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

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

[1] Zhejiang University [2] University of Washington [3] Microsoft Research Asia
[4] Hong Kong University of Science and Technology (GZ)

♠ Equal contribution ♡ Project lead ♢ Data collection ✉ Corresponding author

https://rese1f.github.io/MovieChat

Abstract

Recently, integrating video foundation models and large language models to build a video understanding system overcoming 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 are the remaining challenges. Inspired by Atkinson-Shiffrin memory model, we develop an memory mechanism including a rapidly updated short-term memory and a compact thus sustained long-term memory. We employ tokens in Transformers as the carriers of memory. MovieChat achieves state-of-the-art performace in long video understanding.

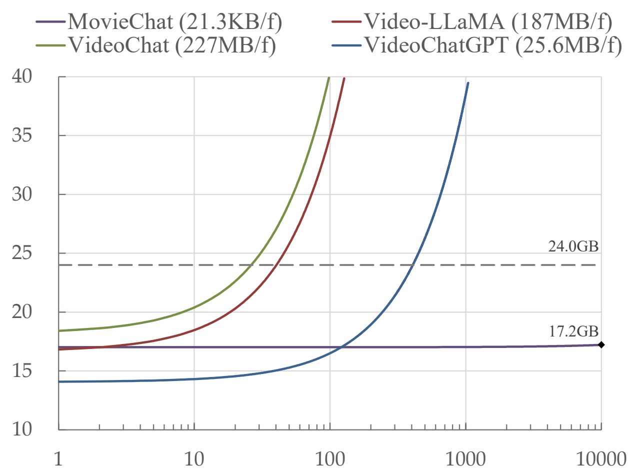


Figure 1. GPU memory cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison. We test the visual-only inference of all methods at a resolution of 224×224 without frame sampling. While the previous method can only support around 100 frames of inference, **MovieChat** can handle videos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

1. Introduction

Recent advances in Large Language Models (LLMs) [11, 17, 60, 78, 82] acheive great success in Natural Language Processing (NLP) field. It is a natural progression to introduce multi-modality [13] into LLMs and turn it into Multi-modal Large Language Models (MLLMs) [1, 18, 24, 27, 29, 40, 41, 43, 50, 53, 55, 76, 88, 100, 101, 110], which is able to conduct multimodal perception and understanding. MLLMs have shown incredible emergent capabilities in various multimodal tasks such as perception (*e.g.*, existence, count, position, OCR), commonsense reasoning, and code reasoning, leading a potential path to Artificial General Intelligence (AGI). Compared to LLMs and other task-specific models, MLLMs provide a more human-like perception of the world, a user-friendly interface for interaction, and a broader range of task-solving capabilities.

Existing vision-centric MLLMs follows the paradigm that utilizing pre-trained LLMs and visual encoder with additional learnable modules (Q-former [18, 41, 43, 106] or simple projection layer [21,50,55,76]). In video field, some previous works [55,106] follow this paradigm to build video MLLMs, while another paradigm [44,85] is that combining existing visual perception tools (*e.g.*, tracking and classification) and LLMs through API to build a system without training. Yet, previously, there was no exploration of a model or system based on long videos (over one minute), and there was also a lack of a standardized benchmark to evaluate the capabilities of these systems in this regard.

In this paper, we present MovieChat, a novel framework that integrating 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 remaining challenges for long video understanding. To this end, we propose a memory mechanism inspired by Atkinson-Shiffrin memory model [4], which including a rapidly updated short-term memory and a compact thus sustained long-term memory.

The contributions of this work are summarized as:

- We present MovieChat, a novel framework that integrating vision models and LLMs, is the first to support long video understanding tasks.
- We propose a kind of memory mechanism to reduce the computation complexity and memory cost, while enhancing the long-term temporal connection.
- We conduct extensive quantitative evaluation and case studies to evaluate the performance of both understanding capability and inference cost.

2. Related Works

2.1. Video Foundation Models

Video foundation models have various applications on downstream tasks (*e.g.*, video question answering [37, 38], video captioning [34, 98], and human action recognition [36, 62, 63]). The common paradigm in the field of video foundation models is now characterized by the combination of extensive large-scale video-language pretraining, followed by fine-tuning on specific downstream tasks [45,46,48,57,77,89,96,111]. Such paradigm depends on end-to-end video-language joint training with pretext pre-training tasks such as masked language modeling [47], masked video modeling [80, 87], video-language masked modeling [25], video-text matching [86], and video-text contrastive learning [96]. These prior arts yield impressive performance in multimodal video tasks. Yet, they can only train with limited video-language pairs or videos without detailed annotations, which leads to difficulties in language-related tasks. With connecting to LLMs, video foundation models serve as visual encoder and achieve state-of-the-art performance in various tasks and become user-friendly.

2.2. Multi-modal Large Language Models

LLMs [11, 17, 60, 78, 82, 83] achieves great success recently. Many works try to build MLLMs [1, 18, 24, 27, 29, 40,41,43,50,53,55,76,88,100,101,110] by combining models of other modalities. Flamingo [1] bridge powerful pretrained vision-only and language-only models and achieve state-of-the-art performance with few-shot learning, simply by prompting the model with task-specific examples. BLIP-2 [41] proposes a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. MiniGPT-4 [110] aligns a frozen visual encoder with a frozen LLM, Vicuna [17], using just one projection layer. Otter [40] is trained on MIMIC-IT [39] and showcasing improved instruction-following ability and in-context learning. In video field, ChatVideo [85] treats tracklets as the basic video unit and allow user interacting with the LLMs. VideoChat [44] integrates video foundation models and LLMs via a learnable neural interface, excelling in spatiotemporal reasoning, event localization, and causal relationship inference. VideoChat Longvideo [61] further incorporates LangChain [33] into VideoChat to support video which more than one minutes. Video-LLaMA [106] further leverage pre-trained models ImageBind [28] and LLaMA [82] bootstraping cross-modal training in video following BLIP-2. Yet, those methods fails in handling long video understanding since high computation complexity, large memory cost, and weak long-term temporal connection. Therefore, our main efforts is to introduce memory mechanism to enhance those aspects.

2.3. Long Video Understanding

Understanding long videos is a challenging task in computer vision. Prior arts use 3D volumn [90], object/human-centric [67, 91], or other forms [71, 92] as video representations. There are also several datasets of video-caption/description pairs among various domains such as cooking (*e.g.*, YouCook [19, 108], MPII Cooking [68–70], and TACoS [64, 65]), instruction (*e.g.*, HowTo100M [58] and HiREST [104]), and movie (*e.g.*, MovieQA [79], M-VAD [81], MPII-MD [66], and MovieNet [32]) from different sources such as YouTube [14, 58, 105], Twitter [5–8], Flirck [5, 6], and internet [9]. Yet, those datasets lack diverse and fine-grained dense captioning for long videos.

2.4. Memory Models in Vision Tasks

There are some prior works exploring memory models [74] in various vision tasks in video, such as video object segmentation (VOS) [15, 16, 31, 52, 56, 59, 72, 73, 93, 102], multi-object tracking (MOT) [2,12,22,30,94] and visual object tracking (VOT) [26,49,54,99,109]. MeMOT [12] build a large spatiotemporal memory that stores the past observations of the tracked objects. XMem [15] develop an architecture that incorporates multiple independent yet deeply-connected feature memory stores to handle long videos with thousands frames. We drew the experience of those prior arts and further adopt memory model combining with LLMs. Unlike using embedded feature given by certain visual encoder, we found that using tokens in Transformers [84] as the carriers of memory suitable for both LLMs

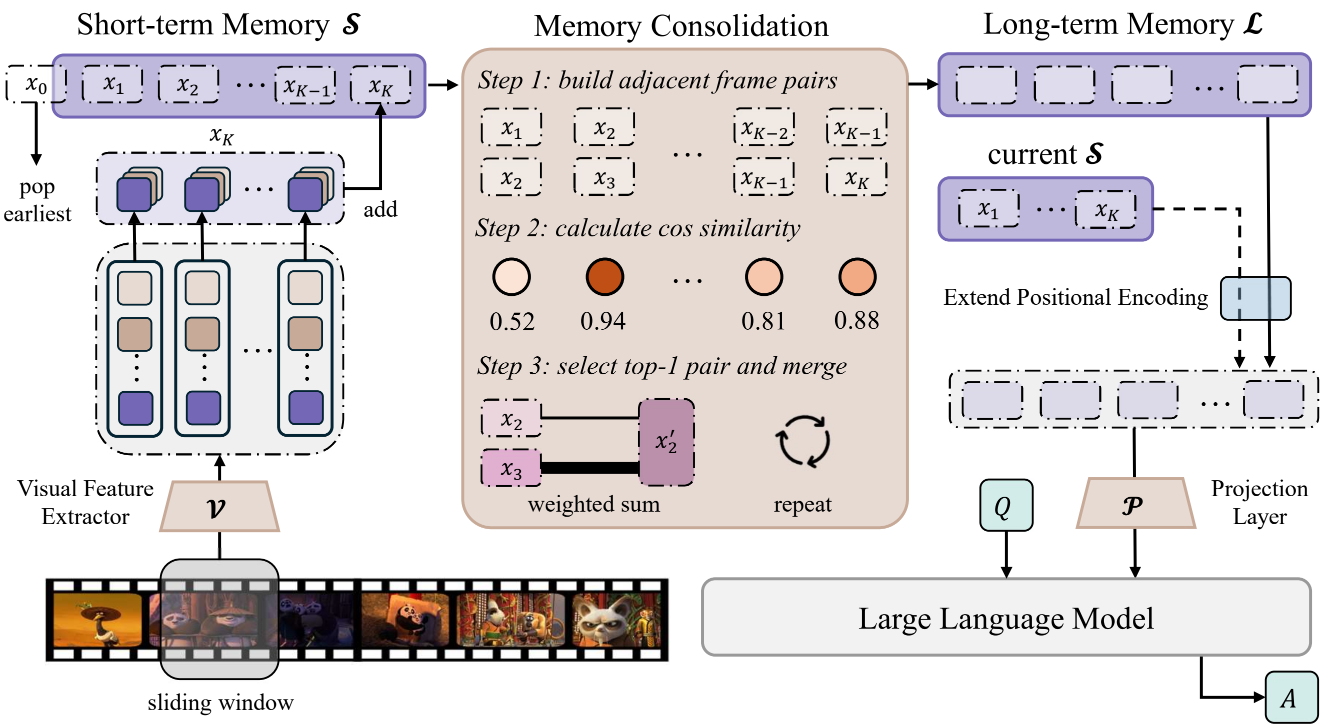


Figure 2. **Illustration of MovieChat.** We use a sliding window approach to extract video features and represent them in token form, which are then sequentially fed into the short-term memory frame by frame. The short-term memory has a fixed length, and when it reaches its set limit, the earliest tokens are popped and consolidated into the long-term memory. We have designed two inference modes: global mode, which exclusively utilizes the long-term memory, and breakpoint mode, which additionally incorporates the current short-term memory as part of the video representation. Breakpoint mode allows for understanding the video at a specific moment in time. Finally, after passing through a projection layer, the video representation is inputted into a large language model for interaction with the user.

and ViT [20] based visual encoder. Our proposed method mainly focus on reducing the redundant of visual tokens in video and building a memory mechanism to pass the information among large temporal range.

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 buffers, the video projection layer, and the Large Language Model (LLM), as illustrated in Figure 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. 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 [3] or Video-Swin [51], we simply use 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 [23] and the Q-former from BLIP-2 [42]. This is mainly because 1) there is no video foundation model makes good alignment with text, and 2) our proposed memory mechanism can effectively capture temporal features. Given input video $\mathbf{v} \in \mathbb{Z}^{T\times 3\times H\times W}$, a sequence of T RGB frames, with height and width H and W, the visual feature extraction by sliding window approach could be formulated as

$$\mathbf{x}_i = \mathcal{V}(\mathbf{v}_i), \mathbf{v}_i \in \mathbb{Z}^{C\times 3\times H\times W}, i = 0, 1, ..., \lceil \frac{C}{T} \rceil, \quad (1)$$

where $\mathcal{V}(\cdot)$ is the visual feature extractor, $\mathbf{v}_i$ are the RGB values of the video clip, $\mathbf{x}_i \in \mathbb{R}^{C\times N\times D}$ are the visual to-

Algorithm 1 Memory consolidation

Require: $\mathcal{S}$ ▷ short-term memory
while $len(\mathcal{S}) > N$ **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

kens, N is the number of tokens, and D is the feature dimension.

3.3. Short-term Memory

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

$$\mathcal{S} = \{\mathbf{x}_i \mid \forall i = 1, 2, ..., K\}, \tag{2}$$

where $\mathcal{S}$ is short-term memory. Note that we set short-term memory to a fixed length since the role of short-term memory is to assist in understanding based on previous short-term contextual information. The update strategy for short-term memory is similar to the First-in-First-out (FIFO) queue. When a new batch of visual tokens enter, we drop the earliest one that was present. The dropped tokens are further consolidated into long-term memory.

3.4. Long-term Memory

Long-term memory can effectively avoid the problem of catastrophic knowledge forgetting, which is crucial for processing long videos. The features stored in short-term memory are dense tokens, but due to GPU memory and computation cost limitations, directly storing all the tokens dropped from short-term memory into long-term memory buffer in sequence is unavailable. Besides, temporally adjacent frames may exhibit significant similarity in the video. To this end, we propose a method to merge temporally adjacent similar frames. This method transforms the dense tokens to the sparse memory and storing them in long-term memory.

To be specific, as shown in Algorithm 1, we conduct memory consolidation by merging the most similar tokens in the adjacent frames following ToMe [10]. We found that the token enbedding in transformers already summarize the information of each frame for use in cos similarity s as:

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

We iteratively conduct this operation until the token count reaches the value set for each consolidation operation.

Extend positional encoding. For long-term memory, the number of tokens exceeds the maximum length of the pretrained model positional encoding. Our model utilizes the positional encoding mechanism following BERT [35], which results in a portion exceeding the length threshold n without available positional encoding. In order to handle long enough long memory, we adopted the hierarchical decomposed positional encoding method proposed by Su *et al*. [75], extending the absolute positional encoding of length n to n^2.

3.5. Inference

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

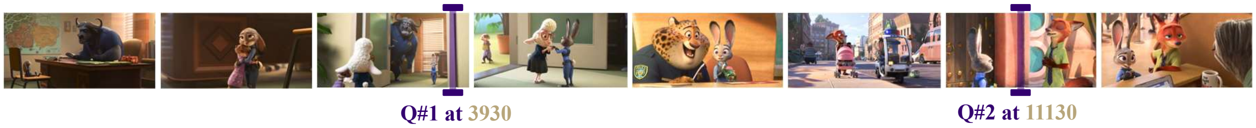
Global mode. Global mode is defined as the understanding and question-answering for the **whole** video. In this case, we only use long-term memory $\mathcal{L}$ as the video representation $\mathbf{V}$.

Breakpoint mode. Breakpoint mode is defined as understanding specific moments in a video. Since events have 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, the video representation $\mathbf{V}$ should be the aggregation of $\mathcal{L}$, $\mathcal{S}$, and the current video clip feature $\mathbf{x}$. We found that simply concatenating these items yields excellent performance. We leave further exploration of additional aggregation choices for future work. future work.

After that, the video representation $\mathbf{V}$ goes through a q-former and a linear projection layer before being fed into the LLM, which can be formulated as:

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

where $\mathcal{P}$ is the projection from visual space to text space, $\mathcal{L}$ is the large language model, $\mathbf{A}, \mathbf{Q}$ are the answer or instruction and the question.



Question #1 **What is Chief Bogo doing?**

Answer #1 In the video, the character Chief Bogo is seen talking to someone on a laptop while sitting at a table in a dark room. Chief Bogo is a cartoon character wearing a brown outfit.

Question #2 **Who is this rabbit named Judy talking to?**

Answer #2 In the video, the rabbit police named Judy is talking to the person in a movie theater, who is a fox.

Question #3 **Can you describe this video?**

Answer #3 The video shows various scenes of an animation with different animals in a restaurant or office environment. At the beginning, we see an animated cartoon rabbit and a deer talking to each other in a restaurant. Next, there is a cartoon bunny talking to a cat at a desk, followed by a scene of a cartoon deer and a mouse talking to each other. The video then continues with the animated cartoon dog and cat sitting at a table in a restaurant, and another scene where a man and a dog are in a room with a computer on a table. There are more scenes of the same characters later on, including one with a bunny and a fox sitting at a desk, and another with a bunny and a dog talking to each other in the background. Finally, there are several more scenes with different animals, including a fox, a mouse, and a cat, all of which are in a different environment.

Figure 3. Question and answer about clips from *Zootopia*, a cartoon, which tells the story of a determined police officer rabbit named Judy who pairs up with a cunning fox to uncover a conspiracy about missing animals and develop an unexpected friendship.

Question #1 **What is she doing?**

Answer #1 In the given video, we see a young woman walking down a street in the city, and a young man also walks down the street behind her. The woman is wearing a red and black outfit.

Question #2 **Where is this couple?**

Answer #2 This couple is standing in the middle of a city street with a black hat on the man's. The scene is shot in a park, and there are a few trees and bushes visible.

Question #3 **Can you tell me what happened in this movie?**

Answer #3 Certainly! In the video, you can see a group of people standing in a park with a green hill in the background. There are also people walking on the street and in a courtyard with trees. Some of the people are holding red umbrellas, and there is a woman wearing a red coat. She is then seen walking down a street at night with a man following her. Next, we see a couple standing on a balcony overlooking a street in the evening with a man walking by and then walking down the street. Towards the end of the video, a group of people are seen walking through a snowy town.

Figure 4. Question and answer about clips from *Goblin*, which tells the story of Kim Shin, an immortal "goblin" who needs to find a human bride to end his endless life but instead meets Ji Eun-tak, a girl fated to die who claims to be the "goblin's bride," leading to a romantic tale unfolding between them.

Method	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [97]	32.2	–	16.8	–	24.7	–
Video Chat [44]	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA Adapter [107]	54.9	3.1	43.8	2.7	34.2	2.7
Video LLaMA [106]	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT [55]	**64.9**	**3.3**	49.3	2.8	35.2	2.7
MovieChat *(Ours)*	61.0	2.9	**49.7**	**2.8**	**51.5**	**3.1**

Table 1. Quantitative evaluation on short video question answering. MovieChat achieves competitive results.

4. Experiments

4.1. Quantitative Evaluation

Short video question answering. We conducted a comprehensive quantitative evaluation in this section. We use several widely used open-ended datasets: MSVD-QA [95], MSRVTT-QA [95], and ActivityNet-QA [103] for short video question answering task. The evaluation process is under the assistant of LLM (details in Appendix A) under default hyper-parameter settings as shown in Appendix B. The accuracy and relative score on a scale of 0 to 5 are reported. Compared to previous method [44, 55, 97], MovieChat achieves competitive result although there is no specific design for short video understanding.

4.2. Case Study

We perform an extensive case study of MovieChat on a variety of open-ended long video (such as cartoon movie in and TV series) for long video question-answering and captioning task, including the **global mode** and the **breakpoint mode** as shown in Figure 4. For Q#1 and Q#2, we annotate timestamps in frames. 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.

5. Conclusion

In conclusion, this paper presents an innovative video understanding system that integrates video foundation models and large language models. By incorporating a memory mechanism inspired by the Atkinson-Shiffrin model, consisting of short-term and long-term memory represented by tokens in Transformers, the system overcomes challenges associated with analyzing long videos. The proposed system, named MovieChat, achieves state-of-the-art performance in long video understanding, surpassing existing systems limited to handling videos with few frames. This approach reduces computation complexity, memory cost, and addresses long-term temporal connections.

The study emphasizes the significance of memory mechanisms in video understanding, enabling the model to retain and retrieve relevant information over extended durations. MovieChat's success has practical implications in domains like video surveillance, content analysis, and video recommendation systems. Future research can explore further improvements to the memory mechanism and the integration of other modalities, such as audio, to enhance video understanding capabilities. This work opens up opportunities for applications requiring a comprehensive understanding of visual information.

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] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

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

[14] 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

[15] 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

[16] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[17] 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

[18] 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, 2

[19] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[21] 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

[22] Kuan Fang, Yu Xiang, Xiaocheng Li, and Silvio Savarese. Recurrent autoregressive networks for online multi-object

tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 466–475. IEEE, 2018. 2

[23] 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

[24] 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, 2

[25] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[26] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13774–13783, 2021. 2

[27] 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, 2

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] hwchase17. langchain. `https://github.com/hwchase17/langchain`, 2023. 2

[34] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020. 2

[35] 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

[36] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 2

[37] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018. 2

[38] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, 2020. 2

[39] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 2

[40] 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

[41] 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

[42] 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

[43] 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, 2

[44] 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, 6

[45] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 2

[46] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023. 2

[47] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129, 2023. 2

[48] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*, 2020. 2

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

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2

[51] 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

[52] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 661–679. Springer, 2020. 2

[53] 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, 2

[54] 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

[55] 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, 2, 6, 12

[56] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9670–9679, 2021. 2

[57] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[58] 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

[59] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2

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

[61] OpenGVLab. Ask-anything. `https://github.com/OpenGVLab/Ask-Anything/tree/long_video_support`, 2023. 2

[62] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010. 2

[63] Zhenting Qi, Ruike Zhu, Zheyu Fu, Wenhao Chai, and Volodymyr Kindratenko. Weakly supervised two-stage training scheme for deep video fight detection model. *arXiv preprint arXiv:2209.11477*, 2022. 2

[64] 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

[65] 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

[66] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 2

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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

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

[76] 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, 2

[77] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 2

[78] 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

[79] 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

[80] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*. 2

[81] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015. 2

[82] 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

[83] 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

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[85] 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, 2

[86] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 2

[87] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 2

[88] 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, 2

[89] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2

[90] 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

[91] 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

[92] 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

[93] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021. 2

[94] 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

[95] 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

[96] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021. 2

[97] 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. 6

[98] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 2

[99] 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

[100] 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

[101] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1, 2

[102] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. *arXiv preprint arXiv:2307.12616*, 2023. 2

[103] 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

[104] 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

[105] 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

[106] 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, 6

[107] 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. 6

[108] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[109] 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

[110] 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

[111] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2

Appendix

A. LLM-Assisted Evaluation

Following [55], we use LLM-Assisted Evaluation for short video question answering task in Section 4.1. Given the question, correct answer, and predicted answer by model, ChatGPT should return the *True* or *False* judgement and relative score (0 to 5). The whole prompt is shown in Figure A1. It takes about 250 tokens per question. We report the baseline results of short video question answering from `https://github.com/mbzuai-oryx/Video-ChatGPT`.

B. Hyper-parameter Setting

Description	Default Value
size of sliding window	10 frames
size of short memory	8 frames $\times$ 32 tokens per frames
size of consolidated memory	64 tokens

Table 2. Hyper-parameter settings of MovieChat.

```
openai.ChatCompletion.create(
  model="gpt-3.5-turbo",
  messages=[
    {
      "role": "system",
      "content":
        "You are an intelligent chatbot designed for evaluating the correctness of generative outputs
        for question-answer pairs. "
        "Your task is to compare the predicted answer with the correct answer and determine if they
        match meaningfully. Here's how you can accomplish the task:"
        "------"
        "##INSTRUCTIONS: "
        "- Focus on the meaningful match between the predicted answer and the correct answer.\n"
        "- Consider synonyms or paraphrases as valid matches.\n"
        "- Evaluate the correctness of the prediction compared to the answer."
    },
    {
      "role": "user",
      "content":
        "Please evaluate the following video-based question-answer pair:\n\n"
        f"Question: {question}\n"
        f"Correct Answer: {answer}\n"
        f"Predicted Answer: {pred}\n\n"
        "Provide your evaluation only as a yes/no and score where the score is an integer value
        between 0 and 5, with 5 indicating the highest meaningful match. "
        "Please generate the response in the form of a Python dictionary string with keys 'pred' and
        'score', where value of 'pred' is  a string of 'yes' or 'no' and value of 'score' is in INTEGER, not
        STRING."
        "DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python
        dictionary string. "
        "For example, your response should look like this: {'pred': 'yes', 'score': 4.8}."
    }
  ]
)
```

Figure A1. Prompt for ChatGPT in LLM-Assisted Evaluation.