term temporal connection are the main challenges for long video understanding. Atkinson-Shiffrin memory model [5] proposes that short-term memory functions as a buffer of long-term memory, serving as a processor for the encoding of information into long-term memory. Inspired by this, we propose a memory mechanism to deal with long video understanding tasks, which includes a rapidly updated short-term memory and a compact thus sustained long-term memory. We use a sliding window approach to extract video features and represent them in token form, which are then sequentially fed into the short-term memory frame by frame. The short-term memory has a fixed length, and when it reaches its set limit, the earliest tokens are popped and consolidated into the long-term memory. After passing through a projection layer, the video representation is inputted into a large language model for interaction with the user. As shown in Fig. 1, our proposed MovieChat mechanism outperforms other existing methods in terms of Video Random Access Memory (VRAM) cost. We also release a new benchmark, MovieChat-1K, with 1K long videos and 13K manual question-answering pairs for validation of the effectiveness of our proposed MovieChat.

The contributions of this work are summarized as:

- We present MovieChat, a novel framework that integrates vision models and LLMs, which is the first to support long video (>10K frames) understanding tasks.
- We propose an effective memory management mechanism to reduce the computation complexity and memory cost, while enhancing the long-term connection.
- We release the first long video understanding benchmark, MovieChat-1K, with manual annotations and conduct extensive quantitative evaluation and case studies to evaluate the comparable performance of both understanding capability and inference cost.

2. Related Works

2.1. Multi-modal Large Language Models

LLMs [12, 18, 44, 59, 61, 62] have achieved great success in natural language processing (NLP) tasks recently. Many works try to build MLLMs [1, 25, 30, 31, 74, 81] by combining models of other modalities. Flamingo [1] bridges powerful pre-trained vision-only and language-only models and achieves state-of-the-art performance with few-shot learning. MiniGPT-4 [81] aligns a frozen visual encoder with a frozen LLM, Vicuna [18], using just one projection layer to realize the system. VideoChat [34] integrates video foundation models and LLMs via a learnable neural interface, excelling in spatiotemporal reasoning, event localization, and causal relationship inference. Video-LLaMA [78] further leverages pre-trained models ImageBind [24] and LLaMA [61], bootstraping cross-modal training in videos following BLIP-2. Yet, these methods fail to handle long video understanding because of high computation complexity, large memory cost, and weak long-term temporal connection. Therefore, our main effort is to introduce an effective memory mechanism to overcome these challenges.

2.2. Long Video Understanding

Understanding long videos is a challenging task in computer vision. Prior arts use 3D CNN for long-term feature bank [66], object/human-centric motion [47, 67], or other forms [51, 68] as video representations. Building long-form video understanding datasets is challenging and rarely explored. [54] captures large scale data from Kinetics-400 [14], but only for generic event boundary detection tasks. [55] creates a language grounding benchmark from audio descriptions of movies, but it lacks long-term understanding evaluation. There are also several datasets of video-caption/description pairs among various domains, such as cooking (*e.g.*, MPII Cooking [48–50] and TACoS [45, 46]), instruction (*e.g.*, HowTo100M [42] and HiREST [76]), Ego [41], and movie (*e.g.*, MovieQA [60] and MovieNet [28]) from different sources such as YouTube [16, 42, 77], Twitter [6–9], and Internet [10]. Yet, those datasets lack diverse and fine-grained dense captioning for long videos.

2.3. Memory Models in Vision Tasks

There are some prior works exploring memory models [56] in various vision tasks in videos, such as video object segmentation (VOS) [17, 27, 52, 53], multi-object tracking (MOT) [2, 13, 26, 69], visual object tracking (VOT) [35, 39, 73, 80], and action understanding [64]. MeMOT [13] builds a large spatiotemporal memory that stores the past observations of the tracked objects. XMem [17] develops an architecture that incorporates multiple independent yet deeply-connected feature memory storage to handle long videos with thousands of frames. We learn from the experience of those prior arts and further adopt an effective memory mechanism in combination with LLMs. Our method focuses on reducing the redundancy of visual tokens in the video and building a memory mechanism to pass the information among a large temporal range.

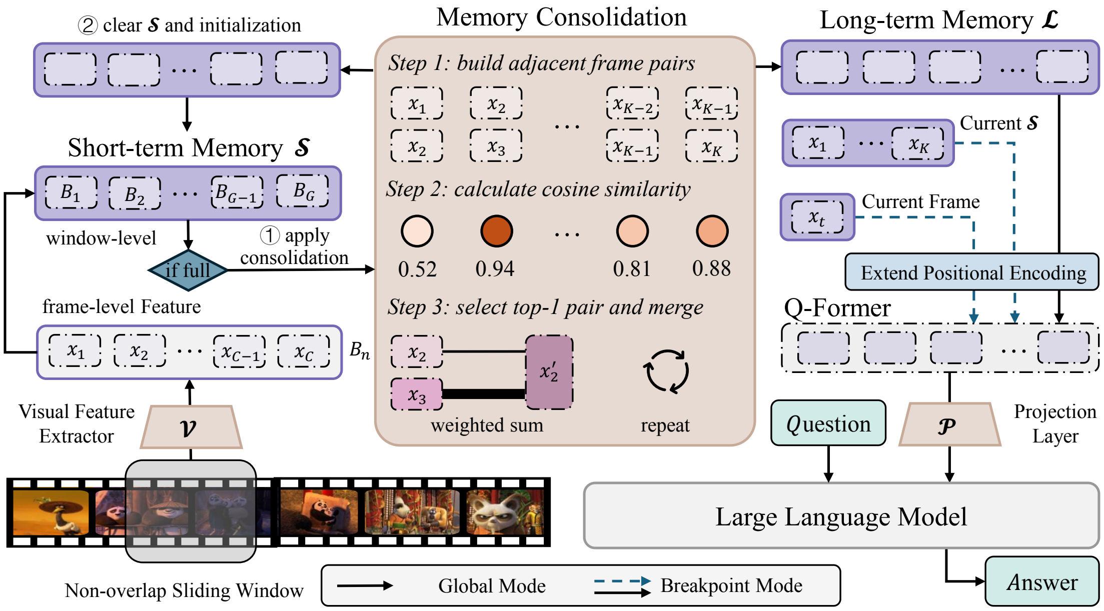


Figure 2. **Illustration of MovieChat.** MovieChat extracts video features with a sliding window and represents them in token form, which are then sequentially fed into the short-term memory frame by frame. When the fixed-length short-term memory reaches its preset limit, the earliest tokens are popped and consolidated into the long-term memory. MovieChat incorporates two distinct inference modes: the global mode, which exclusively utilizes the long-term memory, and the breakpoint mode, which additionally incorporates the current short-term memory as part of the video representation. The breakpoint mode allows for understanding the video at a specific moment in time. After passing through a projection layer, the video representation is inputted into a large language model for interaction with the user.

3. MovieChat

3.1. Overview

Our proposed method, MovieChat, comprises several key components, including the frame-wise visual feature extractor, the short-term and long-term memory modules, the video projection layer, and the LLM, as illustrated in Fig. 2. MovieChat is designed for ultra-long videos (>10K frames) understanding through interactive dialogue with the user. To address the impractical storage demands of concurrently storing a vast number of frames in both GPU memory and RAM, we employ a sliding window approach to efficiently process the video. The short-term memory module embeds dense tokens with sliding window and the long-term memory module periodically updates. MovieChat supports two inference modes: Breakpoint mode is used to understand a specific moment in the video, providing insights and answers based on that particular frame or scene; Global mode, on the other hand, is employed to comprehend the entire video as a whole, enabling a comprehensive understanding of the overall content and context.

3.2. Visual Feature Extraction

For visual feature extraction, instead of utilizing video-based foundational models such as ViViT [4] or Video-Swin [37], we simply use an image-based model to get frame-wise feature in the form of tokens. To be specific, we utilize pre-trained models as our visual feature extractor, including the ViT-G/14 from EVA-CLIP [21] and the Q-former from BLIP-2 [32]. This is mainly because 1) there is few video foundation model that makes good alignment with text, and 2) our proposed memory mechanism can effectively capture temporal features. Given a raw video, the visual input $\mathbf{v} \in \mathbb{Z}^{T\times 3\times H\times W}$ is a sequence of T RGB frames of size $H \times W$ sampled from the video. The visual features are extracted in a sliding window manner, which could be formulated as

$$B_n = \{\mathbf{x}_i = \mathcal{V}(\mathbf{v}_i) \mid \forall i = 1, ..., C\}, n = 1, ..., \lceil \frac{T}{C} \rceil, \quad (1)$$

where B_n is the n-th video clip feature within the sliding window spanning C frames. $\mathcal{V}(\cdot)$ is the visual feature extractor, taking as input a single frame $\mathbf{v}_i \in \mathbb{Z}^{3\times H\times W}$. $\mathbf{x}_i \in \mathbb{R}^{N\times D}$ denotes N extracted visual tokens with respect to each frame, and D is the feature dimension of each token.

3.3. Short-term Memory

Short-term memory stores the frame tokens in a temporary fixed-length buffer. The previously extracted visual features by sliding window G times without further processing are used to construct short-term memory, which can be formulated by:

$$\mathcal{S} = \bigcup_n B_n = \{\mathbf{x}_i \mid \forall i = 1, ..., K\}, n = 1, .., G, \quad (2)$$

Algorithm 1 Memory consolidation

Require: $\mathcal{S}$ ▷ short-term memory
while $len(\mathcal{S}) > R_L$ **do** ▷ iterative merge
  for $\mathbf{x}_i$ in $\mathcal{S}$ **do**
    $s \leftarrow sim(\mathbf{x}_i, \mathbf{x}_{i+1})$ ▷ tokens similarity
  end for
  $m \leftarrow max(s)$ ▷ the maximum value index
  $\mathbf{x}_m \leftarrow merge(\mathbf{x}_m, \mathbf{x}_{m+1})$ ▷ merge
  del $\mathbf{x}_{m+1}$
end while

where $\mathcal{S}$ is short-term memory, and K is equal to $C \times G$. Note that we set short-term memory to contain a fixed length of K frames since the role of short-term memory is to assist in video understanding based on previous short-term contextual information.

The update strategy for short-term memory is based on the First-in-First-out (FIFO) queue. As a new batch of visual tokens enters, when the short-term memory reaches its capacity, we pop the currently stored frames to the memory consolidation module and clear the short-term memory. The output video feature obtained from the consolidation module augments the long-term memory; on the other hand, it reinitializes the short-term memory with this feature. The initialization aims at communicating the information between different sliding windows, thereby achieving more efficient compression.

3.4. Long-term Memory

Long-term memory can effectively avoid the problem of catastrophic knowledge forgetting, which is crucial for handling long video understanding tasks. The features stored in short-term memory are dense tokens, but due to the limitations of GPU memory and computation cost, storing all the tokens dropped from short-term memory into long-term memory buffer in sequence is infeasible. Besides, we observe significant temporal redundancy in videos, where activities span multiple frames with minimal visual changes. To this end, we propose a method to merge adjacent similar frames to simplify video feature representation and accelerate video encoding. This method transforms the dense tokens to the sparse memories, which are stored in long-term memory.

As shown in Algorithm 1, we conduct memory consolidation by merging the most similar tokens in the adjacent frames following ToMe [11] periodically. We calculate the average cosine similarity s among N embedded tokens, as the tokens can well summarize the information of each frame:

$$s = \frac{1}{N} \sum_{j=1}^{N} \left[\cos(\mathbf{x}_i^j, \mathbf{x}_{i+1}^j) \right], \tag{3}$$

Our goal is to keep R_L frames after every merge operation, which also embeds rich information stored in the long-term memory. R_L is the hyper-parameter to control the trade-offs between performance and efficiency. Therefore, we greedily merge each set of adjacent frames with the highest similarity via weighted averaging. The merge operation is iteratively conducted until the token count reaches the predefined value set R_L for each consolidation operation, resulting in the output video feature $\mathbf{v}' \in \mathbb{Z}^{R_L \times 3 \times H \times W}$. The above algorithm is parameter-free, and can be easily plugged into a frame-based video encoder. Although the frame similarity calculation brings additional computing overhead, it is negligible compared to the efficiency gained by reducing stored frames.

Extend positional encoding. For long-term memory, the number of tokens exceeds the maximum length of the positional encoding from the pre-trained model. Thus, our model utilizes the positional encoding mechanism following BERT [29], which results in a portion exceeding the length threshold n without available positional encoding. In order to handle long enough long memory, we adopt the hierarchically decomposed positional encoding method proposed by Su *et al.* [57], which allows to extend the absolute positional encoding of length from n to n^2.

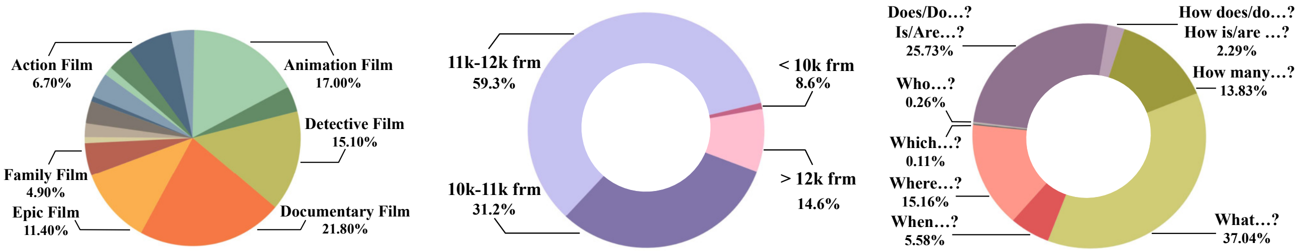
3.5. Inference

Previous methods always use the representation of the whole video to conduct understanding and question-answering, which may fail in localizing specific moment especially in long videos. To this end, we propose two inference modes, global and breakpoint, for long video understanding task as follows.

Global mode. Global mode is defined as the understanding and question-answering for the whole video. In this case, we only use long-term memory $\mathcal{L}$ as the video representation $\mathbf{V}$.

Breakpoint mode. Breakpoint mode is distinctly defined as understanding specific moments in a video. Since events inherently possess continuity, we need to consider not only the information directly related to the moments stored in short-term memory $\mathcal{S}$ but also the information indirectly related stored in long-term memory $\mathcal{L}$. Based on this, we hypothesize that when querying the movie at a specific moment t, the video representation $\mathbf{V}$ should be the aggregation of $\mathcal{L}$, $\mathcal{S}$, and the current video frame feature $\mathbf{x}_t$. We find that simply concatenating these items yields excellent performance and leave further exploration of additional aggregation choices for future work.

Subsequently, the video representation $\mathbf{V}$ goes through a Q-former and a linear projection layer before being fed into



(a) **Category.** (b) **Video length.** (c) **Question type.**

Figure 3. Video-text statistics in MovieChat-1K. It encompasses a diverse set of categories, gathered from multiple question types and containing a diverse distribution of clip durations. We annotate the video categories that account for more than 4.5% of the total (the complete list of video categories and their percentages in Appendix). "frm" represents the number of video frames.

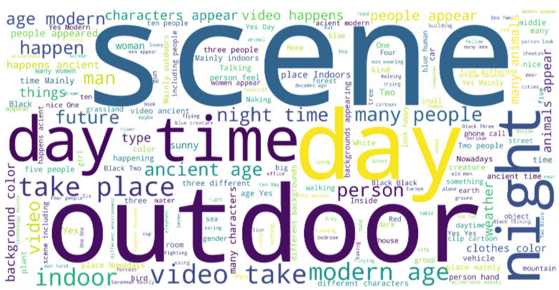


Figure 4. **Word Cloud** of the answer set in MovieChat-1K.

Method	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [72]	32.2	–	16.8	–	24.7	–
Video Chat [34]	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA Adapter [79]	54.9	3.1	43.8	<u>2.7</u>	34.2	2.7
Video LLaMA [78]	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT [40]	<u>64.9</u>	<u>3.3</u>	<u>49.3</u>	**2.8**	<u>35.2</u>	<u>2.7</u>
MovieChat *(Ours)*	**75.2**	**3.8**	**52.7**	2.6	**45.7**	**3.4**

Table 1. Quantitative evaluation for short video question answering with GPT-3.5 [43]. MovieChat achieves comparable performance even it is not specifically designed for for short video question-answering tasks. The best result is highlighted in bold, and the second best is underlined.

the LLM $\mathcal{O}$, which can be formulated as:

$$\mathbf{A} = \mathcal{O}(\mathbf{Q}, \mathcal{P}(\mathbf{V})), \tag{4}$$

where $\mathcal{P}$ is the projection from visual space to text space. $\mathbf{A}$ represents the answer or instruction, and $\mathbf{Q}$ is employed to denote the question, respectively.

4. A New Benchmark: MovieChat-1K

Previous works on building long video understanding benchmarks either focus on non-question-answering tasks (*e.g.*, language grounding [55], generic event boundary detection [54], user engagement and movie metadata prediction [67], *etc.*) or lack long-form understanding evaluation [28]. To better evaluate the performance of MovieChat, we collect a new benchmark for long video understanding tasks, MovieChat-1K, which contains 1K high quality video clips sourced from various movies and TV series with 14K manual annotations.

As shown in Fig. 3a, we collect videos from 15 popular categories with varying distribution, including documentary film, detective film, animation film, and so on. Among these, each video comprises multiple alternating scenes, contributing to a diverse and dynamic visual narrative within the context of the collection. The visual representation in Fig. 3b demonstrates the clip duration distribution of MovieChat-1K. Over 90% of the videos exhibit a duration ranging from 10K to 12K frames, while 14.6% of videos extending beyond 12K frames. Only 8.6% of videos have duration less than 10k frames.

For each video, we manually set and provide 1 dense caption for the whole video, 3 question-answering pairs for global mode and 10 question-answering pairs with timestamps for breakpoint mode. Fig. 3c illustrates the distribution of question types in MovieChat-1K. Note that MovieChat-1K is specifically designed for long video comprehension tasks, the majority of questions are open-ended, with only a quarter classified as multiple-choice questions, marked by initiators such as 'Do,' 'Does,' 'Is,' or 'Are.' We also compute the word distributions of our provided question-answer pairs. As illustrated in Fig. 4, which includes common objects (people, clothes, etc.), time (day, night, etc.), scenes (indoor, outdoor, etc.), and so on. More statistics information can be found in appendix.

5. Experiments

We conduct quantitative and qualitative evaluations between MovieChat and previous methods. Additionally, we perform ablation studies to investigate MovieChat. Experimental settings and analyses can be found in appendix.

Method	CI	DO	CU	TU	CO
Video Chat [34]	2.23	2.50	2.53	1.94	2.24
LLaMA Adapter [79]	2.03	2.32	2.30	1.98	2.15
Video LLaMA [78]	1.96	2.18	2.16	1.82	1.79
Video-ChatGPT [40]	2.40	2.52	2.62	1.98	2.37
MovieChat *(Ours)*	**2.76**	**2.93**	**3.01**	**2.24**	**2.42**

Table 2. Quantitative evaluation for short video generation performance with GPT-3.5 [43]. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

5.1. Quantitative Evaluation

Short video question-answering. We use several widely used open-ended datasets: MSVD-QA [70], MSRVTT-QA [71], and ActivityNet-QA [75] for short video question-answering tasks. The evaluation process is under the assistance of LLM with the default hyper-parameter settings. The accuracy and relative scores on a scale of 0 to 5 are reported. Compared to previous methods [34, 40, 78, 79], MovieChat achieves comparable performance even it is not specifically designed for short video question-answering tasks, as shown in Tab. 1.

Short video generative performance. Following [40], we employ GPT-assisted evaluation to conduct a more comprehensive comparison of the text generation performance between MovieChat and previous methods [34, 40, 72] on processed ActivityNet-QA [75]. The evaluation pipeline covers crucial metrics (including *Correctness of Information*, *Detailed Orientation*, *Contextual Understanding*, *Temporal Understanding* and *Consistency*) and assigns relative scores to the generated predictions on a scale of 1-5. We present the results of the generation performance evaluation in Tab. 2. The results reveal its competitive performance across all key aspects compared to previous methods.

Long video question-answering. We evaluate the long video question-answering performance of MovieChat with our proposed MovieChat-1K. We split 1,000 videos into training set (800), test set (100), validation set (100) and only use test set for final performance evaluation. We select three recent LLM-based video understanding models (*e.g.* Video Chat [34], Video LLaMA [78], and Video-ChatGPT [40]) as the baselines. Yet, none of those methods can support such long video (>10K frames). Therefore, to accommodate their length limitations in global questions, we uniformly sample from the original video up to the maximum frame count which can be officially supported by each individual model. For breakpoint questions, we extend half of the maximum frame count before and after the breakpoint (*i.e.*, placing the breakpoint at the center frame).

Method	# Frames	Global Mode		Breakpoint Mode	
		Accuracy	Score	Accuracy	Score
Video Chat [34]	32	57.8	3.00	46.1	2.29
Video LLaMA [78]	32	51.7	2.67	39.1	2.04
Video-ChatGPT [40]	100	47.6	2.55	48.0	2.45
MovieChat *(ours)*	2048	**62.3**	**3.23**	**48.3**	**2.57**

Table 3. Quantitative evaluation for long video question answering on MovieChat-1K test set in global mode with the average of GPT-3.5 [43], Claude [3] and human bling rating. HBR stands for human blind rating. The best result is highlighted in bold, and the second best is underlined.

Method	CI	DO	CU	TU	CO
Video Chat [34]	3.04	2.75	3.09	3.00	3.21
Video LLaMA [78]	2.75	2.24	2.83	2.62	2.97
Video-ChatGPT [40]	2.37	2.30	2.58	2.49	2.69
MovieChat *(Ours)*	**3.11**	**2.93**	**3.24**	**3.17**	**3.25**

Table 4. Quantitative evaluation for long video generation performance in global mode with the average of GPT-3.5 [43], Claude [3] and human blind rating. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is in bold, and the second best is underlined.

To enhance the robustness of the results, we simultaneously employ GPT-3.5 [43] and Claude [3] as LLM assistants, with the additional support of human blind rating. We observe a discrepancy between the accuracy and relative score generated by the previously LLM-assisted evaluation method [40] for video question-answering tasks. However, merely adjusting the prompt for the LLM cannot effectively address this issue. Therefore, after obtaining the accuracy and score from the LLM-assisted evaluation method, we implement manual filtering to remove results with inconsistent values, thus improving the reliability of our outcomes.

As shown in Tab. 3, compared to previous methods [34, 40, 78], MovieChat reads more video frames. In both global mode and breakpoint mode, our method maintains a performance gain in terms of the average accuracy and score provided by LLM assistants and human blind rating. We comprehensively evaluate MovieChat's question-answering performance across different question types compared to baselines. The results indicate that our approach outperforms the baselines in both open-ended and true-false questions.

Long video generative performance. We compare the quality of answers generated by MovieChat and previous methods [34, 40, 78] in long video question-answering on MovieChat-1K. As shown in Tab. 4, with the average score

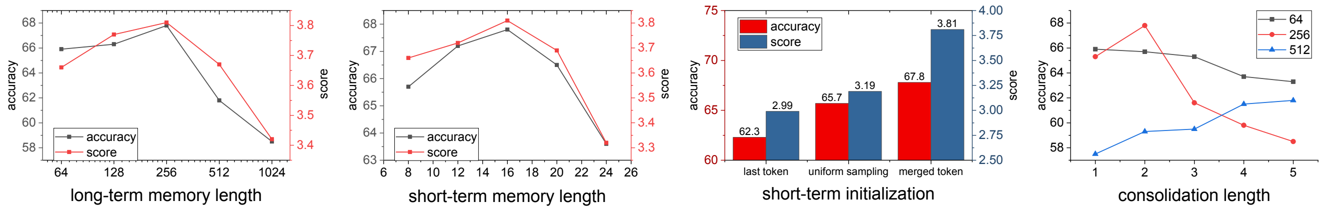


Figure 5. Hyperparameter ablation studies on how length of long-term memory buffer l_{length}, short-term memory buffer l_{short}, consolidation length l_{merge} and short-term initialization affect the performance of MovieChat on long video understanding. We set $l_{short} = 16$, $l_{merge} = 2$ in ablation study of long-term memory, $l_{long} = 256$, $l_{merge} = 2$ in ablation study of short-term memory and $l_{short} = 16$ in ablation study of consolidation length and short-term initialization.

Method	Global Mode		Breakpoint Mode	
	Accuracy	Score	Accuracy	Score
w/o MM	51.4	3.10	38.2	2.31
base	**67.8**	**3.81**	**50.4**	**2.96**

Table 5. Ablation study on how memory mechanism (MM) affects the long video question answering. The best result is in bold.

Method	Global Mode					Breakpoint Mode				
	CI	DO	CU	TU	CO	CI	DO	CU	TU	CO
w/o MM	3.30	2.53	3.28	2.77	3.42	2.42	2.85	2.87	2.00	2.87
base	**3.32**	**3.28**	**3.40**	**2.97**	**3.48**	**2.97**	**3.24**	**3.31**	**2.70**	**3.45**

Table 6. Ablation study on how memory mechanism (MM) affects the long video generative performance. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is in bold.

provided by GPT-3.5 [43], Claude [3] and human bling rating, our approach continues to generate higher-quality answers even as the video contents become more extensive.

5.2. Ablation Study

Short-term and long-term memory buffers. As MovieChat incorporates a memory mechanism including short-term memory and long-term memory, it is imperative to evaluate how the proposed memory mechanism influences the performance. Tab. 5 and Tab. 6 provide the memory-dependent performance of MovieChat for long video question-answering and generative tasks with the average results of GPT-3.5 [43], Claude [3], and human blind rating. MovieChat with the memory mechanism significantly outperforms the memory-independent variant, which signifies the importance of memory mechanisms.

Hyper-parameter ablations. We perform a series of hyperparameter ablations based on the MovieChat-1K dataset to better understand MovieChat. Fig. 5 shows the performance when ablating the length of memory buffers, consolidation length and short-term initialization with the average results of GPT-3.5 [43], Claude [3], and human blind rating. The performance of MovieChat degrades when all four are significantly changed, showing the validity of our empirically chosen hyperparameyers. Fig. 5 demonstrates that information obtained from the video expands with the growing length of memory buffers, while the loss of finer details intensifies with the fixed length of consolidation. Furthermore, using merged tokens for short-term initialization outperforms last few tokens and uniform sampling. Additionally, the length of merged tokens and the memory buffer size have a combined effect on MovieChat's performance.

5.3. Case Study

We perform an extensive case study of MovieChat on a variety of open-ended long video (such as cartoon movie and TV series) for long video question-answering, including the ■ breakpoint mode (Q#1) and the ■ global mode (Q#2). The evaluation is conducted between MovieChat and previous methods [34, 40, 79] as shown in Fig. 6 . For Q#1 in breakpoint mode, we mark the timestamp when the question is asked. For long videos over 10K frames, MovieChat is still capable of providing excellent responses to questions regarding both the current moment and the entire video content with less hallucination. More examples to show long video scene understanding and temporal understanding ability of MovieChat are available in appendix.

6. Limitation

Although MovieChat has demonstrated impressive abilities in long video understanding, it is still an early-stage prototype and has some limitations, including: 1) Limited perception capacities. MovieChat's performance is hindered by the pretrained short video understanding model. 2) Inadequate Time Processing. MovieChat provides only rough estimates of the duration proportions of events within long videos, lacking precision in temporal details.

Question #1 **What may happen next step?** **Breakpoint Mode**

Video Chat: Based on the video, it appears that the person is cooking meat and vegetables on a stove top using different utensils and ingredients such as onions, garlic, mustard, and olive oil. It is likely that the person will continue cooking the meat until it is fully cooked, and then serve it to their family or friends. Additionally, they may add other ingredients such as mushrooms or potatoes to the dish. It is possible that they will use different types of cooking methods such as sautéing or grilling to ensure the meat is cooked evenly and to perfection.

Video LLaMA: At frame 4250, the video shows the hand of a man cutting meat on a cutting board. It's not clear what happens next, but it is possible that the man continues cutting the meat.

Video-ChatGPT: The next step in the video is to place the cooked meat on a plate and serve it. The person in the video is seen cutting the meat into smaller pieces and placing it on a plate. The person then takes a fork and begins to eat the meat.

MovieChat: I think the man will prepare the food in the kitchen. It seems like he is going to cook it with spices and herbs. He has a large wooden board on the kitchen counter and a large white pot on the stove, and he will probably use some tools to prepare the food. He might also use his hands or a spatula to cook the meat.

Question #2 **Which step took the longest to complete this dish?** **Global Mode**

Video Chat: Based on the video, it appears that the longest step in making the dish is preparing the pan for cooking the steak. Specifically, the man takes the pan out of the oven and starts cleaning it by wiping it with a rag, then rubbing it with olive oil, and finally heating the pan to high heat. It is also mentioned that the man has to wait for the pan to heat up before putting the steak in the pan.

Video LLaMA: The dish that this video portrays took several minutes to complete. However, the most time-consuming task seems to have been preparing the ingredients on the white kitchen counter. The man in the video is seen chopping vegetables and using a sharp knife to prepare a meal. Finally, he will sit down to enjoy the meal, savoring each bite and reflecting on the process that led to this culinary creation.

Video-ChatGPT: The longest step in preparing the dish was the time it took to cook the meat. The man in the video cooks the meat on a stove for a long time, which is a crucial step in the process. The video shows the man cooking the meat on a stove for a considerable amount of time, which indicates that it took a considerable amount of effort and patience to achieve the desired result.

MovieChat: The longest step to complete this dish would be the preparation of the food on the cutting board. It requires chopping the meat and adding other ingredients like broccoli and carrots, and mixing them together before putting them in the pan. The chef had to carefully chop and arrange the ingredients before cooking the dish.

Figure 6. Question and answer about a clip from *YouTube*, which is a tutorial on how to cook steak. The entire instructional process begins with marinating the steak, followed by pan-searing it, preparing side dishes, and ultimately plating the meal. Green (Red) highlights the correct (wrong) answer and yellow indicates that the model is hallucinating.

7. Conclusion

Conclusively, we presents an innovative video understanding system integrating video foundation models and large language models. By incorporating a memory mechanism represented by tokens in Transformers, MovieChat tackles challenges in analyzing long videos. MovieChat achieves state-of-the-art performance in long video understanding, surpassing existing systems limited to handling videos with few frames. This work opens up opportunities for applications requiring a comprehensive understanding of long-term visual information.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2022ZD0162000), the National Natural Science Foundation of China (No. 62106219, No. 62206046), the Zhejiang Provincial Natural Science Foundation of China (No. LZ24F030005), and the Shanghai Sailing Program (No. 21YF1401300).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2

[2] Roy Allen, Peter Mcgeorge, David G Pearson, and Alan Milne. Multiple-target tracking: A role for working memory? *Quarterly journal of experimental psychology*, 59(6):1101–1116, 2006. 2

[3] Anthropic. Meet claude, 2023. 6, 7

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[5] Richard C Atkinson and Richard M Shiffrin. Chapter: Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2:89–195, 1968. 2

[6] George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. Trecvid 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. *arXiv preprint arXiv:2104.13473*, 2021. 2

[7] George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, et al. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. *arXiv preprint arXiv:2009.09984*, 2020. 2

[8] George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzad Godil, David Joy, Andrew Delgado, Alan F Smeaton, Yvette Graham, et al. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*, 2018. 2

[9] George Awad, Asad A Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Willie Mcclinton, Martial Michel, Alan F Smeaton, Yvette Graham, Wessel Kraaij, et al. Trecvid 2017: evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking. In *TREC Video Retrieval Evaluation (TRECVID)*, 2017. 2

[10] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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, 2021. 2

[11] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[13] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8090–8100, 2022. 2

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[15] Wenhao Chai and Gaoang Wang. Deep vision multimodal learning: Methodology, benchmark, and trend. *Applied Sciences*, 12(13):6588, 2022. 1

[16] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2

[17] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 640–658. Springer, 2022. 2

[18] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 1, 2

[19] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1

[20] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. 2022. 3

[22] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Zhenyu Qiu, Wei Lin, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[23] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1

[24] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2

[25] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 1, 2

[26] Zhicheng Hao, Jun Qiu, Haimiao Zhang, Guangbo Ren, and Chang Liu. Umotma: Underwater multiple object tracking with memory aggregation. *Frontiers in Marine Science*, 9:1071618, 2022. 2

[27] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021. 2

[28] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 2, 5

[29] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 4

[30] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 1, 2

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2023. 3

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[34] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1, 2, 5, 6, 7

[35] Boyu Liu, Yanzhao Wang, Yu-Wing Tai, and Chi-Keung Tang. Mavot: Memory-augmented video object tracking. *arXiv preprint arXiv:1711.09414*, 2017. 2

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[38] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023. 1

[39] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Adaptive correlation filters with long-term and short-term memory for object tracking. *International Journal of Computer Vision*, 126:771–796, 2018. 2

[40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 5, 6, 7

[41] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. 2

[42] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[43] openai. Gpt3.5, 2021. 2021. 5, 6, 7

[44] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[45] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2

[46] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 184–195. Springer, 2014. 2

[47] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2017. 2

[48] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE, 2012. 2

[49] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 144–157. Springer, 2012. 2

[50] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119:346–373, 2016. 2

[51] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 154–171. Springer, 2020. 2

[52] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 629–645. Springer, 2020. 2

[53] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021. 2

[54] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8075–8084, 2021. 2, 5

[55] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 2, 5

[56] Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766, 2015. 2

[57] Jianlin Su. Bert position encoding. `https://kexue.fm/archives/7947`, 2023. 4

[58] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 1

[59] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 1, 2

[60] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[62] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[63] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*, 2023. 1

[64] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. *arXiv preprint arXiv:2308.07863*, 2023. 2

[65] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 1

[66] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 2

[67] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 2, 5

[68] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[69] Ming Xin, Wenjie Sun, Kaifang Li, and Guancheng Hui. Multi-object tracking with spatial-temporal correlation memory networks. In *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, pages 616–619. IEEE, 2022. 2

[70] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 6

[71] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. June 2016. 6

[72] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. 5, 6

[73] Tianyu Yang and Antoni B Chan. Learning dynamic memory networks for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 152–167, 2018. 2

[74] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 2

[75] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 6

[76] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical

video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023. 2

[77] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Title generation for user generated videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 609–625. Springer, 2016. 2

[78] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 2, 5, 6

[79] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 5, 6, 7

[80] Zechu Zhou, Xinyu Zhou, Zhaoyu Chen, Pinxue Guo, Qian-Yu Liu, and Wenqiang Zhang. Memory network with pixel-level spatio-temporal learning for visual object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[81] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2