MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

Bo He[1,2]* Hengduo Li[2] Young Kyun Jang[2] Menglin Jia[2] Xuefei Cao[2]
Ashish Shah[2] Abhinav Shrivastava[1] Ser-Nam Lim[3]

[1]University of Maryland, College Park [2]Meta [3]University of Central Florida

https://boheumd.github.io/MA-LMM/

Abstract

*With the success of large language models (LLMs), integrating the vision model into LLMs to build vision-language foundation models has gained much more interest recently. However, existing LLM-based large multimodal models (*e.g.*, Video-LLaMA, VideoChat) can only take in a limited number of frames for short video understanding. In this study, we mainly focus on designing an efficient and effective model for long-term video understanding. Instead of trying to process more frames simultaneously like most existing work, we propose to process videos in an online manner and store past video information in a memory bank. This allows our model to reference historical video content for long-term analysis without exceeding LLMs' context length constraints or GPU memory limits. Our memory bank can be seamlessly integrated into current multimodal LLMs in an off-the-shelf manner. We conduct extensive experiments on various video understanding tasks, such as long-video understanding, video question answering, and video captioning, and our model can achieve state-of-the-art performances across multiple datasets.*

1. Introduction

Large language models (LLMs) have gained significant popularity in the natural language processing field. By pre-training on large-scaled textual data, LLMs (*e.g.* GPT [1–4], LLaMA [5, 6]) have demonstrated remarkable abilities to perform both generative and discriminative tasks with a unified framework. Recently, there has been a growing interest in utilizing LLMs on multimodal tasks. By integrating LLMs with visual encoders, they can take images and videos as input and show incredible capabilities in various visual understanding tasks, such as captioning, question answering [7–13], classification, detection, and segmentation [14–20].

*Work done during Bo's internship at Meta.

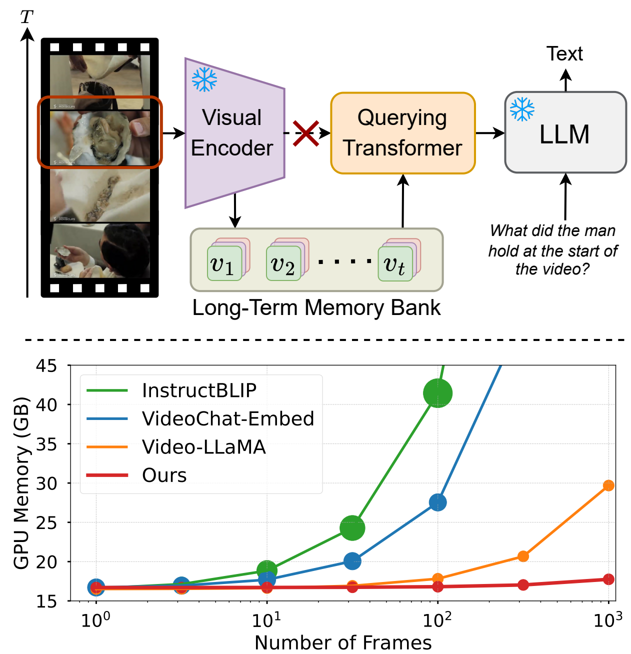


Figure 1. (a) We propose the long-term memory bank to auto-regressively store and accumulate past video information, different from previous methods directly feeding the visual encoder's outputs into the querying transformer. (b) GPU memory consumption v.s. video frame length of existing multimodal methods and MA-LMM during inference. Circle sizes represent the number of text tokens.

To handle video inputs, some prior large multimodal models [7, 9] directly feed the concatenated query embeddings of each frame along the temporal axis into LLMs. However, the inherent context length limitation of LLMs and GPU memory consumption restrict the number of video frames that can be processed. For example, LLaMA has a context length limitation of 2048 while large multimodal models like LLaVA [8] and BLIP-2 [7, 9] take in 256 and 32 tokens per image respectively. Therefore, this design is not practical and feasible when video duration is much longer (*e.g.* movies and TV shows). To address these issues, a naive solution is to apply average pooling along the temporal axis like Video-ChatGPT [21], but this leads to inferior performances as it lacks explicit temporal modeling. An alternative method involves adding a video modeling component to capture temporal dynamics, as seen in Video-LLaMA [12], which employs an extra video querying transformer (Q-Former) to obtain video-level representation. However, this design adds model complexities, increases the training parameters, and is not suitable for online video analysis.

With these in mind, we introduce a **M**emory-**A**ugmented **L**arge **M**ultimodal **M**odel (MA-LMM), aiming for efficient and effective long-term video modeling. MA-LMM adopts a structure similar to existing large multimodal models [7, 9, 12], which comprise a visual encoder to extract visual features, a querying transformer to align the visual and text embedding spaces, and a large language model. As illustrated in Figure 1(a), as opposed to directly feeding visual encoder outputs to the querying transformer, we opt for an online processing approach that takes video frames sequentially and stores the video features in the proposed long-term memory bank. This strategy of sequentially processing video frames and leveraging a memory bank significantly reduces the GPU memory footprint for long video sequences. It also effectively addresses the constraints posed by the limited context length in LLMs as demonstrated in Figure 1(b). Our design provides a solution for long-term video understanding with large multimodal models with great advantages over prior approaches [7, 9, 12, 13, 21] which consume huge GPU memory and require a large number of input text tokens.

The core contribution of our approach is the introduction of a long-term memory bank that captures and aggregates historical video information. Specifically, the memory bank aggregates past video features in an auto-regressive manner, which can be referenced during subsequent video sequence processing. Also, our memory bank is designed to be compatible with the Q-Former, where it acts as the key and value in the attention operation for long-term temporal modeling. As a result, it can be seamlessly integrated into existing large multimodal models in an off-the-shelf manner to enable long-term video modeling ability. To further enhance efficiency, we propose a memory bank compression method that maintains the length of the memory bank constant relative to the input video length. By selecting and averaging the most similar adjacent frame features, it can preserve all the temporal information while significantly reducing the temporal redundancies in long videos.

We summarize our main contributions as follows:

- We introduce a novel long-term memory bank design to enhance existing large multimodal models, equipping them with long-term video modeling capability.
- Our model significantly reduces the GPU memory usage and addresses LLMs' context length limitations by processing video sequences in an online fashion.
- Our approach has achieved new state-of-the-art performances on various downstreaming video tasks, including long-term video understanding, video question answering, and video captioning.

2. Related Work

Image-language models. Inspired by the success of powerful language models [1–6], recent image-language models tend to incorporate pre-trained language models with image encoders to support the multimodal reasoning ability [7–10, 22]. Flamingo [22] proposes to connect powerful pre-trained vision-only and language-only models and achieve state-of-the-art performance in few-shot learning tasks. BLIP-2 [7] introduces a lightweight querying transformer to bridge the modality gap between the frozen pre-trained image encoder and frozen LLMs. Despite having significantly fewer trainable parameters, it performs well on various multimodal tasks. LLaVA [8] employs a simple linear layer to project image features into the text embedding space and efficiently finetunes LLMs [23] for better performance. Building upon BLIP-2, MiniGPT-4 [10] collects a large-scale high-quality dataset of image-text pairs and achieves better language generation ability. VisionLLM [15] leverages the reasoning and parsing capacities of LLMs, producing strong performance on multiple fine-grained object-level and coarse-grained reasoning tasks.

Video-language models. Current image-language models such as Flamingo [22] and BLIP-2 [7, 9] can also support video inputs. They simply flattened the spatio-temporal features into 1D sequences and then fed them into the language models for video inputs. However, these approaches can not effectively capture the temporal dynamics of videos. Recent video-language models [11–13] attempt to address this by incorporating mechanisms to better understand these dynamics. For instance, Video-LLaMA [12] enhances BLIP-2 structure by adding an additional video querying transformer to explicitly model the temporal relationship. Similarly, building on LLaVA [8], Video-ChatGPT [21] simply average pools the frame-level features across spatial and temporal dimensions to generate video-level representation. ChatVideo [24] treats tracklets as the basic video elements and relies on several video foundation models to annotate these tracklets with textual descriptions. Then, it uses ChatGPT [4] to achieve various video understanding tasks. VideoChat [13] utilizes perception models to generate action and object annotations, which are then forwarded to LLMs for further reasoning. Despite these advancements, these models are primarily designed for short videos, as longer videos pose significant challenges due to language model context length and GPU memory restrictions.

Long-term video models. Long-term video understanding methods focus on capturing long-range patterns in long videos, which typically exceed 30 seconds. To mitigate the computational demands of processing long videos, a

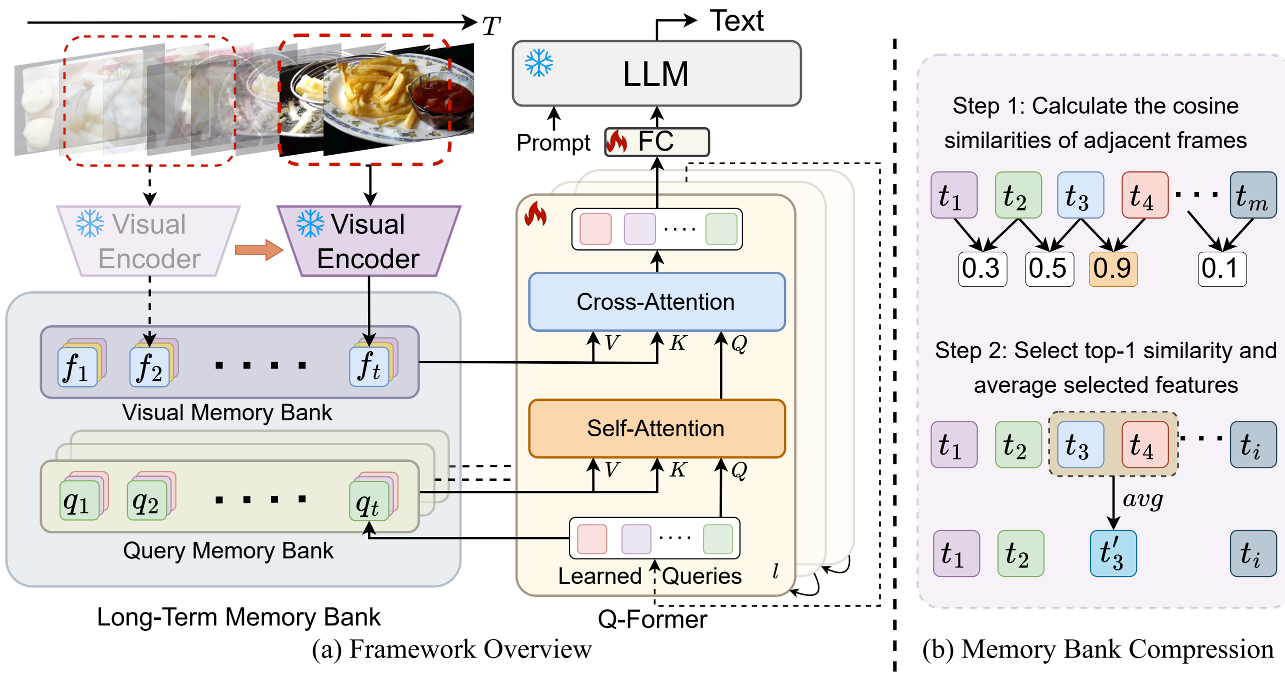


Figure 2. (a) Framework overview. MA-LMM auto-regressively processes video frames in an online manner. Two long-term memory banks are designed to store the raw visual features and learned queries at each timestep, which are used for future reference. The Q-Former is composed of several cascaded blocks, indexed by l. LLM outputs text for various video understanding downstream tasks. The snowflake icon indicates components with fixed parameters, while the flame icon denotes parts of the model that are fine-tuned. (b) Illustration of the memory bank compression technique, which is applied to maintain the length of the memory bank constant.

prevalent approach involves using pre-extracted features, sidestepping the need for joint training of backbone architectures [25–29]. Alternatively, some research works aim to devise sparse video sampling methods [30, 31], reducing the number of input frames by only preserving salient video content. Other works like Vis4mer [32] and S5 [33] leverage the streamlined transformer decoder structure of S4 [34] to enable long-range temporal modeling with linear computation complexity. Inspired by the memory bank design [35–38], we propose to integrate the long-term memory bank with large multimodal models to enable efficient and effective long-term temporal modeling capabilities.

3. Method

We introduce MA-LMM, a memory-augmented large multimodal model for long-term video understanding. Figure 2(a) illustrates the overview of our MA-LMM framework. Following similar practices of large multimodal models [7–9, 12], the overall model architecture can be divided into three parts: (1) visual feature extraction with a frozen visual encoder (Sec. 3.1), (2) long-term temporal modeling with a trainable querying transformer (Q-Former) to align the visual and text embedding spaces (Sec. 3.2), and (3) text decoding with a frozen large language model (Sec. 3.3).

3.1. Visual Feature Extraction

We propose to auto-regressively process video frames in an online manner. This design draws inspiration from the cognitive processes humans use to handle long-term visual information. Instead of concurrently processing extensive duration of signals, humans process them in a sequential manner, correlate current visual inputs with past memories for comprehension, and selectively retain salient information for subsequent reference [38]. Similarly, our MA-LMM processes video frames sequentially, dynamically associating new frame input with historical data stored in the long-term memory bank, ensuring that only discriminative information is conserved for later use. This selective retention facilitates a more sustainable and efficient approach to video understanding, which further allows the model to automatically support online video reasoning tasks.

Formally, given a sequence of T video frames, we pass each video frame into a pre-trained visual encoder and obtain the visual features $V = [v_1, v_2, .., v_T], v_t \in \mathbb{R}^{P\times C}$, where P is the number of patches for each frame and C is the channel dimension for the extracted frame feature. Then we inject temporal ordering information into the frame-level features by a position embedding layer (PE) as

$$f_t = v_t + PE(t), f_t \in \mathbb{R}^{P\times C}. \quad (1)$$

3.2. Long-term Temporal Modeling

For aligning the visual embedding to the text embedding space, we use the same architecture as the Querying Transformer (Q-Former) in BLIP-2 [7, 9]. Q-Former takes in the learned queries $z \in \mathbb{R}^{N\times C}$ to capture video temporal information, where N is the number of learned queries, and C is the channel dimension. In our experiments, Q-Former outputs 32 tokens for each image, which is more efficient than 256 tokens produced by LLaVA [8]. Each Q-Former block consists of two attention submodules: (1) cross-attention layer, which interacts with the raw visual embedding extracted from the frozen visual encoder, and (2) self-attention layer, which models interactions within the input queries. Different from the original Q-Former in BLIP-2 that only attends to the current frame's embedding, we design a long-term memory bank consisting of the visual memory bank and the query memory bank, which accumulates the past video information and augments the input to cross- and self-attention layers for effective long-term video understanding.

Visual Memory Bank. The visual memory bank stores the raw visual features of each frame extracted from the frozen visual encoder. Specifically, for the current time step t, the visual memory bank contains the concatenated list of past visual features $F_t = \texttt{Concat}[f_1, f_2, .., f_t], F_t \in \mathbb{R}^{tP \times C}$. Given the input query z_t, the visual memory bank acts as the key and value as:

$$Q = z_t W_Q,\ K = F_t W_K,\ V = F_t W_V. \tag{2}$$

Then we apply the cross-attention operation as:

$$O = Attn(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V. \tag{3}$$

In this way, it can explicitly attend to past visual information through the cached visual memory bank with long-term context. Since all the cross-attention layers in the Q-Former attend to the same visual feature, there is only one visual memory bank that is shared across all the Q-Former blocks.

Query Memory Bank. Different from the fixed visual memory bank which stores the raw and static visual features, the query memory bank accumulates input queries of each timestep, represented as $Z_t = \texttt{Concat}[z_1, z_2, .., z_t], Z_t \in \mathbb{R}^{tN \times C}$. By storing these queries, we maintain a dynamic memory of the model's understanding and processing of each frame up to the current timestep via the Q-Former. The query memory bank also acts as key and value as:

$$Q = z_t W_Q,\ K = Z_t W_K,\ V = Z_t W_V. \tag{4}$$

similar to the Eq 2. Then we apply the same attention operation as Eq. 3. At each time step, z_t contains the learned important information specifically for each video till the current timestep t. Different from the static visual memory bank, the input queries z_t evolve through cascaded Q-Former blocks during the model training, capturing distinct video concepts and patterns at increasing levels of abstraction. As a result, each self-attention layer has a unique query memory bank, where the contained input queries are updated during the training time.

Memory Bank Compression. Given that our model directly stores past video information in the memory banks, the GPU memory cost and computational complexity increase linearly as the number of past video frames. This becomes particularly challenging for long videos, and thus it is essential to further compress the memory bank to a smaller size. A straightforward solution is to use a first-in-first-out queue to manage the temporal sequence. In this scenario, the features of the earliest timestep are removed if the current length of the memory bank exceeds a pre-defined limit. However, this approach inevitably leads to the loss of earlier historical information as new frames are added and old features are popped when the memory bank is full. To address this issue, we take advantage of the temporal redundancy in long videos and introduce a novel memory bank compression (MBC) technique. This method aggregates and compresses video information over time based on the similarity of adjacent features, while preserving early historical information. In this manner, repetitive information is consolidated in the memory bank, while discriminative features remain intact.

We apply the compression algorithm at each auto-regressive iteration if the current memory bank length is larger than the pre-defined threshold M. Formally, given the visual memory bank containing a list of $[f_1, f_2, .., f_M], f_t \in \mathbb{R}^{P \times C}$, when a new frame feature f_{M+1} comes in, we need to compress the memory bank by reducing the length by 1. At each spatial location i, we first calculate the cosine similarity between all the temporally adjacent tokens as

$$s_t^i = \cos(f_t^i, f_{t+1}^i), t \in [1, M], i \in [1, P]. \tag{5}$$

Then we select the highest similarity across time, which can be interpreted as the most temporally redundant features:

$$k = \text{argmax}_t(s_t^i). \tag{6}$$

Next, we simply average the selected token features at all the spatial locations to reduce the memory bank length by 1:

$$\hat{f}_k^i = (f_k^i + f_{k+1}^i)/2. \tag{7}$$

In this way, we can still preserve the most discriminative features while keeping the temporal ordering unchanged as depicted in Figure 2(b). The same procedure is adopted to compress the query memory bank.

3.3. Text Decoding

As we process video frames in an auto-regressive manner, the Q-Former output at the final timestep contains all historical information, which is then fed into the LLM. Therefore, we can significantly reduce the number of input text tokens from $N * T$ to N, addressing the context length limitation of the current LLMs and substantially easing the GPU memory requirements. During training, given a labeled dataset consisting of video and text pairs, our model is supervised with the standard cross entropy loss as:

$$\mathcal{L} = -\frac{1}{S}\sum_{i=1}^{S} \log P(w_i|w_{<i}, V). \tag{8}$$

in which V represents the input video, and w_i is the i-th ground-truth text token. During training, we update the parameters of the Q-Former while keeping the weights of both the visual encoder and the language model frozen.

4. Experiments

4.1. Tasks and Datasets

To validate the effectiveness of the proposed MA-LMM, we mainly focus on the long-term video understanding task. We also extend the evaluation to standard video understanding tasks (*e.g.*, video question answering, video captioning) to further compare with existing multimodal methods.

Long-term Video Understanding. We conduct experiments on three widely used long-term video datasets including LVU [29], Breakfast [42], and COIN [43]. We report the top-1 classification accuracy as the evaluation metric. The LVU dataset contains ∼30K videos extracted from ∼3K movies, with each video lasting 1 to 3 minutes. Given that current large multimodal models generally perform text generation and lack regression capability, we limit our experiments to seven classification tasks: relationship, speaking style, scene, director, genre, writer, and release year. The Breakfast [42] dataset includes videos related to breakfast preparation, which consists of 1712 videos with an average length of around 2.7 minutes. COIN [43] is a large-scale dataset for comprehensive instructional video analysis, which comprises 11827 instructional videos from YouTube, covering 180 distinct tasks in 12 domains related to daily life. The average length of a video is 2.36 minutes.

Video Question Answering. We conduct evaluation on three open-ended video question answering datasets including MSRVTT-QA [48], MSVD-QA [48], and ActivityNet-QA [49]. ActivityNet-QA contains long videos with average durations of 2 minutes, while MSRVTT-QA and MSVD-QA consist of short videos with 10-15 seconds duration.

Video Captioning. We report the video captioning results of METEOR [50] and CIDEr [51] metrics on three popular datasets: MSRVTT [52], MSVD [53] and Youcook2 [54].

Online Action Prediction. We further evaluate the online prediction capability of our model by conducting experiments on the EpicKitchens-100 [55] dataset, which consists of 700 long videos of cooking activities with 100 total hours. It includes 97 verbs, 300 nouns, and 3807 action types. Following the same experimental setting in [56], we report the top-5 accuracy and recall results on the validation dataset.

4.2. Implementation Details

For the visual encoder, we adopt the pre-trained image encoder ViT-G/14 [57] from EVA-CLIP [58], it can be further changed to other clip-based video encoders. We use the pre-trained Q-Former weights from InstructBLIP [9] and adopt Vicuna-7B [59] as the LLM. All the experiments are conducted on 4 A100 GPUs. More details about training and evaluation are described in the supplementary material.

4.3. Main Results

Long-term Video Understanding. We compare MA-LMM with previous state-of-the-art (SOTA) methods on the LVU benchmark [29] in Table 1. Notably, MA-LMM outperforms existing long-term video models (S5 [33], ViS4mer [32], VideoBERT [41], and Object Transformer [29]) in both content understanding and metadata prediction tasks. This results in significant improvement in most tasks, enhancing the average top-1 accuracy by 3.8% compared to the S5 [33] model. Unlike previous video-based models which process all video frames simultaneously in an offline manner and predict probabilities for each class, our MA-LMM processes video frames in an online fashion and directly outputs the text label for each class type.

We also evaluate our MA-LMM on the Breakfast [42] and COIN [43] datasets that pose a challenge for the long-term video activity classification task. We show the results in Table 2. Our method improves upon the previous best method, S5[33], by 2.3% and 2.4% respectively on the top-1 accuracy metric. This result further proves the superior long-term video understanding capability of our approach.

Video Question Answering. To compare with existing multimodal video understanding methods, we conduct experiments on the open-ended video question answering datasets in Table 3 to demonstrate the generalization ability of our model. Given that these are mostly short videos, it is expected that our memory bank will be less effective. Interestingly, we observe that our MA-LMM achieves new state-of-the-art performances on the MSRVTT and MSVD datasets while falling short of VideoCoCa's performance on the ActivityNet dataset. On the latter, it is not surprising, since VideoCoCa [67] leverages large-scale video-text datasets for pre-training (*e.g.*, HowTo100M [70] and VideoCC3M [71]) while our MA-LMM uses model weights only pre-trained on the image-text datasets.

Notably, our MA-LMM significantly outperforms the recent LLM-based model Video-LLaMA [12] on all three datasets. Video-LLaMA concatenates all the query embeddings from the frozen image Q-Former and trains an additional video Q-Former from scratch to model temporal dependencies, consuming too much GPU memory to be feasible for long video inputs. In contrast, our MA-LMM simply fine-tunes the weights from the pre-trained image Q-

Table 1. Comparison with state-of-the-art methods on the LVU [29] dataset. **Bold** and underline represent the top-1 and top-2 results.

Model	Content			Metadata				Avg
	Relation	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer [29]	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.0
Performer [39]	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer [40]	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT [41]	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
LST [32]	52.5	37.3	62.8	56.1	52.7	42.3	39.2	49.0
VIS4mer [32]	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
S5 [33]	**67.1**	42.1	73.5	67.3	**65.4**	51.3	48.0	59.2
Ours	58.2	**44.8**	**80.3**	**74.6**	61.0	**70.4**	**51.9**	**63.0**

Table 2. Comparison on the Breakfast [42] and COIN [43] datasets. The top-1 accuracy is reported here.

Model	Breakfast	COIN
TSN [44]	-	73.4
VideoGraph [45]	69.5	-
Timeception [28]	71.3	-
GHRM [46]	75.5	-
D-Sprv. [47]	89.9	90.0
ViS4mer [32]	88.2	88.4
S5 [33]	90.7	90.8
Ours	**93.0**	**93.2**

Table 3. Comparison with state-of-the-art methods on the video question answering task. Top-1 accuracy is reported.

Model	MSRVTT	MSVD	ActivityNet
JustAsk [60]	41.8	47.5	38.9
FrozenBiLM [61]	47.0	54.8	43.2
SINGULARITY [62]	43.5	–	44.1
VIOLETv2 [63]	44.5	54.7	–
GiT [64]	43.2	56.8	–
mPLUG-2 [65]	48.0	58.1	–
UMT-L [66]	47.1	55.2	47.9
VideoCoCa [67]	46.3	56.9	**56.1**
Video-LLaMA [12]	46.5	58.3	45.5
Ours	**48.5**	**60.6**	49.8

Table 4. Comparison with state-of-the-art methods on the video captioning task. METEOR (M) and CIDEr (C) results are reported.

Model	MSRVTT		MSVD		YouCook2	
	M	C	M	C	M	C
UniVL [68]	28.2	49.9	29.3	52.8	–	127.0
SwinBERT [69]	29.9	53.8	41.3	120.6	15.6	109.0
GIT [64]	32.9	73.9	**51.1**	**180.2**	17.3	129.8
mPLUG-2 [65]	**34.9**	**80.3**	48.4	165.8	–	–
VideoCoca [67]	–	73.2	–	–	–	128.0
Video-LLaMA	32.9	71.6	49.8	175.3	16.5	123.7
Ours	33.4	74.6	51.0	179.1	**17.6**	**131.2**

Former without introducing an additional video Q-Former, yet is able to effectively capture temporal relationships by virtue of the long-term memory bank. This result strongly justifies the superiority of our design on the general video question answering task, and reveals that even a few frames and queries captured in the memory banks can have significant beneficial effects.

Video Captioning. To further evaluate the capabilities of our MA-LMM in generating free-form text, we conduct experiments on the standard video captioning datasets including MSRVTT [52], MSVD [53] and YouCook2 [54] in Table 4. Although these datasets only consist of videos with short duration and our model is initially pre-trained merely on image-text dataset pairs, our MA-LMM exhibits outstanding performances across all the metrics. It consistently ranks among the top-2 positions compared to current leading methods. Remarkably, our results also surpass the recent Video-LLaMA [12] on these datasets, highlighting the significant improvements our model offers in both video captioning and question-answering tasks.

Online Action Prediction. Since our model can naturally support the online video understanding task, we compare our MA-LMM with Video-LLaMA on the EpicKitchens-100 [55] dataset to investigate the online action prediction capability. In Table 5, our MA-LMM outperforms Video-LLaMA, achieving more accurate results in both top-5 accuracy and recall measures. This highlights our model's superior capacity to anticipate actions in an online manner, showcasing its effectiveness for applications that require real-time analytical capabilities.

Table 5. Action anticipation results on EpicKitchens-100.

Model	Accuracy@Top-5			Recall@Top-5		
	Verb	Noun	Act.	Verb	Noun	Act.
Video-LLaMA	73.9	47.5	29.7	**26.3**	27.3	11.7
Ours	**74.5**	**50.7**	**32.7**	25.9	**29.9**	**12.2**

4.4. Ablation Studies

Contribution of each component. To further investigate the contribution of the visual memory bank and query memory bank, we conduct ablation studies in Table 6. Initially, we observe that without any memory bank module, the performances across all three datasets are notably worse, due to the lack of temporal context. The introduction of either memory bank results in substantial improvements, confirming their roles in enhancing the model's ability to understand tempo-

Table 6. Contribution of visual and query memory banks.

Visual	Query	LVU	Breakfast	COIN
✗	✗	48.3	74.6	72.3
✓	✗	61.5	91.8	92.4
✗	✓	58.0	81.4	88.5
✓	✓	**63.0**	**93.0**	**93.2**

Table 7. Contribution of the long-term memory bank (MB) under off-the-shelf evaluation without training.

MB	MSRVTT	MSVD	ActivityNet	LVU
✗	19.5	38.8	29.9	23.6
✓	**20.3**	**40.0**	**37.2**	**32.8**

ral sequences. We also find that the visual memory bank achieves better performance than the query memory bank. We hypothesize that the explicit method of storing historical raw video features in the visual memory bank is more effective than the query memory bank which implicitly captures video information through the input learned queries. And two memory banks are complementary to each other. When incorporating two memory banks together, our approach can boost the final performance by 14.7%, 18.4%, and 20.9% on the LVU, Breakfast, and COIN, respectively.

Long-term temporal modeling ablation. We compare different temporal modeling approaches in Table 8. In our setup, the Q-Former outputs 32 text tokens per frame. The most straightforward approach for temporal feature integration is either concatenating or averaging frame-level features. However, they resulted in inferior performances. Notably, concatenation requires a significantly higher number of text tokens and computational cost compared to other variants, which also introduces higher GPU memory consumption since they need to takes in all the video frames simultaneously. In addition, we conduct experiments using ToMe [72] to reduce the number of text tokens per frame from 32 to 2. However, without our auto-regressive strategy, it still requires 200 text tokens for 100-frame input. The second part of this table presents the performances of different memory bank compression approaches. The first-in-first-out (FIFO) technique removes the oldest features to main the length of the memory bank fixed, while the memory bank compression (MBC) strategy merges temporally consecutive features with the highest similarity, effectively reducing the most redundant information while keeping the temporal ordering unchanged. With this design that theoretically keeps all historical information, MBC outperforms FIFO by 1.7%, 4.5%, and 2.8% accuracy across three datasets. This experimental result validates the superior efficiency and effectiveness of our approach in modeling long-term temporal information.

Off-the-shelf evaluation. A key advantage of MA-LMM is that our long-term memory bank can be inserted into existing large multimodal models in an off-the-shelf manner, thereby endowing them with effective temporal modeling capabilities without retraining. As presented in Table 7, MA-LMM can consistently boost the final performance when incorporating the long-term memory bank to the baseline method [9]. Particularly, on long-term video datasets like ActivityNet and LVU, MA-LMM can largely improve the results by 7.3% and 9.2%. This highlights the robustness of long-term memory banks in temporal modeling under the off-the-shelf setting.

Table 8. Ablation of different temporal modeling methods.

Method	#Frame	#Token	GPU	LVU	Breakfast	COIN
Concat	60	1920	49.2	62.6	90.4	93.0
Avg Pool	100	32	21.2	57.6	80.6	87.6
ToMe	100	200	22.2	61.5	91.3	91.5
FIFO	100	32	19.1	61.3	88.5	90.4
MBC	100	32	19.1	**63.0**	**93.0**	**93.2**

Table 9. The comparison of using different LLMs.

LLM	MSRVTT	MSVD	ActivityNet	LVU
FlanT5-XL	46.5	57.6	48.2	62.0
Vicuna-7B	**48.5**	**60.6**	**49.8**	**63.0**

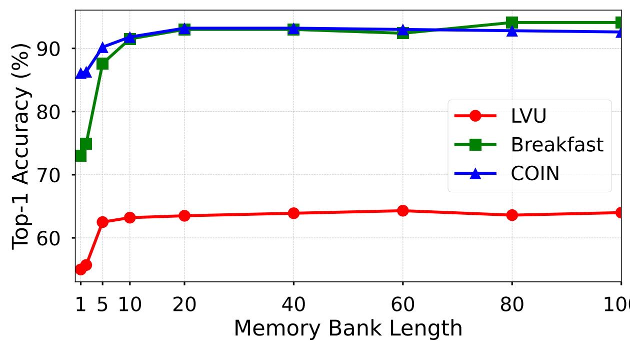
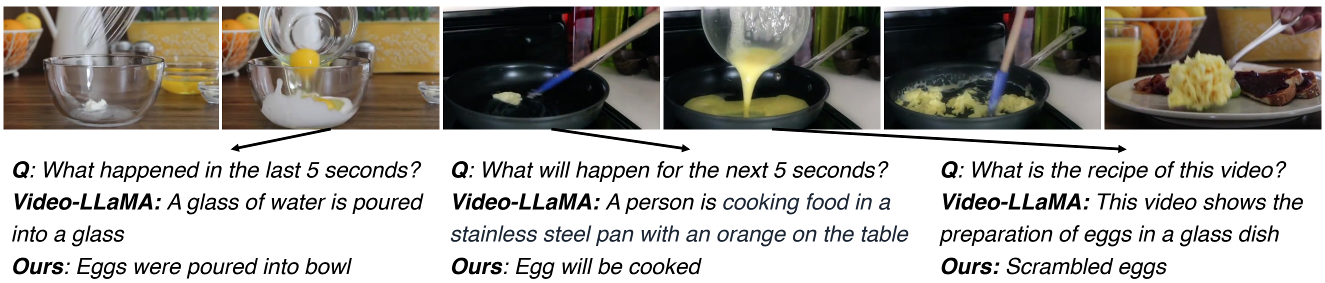


Figure 3. Impact of different memory bank lengths.

Different language model architectures. Our MA-LMM can utilize different language model architectures including but not limited to encoder-decoder models and decoder-only models. We experimented with two popular models FlanT5-XL [73] and Vicuna-7B [59], and show the results in Table 9 that the Vicuna-7B marginally outperforms the FlanT5-XL on these video tasks.

Memory bank length ablation. In Figure 3, we conduct experiments to evaluate the effect of varying the memory bank length. Given an input of 100 video frames, the top-1 accuracy first increases as the feature bank length becomes larger. This rise can be attributed to the augmented storage capacity of the memory bank, which can preserve more

(a) Video question answering Task (ActivityNet-QA)



(b) Online off-the-shelf setting with custom questions

Figure 4. Visualization results on the video question answering task and the online off-the-shelf setting.

Figure 5. Visualization of the compressed visual memory bank.

historical data and consequently boost the final performance. However, we observe that performances begin to saturate when the memory bank length is around 10 to 20. This supports our hypothesis that there are prevalent temporal redundancies in long videos, and we can significantly reduce the frame length without sacrificing the performance.

4.5. Visualization

In Figure 4, we provide a comprehensive visual comparison between MA-LMM and Video-LLaMA [12]. In the video question answering task, our MA-LMM exhibits superior memorization and recognition capabilities. Specifically, it can accurately memorize historical information and recognize fine-grained information, such as the color of the man with No.7, and precisely count the number of goalkeepers who appeared in the video. With the auto-regressive design, our model supports online reasoning directly. This capability is further exemplified in our experiments on off-the-shelf evaluations using custom questions. We can see that, our model can correctly anticipate the next step of the video (*"egg will be cooked"*) and predict the correct recipe of the video (*"scrambled egg"*). More visualization examples are shown in the supplementary material.

Figure 5 provides a visualization of the compressed visual memory bank. We set the memory bank length to 5 for this illustration. The compressed visual memory bank appears to group consecutive frames with similar visual content. For instance, in the presented video, the video frames are effectively grouped into five clusters, each capturing a distinct yet semantically consistent activity, which is similar to the effect of temporal segmentation.

5. Conclusion

In this paper, we introduce a long-term memory bank designed to augment current large multimodal models, equipping them with the capabilities to effectively and efficiently model long-term video sequences. Our approach processes video frames sequentially and stores historical data in the memory bank, addressing LLMs' context length limitation and GPU memory constraints posed by the long video inputs. Our long-term memory bank is a plug-and-play module that can be easily integrated into existing large multimodal models in an off-the-shelf manner. Experiments on various tasks have demonstrated the superior advantages of our method. We believe our MA-LMM offers valuable insights for future research in the long-term video understanding area.

Acknowledgements. This project was partially funded by NSF CAREER Award (#2238769) to AS.

References

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018. 1, 2

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[3] 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.

[4] OpenAI. Chatgpt. *https://openai.com/blog/chatgpt*, 2023. 1, 2

[5] 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

[6] 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. 1, 2

[7] 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, 3

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 4

[9] 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, 2023. 1, 2, 3, 5, 7

[10] 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. 2

[11] 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. 2

[12] 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. 2, 3, 5, 6, 8

[13] 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

[14] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 1

[15] 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. 2

[16] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[17] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, 2022.

[18] Junke Wang, Xitong Yang, Hengduo Li, Li Liu, Zuxuan Wu, and Yu-Gang Jiang. Efficient video transformers with spatial-temporal token selection. In *ECCV*, 2022.

[19] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *arXiv preprint arXiv:2303.12079*, 2023.

[20] Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Omnivid: A generative framework for universal video understanding. In *CVPR*, 2024. 1

[21] 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

[22] 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. 2

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[24] 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. 2

[25] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3

[26] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[27] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal

aggregation for action classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017.

[28] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 6

[29] 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. 3, 5, 6

[30] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019. 3

[31] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 3

[32] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 3, 5, 6

[33] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023. 3, 5, 6

[34] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[35] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10337–10346, 2020. 3

[36] Sangho Lee, Jinyoung Sung, Youngjae Yu, and Gunhee Kim. A memory network approach for story-based temporal summarization of 360 videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1419, 2018.

[37] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3054–3063, 2021.

[38] 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. 3

[39] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 6

[40] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 6

[41] 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. 5, 6

[42] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 5, 6

[43] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 5, 6

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 6

[45] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*, 2019. 6

[46] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. Graph-based high-order relation modeling for long-term action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8993, 2021. 6

[47] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 6

[48] 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. 5

[49] 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. 5

[50] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evalu-

ation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[52] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5, 6

[53] David L 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-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011. 5, 6

[54] 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. 5, 6

[55] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 5, 6

[56] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023. 5

[57] 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. 5

[58] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 5

[59] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 5, 7

[60] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021. 6

[61] 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

[62] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 6

[63] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909, 2023. 6

[64] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 6

[65] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*, 2023. 6

[66] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023. 6

[67] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022. 5, 6

[68] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 6

[69] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 6

[70] 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. 5

[71] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022. 5

[72] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 7

[73] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 7