Streaming Long Video Understanding with Large Language Models

Rui Qian[1,2] **Xiaoyi Dong**[2] **Pan Zhang**[2] **Yuhang Zang**[2] **Shuangrui Ding**[1,2]
Dahua Lin[1,2] **Jiaqi Wang**[2*]
[1] The Chinese University of Hong Kong
[2] Shanghai AI Laboratory

Abstract

This paper presents **VideoStreaming**, an advanced vision-language large model (VLLM) for video understanding, that capably understands arbitrary-length video with a constant number of video tokens streamingly encoded and adaptively selected. The challenge of video understanding in the vision language area mainly lies in the significant computational burden caused by the great number of tokens extracted from long videos. Previous works rely on sparse sampling or frame compression to reduce tokens. However, such approaches either disregard temporal information in a long time span or sacrifice spatial details, resulting in flawed compression. To address these limitations, our VideoStreaming has two core designs: Memory-Propagated Streaming Encoding and Adaptive Memory Selection. **The Memory-Propagated Streaming Encoding architecture** segments long videos into short clips and sequentially encodes each clip with a propagated memory. In each iteration, we utilize the encoded results of the preceding clip as historical memory, which is integrated with the current clip to distill a condensed representation that encapsulates the video content up to the current timestamp. This method not only incorporates long-term temporal dynamics into the streaming encoding process but also yields a fixed-length memory as a global representation for arbitrarily long videos. After the encoding process, the **Adaptive Memory Selection strategy** selects a constant number of question-related memories from all the historical memories, and feeds them into the LLM to generate informative responses. The question-related selection reduces redundancy within the memories, enabling efficient and precise video understanding. Meanwhile, the disentangled video extraction and reasoning design allows the LLM to answer different questions about a video by directly selecting corresponding memories, without the need to encode the whole video for each question. Through extensive experiments, our model achieves superior performance and higher efficiency on long video benchmarks, showcasing precise temporal comprehension for detailed question answering.

1 Introduction

The evolution of Large Language Models (LLMs) has significantly advanced artificial intelligence, encompassing text generation and reasoning in complex language environments [9, 81, 14, 67, 16, 75, 2, 77]. Later, the community extends LLMs to multi-modal domains, demonstrating promising results in captioning and question-answering tasks that integrate diverse visual signals [49, 44, 15, 65]. Yet, within the domain of video understanding, long video sequences pose a formidable challenge. Incorporating such long visual contents into LLMs requires a substantial number of tokens, which not only amplifies computational demands but also risks early contextual information loss [52].

*Corresponding Author

Preprint. Under review.

Among the recent works on general video understanding with LLMs [51, 45, 47, 95, 53, 46, 72, 68], a prevalent strategy is using sparse temporal sampling [47, 93] or spatio-temporal pooling [53, 51] to reduce tokens. Unfortunately, this paradigm explicitly loses substantial information in the long time span. To address this limitation, [46, 45, 95] develop frame-wise compression, with LLaMA-VID [46] as a typical example. It compresses each frame into only two tokens but overlooks the inter-frame temporal dynamics which are vital in compressing temporal redundancy within videos. Besides, its question-dependent compression pipeline limits the ability to produce a general representation that can handle diverse instructions. Another line of works employ memory banks [84, 7] to store history information [72, 26]. Whereas, these methods rely on explicit timestamps to recall the historical details, limiting the ability to generate comprehensive responses without specific time indicators.

In this work, we propose VideoStreaming, a novel Memory-Propagated Streaming Encoding architecture with Adaptive Memory Selection to sequentially encode a long video into condensed memories and generate responses referring to relevant timestamps. The core idea behind the memory-propagated streaming encoding is to preserve representative spatial cues and temporal dynamics while reducing temporal redundancy in videos. To achieve this goal, we segment the long video into multiple short clips and sequentially encode each clip. When encoding each clip, we first refer to the encoded results of its preceding clip as historical memory, then concatenate it with the current clip features and feed them into a small decoder-only language model [25]. Due to its autoregressive nature, the information of the sequence naturally accumulates to the last few tokens [42, 39]. Consequently, we take these last few tokens as an updated memory that encapsulates the video information up to the current timestamp. Through this streaming encoding, we explicitly take long-term temporal relations into consideration and maintain a fixed-length memory to represent an arbitrarily long video.

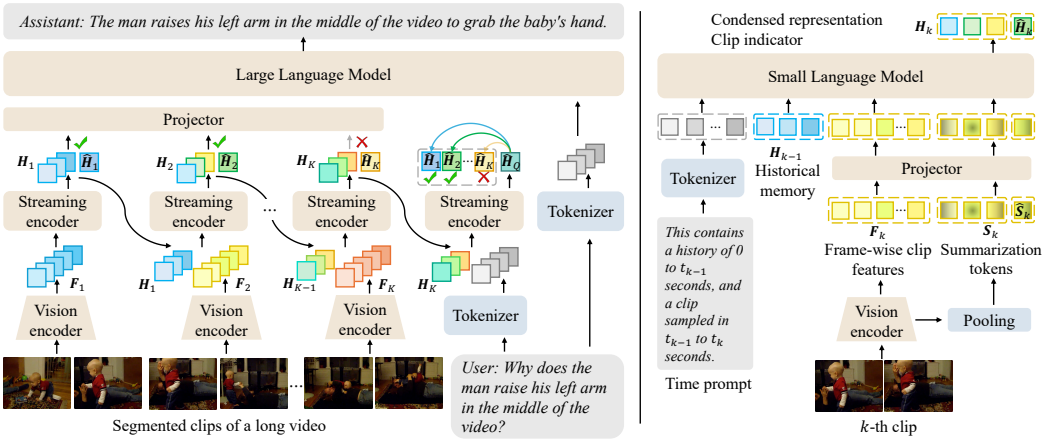
However, this fixed-length memory inevitably loses detailed information, especially in early contexts. To address this problem, we store the historical memories of all clips and select a constant number of subsets that are closely related to the question. To accomplish this, when streaming encoding each clip, we additionally append a summary token at the end of the sequence as a clip indicator that summarizes the clip contents within one token. Then, given a specific question, we concatenate the condensed memory from the final iteration with the question and pass it through the same small language model used in streaming encoding. We take the final token as the question indicator and calculate its similarity with all historical clip indicators, the clip indicator with higher similarity means its corresponding memory is more related to the question. Finally, we feed the adaptively selected memories into the LLM for detailed question answering.

In practice, we realize our VideoStreaming with a carefully designed two-stage progressive training process and long-video data construction strategy. In the first stage, we empower a small language model with the single-clip encoding capability by a specialized prefix task. In the second stage, it serves as the streaming encoder and we jointly train it with the LLM for long video understanding. Due to the lack of long video QA data, we manually constructed a set of long video QA pairs in two ways. On the one hand, we concatenate short videos from existing datasets [83, 88] into longer ones, where the original questions correspond to different segments. On the other hand, we curate a subset of Panda-70M [11] which includes captions for segmented clips as well as the original long videos, and use this to create multi-round long video QA pairs with explicit timestamps. These long video QA data not only optimize the responses from the LLM but also guide accurate memory selection.

In summary, our contributions are as follows: (1) We analyzed the challenge of long video understanding in the vision language area, and pointed out that the problem of current methods lies in the inefficient video encoding. (2) In response to the challenges, we propose two efficient designs: Memory-Propagated Streaming Encoding and Adaptive Memory Selection, which result in our advanced video understanding model VideoStreaming. (3) The extensive experiments demonstrate that our model achieves precise temporal grounding with respect to specific questions, attains superior performance, and exhibits higher inference efficiency on long video benchmarks.

2 Related Work

Large Language Models (LLMs) have revolutionized natural language processing. Early works establish encoder-decoder models with masked language modeling [16, 67], while later decoder-only models like GPT [66] showcase remarkable performance and scalability. Recent groundbreaking works, such as PaLM [14], LLaMA [77] and GPT-4 [59], have pushed the boundaries by developing



(a) An overview of streaming video understanding framework. (b) Details of the streaming encoder.

Figure 1: Fig. 1a shows an overview of VideoStreaming, where we segment a long video into short clips and iteratively encode each clip into compact memories. Then, according to specific questions, we select a constant number of subsets of relevant memories as input to an LLM to produce responses. The ✓ and ✗ respectively denote selected and unselected memories. Fig. 1b illustrates the detailed process of each streaming encoding iteration. We encode current clip features with reference to specific timestamps and historical memory from the preceding clip into a condensed representation.

significantly larger models with billions of parameters. To harness the full potential of LLMs, a series of works [58, 60, 13] adopt supervised instruction tuning [81] to guide models towards generating more natural and contextually relevant responses. Inspired by the powerful reasoning capacities of LLMs, we explore using LLMs for challenging long video understanding.

Vision Language Models like CLIP [65] employ contrastive learning on image-text pairs to formulate a unified embedding space [65, 33, 44]. Later, [49, 43, 59, 97, 3, 4, 92, 96, 20] integrate image features into LLMs and achieve promising visual reasoning in image domain. Considering video as a prevalent visual signal [17, 19, 22, 18, 64, 62, 76, 63], some works further expand the application to process more complex spatio-temporal video data. [53, 51, 47, 29, 93] use sparse sampling or simple temporal pooling to obtain compact video tokens for LLMs. [45, 95] employ Q-Former [43] to project frame-wise features into the textual space. To handle longer videos, [36, 82] utilize token merging [8] to reduce redundancy and alleviate computational burden. LLaMA-VID [46] proposes an instruction-aware compression strategy to represent each frame with only two tokens, but it overlooks the temporal relations in the compression step. [72, 26] develop memory banks to accumulate information in long videos and excel in global video comprehension. However, these methods struggle with moment-specific questions without explicit time indicators. To address these limitations, we propose a memory-propagated streaming encoding architecture with adaptive memory selection, which effectively reduces temporal redundancy and accurately selects relevant information for detailed question answering.

Long Video Understanding is a challenging task in computer vision. The most prevalent strategy is to maintain a memory bank to store history information in long videos [84, 7, 85, 57, 12, 78]. To ensure computation efficiency, it is crucial to compress the history into a finite-length memory, which is typically done by parametric [85] or non-parametric [7] compression modules. More recently, [37, 31, 94] use language as a bridge for long-term video understanding. They first divide a long video into short clips, generate textual descriptions for each clip, and then employ an LLM to aggregate the short captions for long video analysis. However, this architecture cannot be trained end-to-end, and the long video understanding quality depends on the short clip captions. In contrast, we employ a trainable small language model to iteratively encode short clips into compact memories, which can be jointly optimized with the subsequent LLM on long video understanding tasks.

3 VideoStreaming

In this section, we introduce VideoStreaming, a streaming long video understanding framework with LLM. As illustrated in Fig. 1a, given a long video input, VideoStreaming segments it into

multiple short clips and iteratively encodes each clip into compact historical memory. To enhance the reasoning ability to specific questions, we design an adaptive memory selection strategy to select a subset of relevant memories and feed them into an LLM to produce detailed responses.

3.1 Single Clip Encoding

To effectively distill the information within a sequence into a compact set of tokens, we take inspiration from recent advanced decoder-only language models [2, 77, 13, 5, 34] and employ a comparatively small language model, Phi-2 [25], for efficient encoding. Due to the causal attention and autoregressive nature, the language model spontaneously aggregates the sequence information onto the last few tokens [42, 39], which naturally serve as a compact representation that provides a high-level summary of the input sequence.

Mathematically, given a T-frame video clip, we first use a pre-trained CLIP ViT-L [65] to extract frame-wise features and concatenate every four spatially adjacent visual tokens along channel dimension to reduce the number of tokens by 75%. The resulting clip features are denoted as $\boldsymbol{F} \in \mathbb{R}^{TN \times C}$, where N denotes the per-frame spatial token number, and C is the channel dimension. To produce the condensed representations, we initialize a set of summarization tokens $\boldsymbol{S} \in \mathbb{R}^{TP \times C}$ by adaptively pooling each frame into P tokens, where $P \ll N$. Intuitively, $\boldsymbol{S}$ can be regarded as a coarse encapsulation of the given clip, making it well-suited to serve as the summarization tokens for consolidating the clip information. To this end, we concatenate $\boldsymbol{F}$ with $\boldsymbol{S}$ and feed them into the encoder $g(\cdot)$, which consists of an MLP projector and a language model Phi-2. We utilize the output of the last $T \times P$ tokens as the condensed representation of the given clip:

$$\boldsymbol{H} = g([\boldsymbol{F} \circ \boldsymbol{S}]) \in \mathbb{R}^{TP \times D}, \tag{1}$$

where $\circ$ denotes concatenation operation, D is the channel dimension of Phi-2.

To reinforce the visual consolidation ability, we design a prefix task to train the encoder on visual captioning and question-answering tasks. In particular, to guarantee that the clip information is distilled into the summarization tokens, we enforce the language model to generate the response only with reference to these few tokens. To achieve this goal, a straightforward way is to modify the attention mask in each Transformer decoder layer. As depicted in Fig. 2, we take a sequence covering TN clip feature tokens, TP summarization tokens, and TT text response tokens as an example. Based on the standard causal attribute, the binary attention mask $\mathbf{M}$ is modified as shown in Figure 3: with the modified attention mask, the TT text tokens can only get video-related information from the

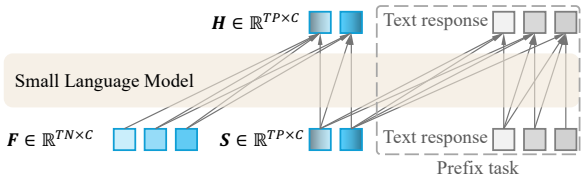


Figure 2: Illustration of the prefix task format.

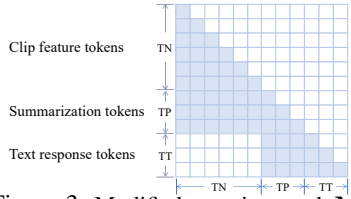


Figure 3: Modified attention mask $\mathbf{M}$.

TP summarization tokens to predict the next token. This encourages the summarization tokens to extract more video information from previous TN video clip tokens, *ie.* learns better video encoding.

3.2 Memory-Propagated Streaming Long Video Encoding

Till this point, we have obtained an encoder capable of distilling short video clips into condensed representations. The next step is to comprehensively consider the long-term temporal relations within the complete videos, leveraging the historical information from previous clips to facilitate the encoding of subsequent segments as depicted in Fig. 1b.

To accomplish this objective, we divide a long video into K clips, each containing T frames, and propose a memory-propagated streaming encoding mechanism to iteratively encode each clip in sequence. In each iteration, we employ the encoded results from the last iteration as historical memory and integrate them with current clip features to produce an updated memory for subsequent encoding. Specifically, given the k-th clip, we denote the current clip features as $\boldsymbol{F}_k \in \mathbb{R}^{TN \times C}$, the summarization tokens as $\boldsymbol{S}_k \in \mathbb{R}^{TP \times C}$, and an additional global token as $\hat{\boldsymbol{S}}_k \in \mathbb{R}^{1 \times C}$. This global token, initialized by global average pooling on the clip features $\boldsymbol{F}_k$, is expected to summarize

the entire clip contents and serve as a clip indicator for memory selection in the next subsection. To enrich the temporal contexts, we refer to the encoded representations from the previous clip $\boldsymbol{H}_{k-1} \in \mathbb{R}^{TP \times D}$ to provide historical information. Then we jointly feed them into the streaming encoder to produce the condensed representation $\boldsymbol{H}_k \in \mathbb{R}^{TP \times D}$ and the clip indicator $\hat{\boldsymbol{H}}_k \in \mathbb{R}^{1 \times D}$ of the k-th clip:

$$\boldsymbol{H}_k, \hat{\boldsymbol{H}}_k = g([\boldsymbol{H}_{k-1} \circ \boldsymbol{F}_k \circ \boldsymbol{S}_k \circ \hat{\boldsymbol{S}}_k]). \tag{2}$$

Note that for the first clip encoding, the historical memory is not used. Through this streaming encoding process, $\boldsymbol{H}_k$ not only encompasses the current clip information but encapsulates the overall video content up to the k-th clip. To this end, we manage to maintain a fixed length of memory to represent arbitrarily long videos.

Discussion. In this architecture, we use a language model for video encoding, which has the unique advantage that we can flexibly provide the encoder with diverse prompts to guide the encoding process. Hence, the summarization tokens capture not only the core content but also additional contextual information. Typically, the explicit timestamp is an important cue in videos [68]. As shown in Fig. 1b, we incorporate a text prompt indicating the specific timestamps of each clip and historical memory to enhance temporal awareness. Besides, this prompt-based approach also allows the user to tailor the condensed output to better suit the needs of downstream tasks, going beyond a purely extractive summarization.

Another noteworthy point is that in the language model, the feature space of the final decoder layer is designed for the next token prediction, which may not perfectly align with the objective of producing condensed video representations. Considering that we modify the attention masks in each decoder layer to encourage information consolidation, this allows us to leverage the intermediate outputs from partial attention layers as the encoded results. Similar to the techniques in vision domain [89, 49, 21], this strategy potentially enables the model to capture a richer set of semantic and contextual features as the condensed representations, bridging the gap between the language model's original training objective and the requirements for video encoding.

3.3 Adaptive Memory Selection

Through the streaming video encoding, it is feasible to use the encoded results from the final iteration, i.e., $\boldsymbol{H}_K$, as a compact global memory that concludes the entire video. However, this fixed-length memory inevitably loses details, especially the information from early segments. Hence, this global memory alone is insufficient for comprehensive long video understanding.

To address this limitation, we make use of the encoded results of all historical clips of the input video, i.e., $\mathcal{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_K\}$. Given a specific question or instruction, we first generate an adaptive indicator that summarizes relevant video content for that particular instruction. We accomplish this by reusing the language model in the streaming encoder, where we concatenate the global memory from the final iteration, $\boldsymbol{H}_K$, and the instruction texts, then pass the sequence into the model. We employ the output of the final token as the instruction indicator, denoted as $\hat{\boldsymbol{H}}_Q \in \mathbb{R}^{1 \times D}$. Thereafter, we calculate the cosine similarity between this instruction indicator and all historical clip indicators $\{\hat{\boldsymbol{H}}_1, \hat{\boldsymbol{H}}_2, ..., \hat{\boldsymbol{H}}_K\} \in \mathbb{R}^{K \times D}$ and obtain the similarity distribution $\boldsymbol{s} \in \mathbb{R}^K$. To achieve a differentiable discrete selection, we develop a variant of Gumbel-Softmax [32], denoted as Gumbel-Topk$(\cdot)$, to produce a binary index $\mathbf{I}$ that activates a subset of V out of K positions with the highest similarities:

$$\mathbf{I} = \text{Gumbel-Topk}(\boldsymbol{s}, V) \in \{0, 1\}^K. \tag{3}$$

Based on $\mathbf{I}$, we select the corresponding encoded results from $\mathcal{H}$ to formulate a subset of memories that are related to the instruction:

$$\hat{\mathcal{H}} = \{\mathbf{I}_k \cdot \boldsymbol{H}_k | \mathbf{I}_k = 1\}, \tag{4}$$

where $\mathbf{I}_k$ denotes the selected indexes. We concatenate the selected memories $\hat{\mathcal{H}}$ in temporal order, resulting in a sequence consisting of $V \times T \times P$ tokens. Then, we feed the sequence with instruction texts into an LLM for comprehensive reasoning.

Our adaptive memory selection allows the model to dynamically access historical memories relevant to specific instructions, which mitigates the information loss inherent in the streaming encoding process. By drawing upon fine-grained details across the full video duration, the LLM can provide detailed and informative responses, while preserving high computational efficiency.

Table 1: The statistics of the average video duration time of each evaluation dataset.

Dataset	Duration
Next-QA [86]	42.23 sec
Next-GQA [87]	39.60 sec
VideoChatGPT [53]	1.81 min
EgoSchema [54]	3.00 min
MovieChat-1K [72]	7.66 min
MovieNet-QA [74]	108.26 min

Table 2: Results on VideoChatGPT benchmark [53].

Method	Params	CI	DO	CU	TU	CO
Video-LLaMA [95]	7B	1.96	2.18	2.16	1.82	1.79
VideoChat [45]	7B	2.23	2.50	2.53	1.94	2.24
VideoChatGPT [53]	7B	2.40	2.52	2.62	1.98	2.37
MovieChat [72]	7B	2.76	2.93	3.01	2.24	2.67
LongVLM [82]	7B	2.76	2.86	3.34	2.39	3.11
LLaMA-VID [46]	13B	3.07	3.05	3.60	2.58	2.63
PLLaVA [90]	13B	3.27	2.99	3.66	2.47	3.09
Ours	7B+1.3B	**3.33**	**3.27**	**3.73**	**2.74**	**3.15**

3.4 Progressive Training

To train VideoStreaming, we design a progressive two-stage paradigm. First, we train single clip encoding on image and short video understanding tasks. Next, we train memory-propagated streaming encoding and adaptive memory selection as well as the LLM for long video understanding.

Single Clip Training. In this stage, both image- and video-text pairs are used to train the encoder to handle general visual signals. Following [47, 53, 95, 46], we employ 790K image and short video caption data [70, 6] to train the MLP projector for modality alignment. After that, we employ 763K image and video instruction data from [49, 53, 48] to finetune the small language model. For video input, we uniformly sample $T = 16$ frames with spatial resolution 224×224 and use a frozen CLIP ViT-L/14 [65] to extract frame-wise features. After adjacent token merging, we obtain $16 \times 64 = 1024$ tokens as the clip feature representation. Then, the encoder, a two-layer MLP and a small language model Phi-2 2.7B [25], distills each frame into $P = 4$ tokens, resulting in $16 \times 4 = 64$ tokens as the condensed representation with a compression ratio of $16 : 1$. For image-text pairs, we regard the images as single-frame clips and encode each into 4 tokens. We use standard next token prediction to consolidate visual contents into compact summarization tokens as illustrated in Fig. 2.

Streaming Long Video Training. In the second stage, we use long video QA pairs to finetune the whole architecture, including ViT, the streaming encoder, and the LLM, as shown in Fig. 1a. The long video QA data encompasses three parts. (1) We adopt 25K movie QA pairs from [46, 28, 72]. (2) We curate a subset from Panda-70M [11], which provides the original long videos and the captions of segmented clips. Based on this subset, we create 300K multi-round long video QA pairs with explicit timestamps. (3) We synthesize 20K long videos by concatenating short videos from existing QA datasets [88, 83], and the original QA pairs correspond to different segments in the synthesized long videos. For each video, we extract 16-frame clips at 1 FPS, and the number of clips varies with the video duration. In streaming encoding, we employ the intermediate outputs from the first 16 layers of Phi-2 as the condensed memories. Finally, we select $V = 4$ most relevant timestamps and feed the selected memories of $V \times T \times P = 256$ tokens into the LLM, Vicuna-7B [13], for long video reasoning. Since our curated long video data could provide pseudo temporal grounding labels of specific questions, we utilize 30K QA pairs to warm up memory selection via a KL divergence loss. Subsequently, we use the rest 315K QA pairs to optimize the responses from the LLM and guide memory selection in a weakly-supervised manner. More training details are included in Appendix A.

4 Experiments

4.1 Datasets

We evaluate our model on long video QA datasets and present the statistics on the temporal duration of individual datasets in Table. 1. Among them, Next-QA [86], Next-GQA [87] and VideoChatGPT [53] encompass minute-long videos with thousands of frames. EgoSchema [54] contains over 5K three-minute videos with multiple-choice questions. Each question has a long temporal certificate, requiring more than 100 seconds within a video to produce a correct answer. MovieChat-1K [72] and MovieNet-QA [74] consist of around ten-minute-long or even hour-long movies, posing significant challenges for the model to comprehend the visual contents across such long time spans.

Table 3: Results on the fullset test split of EgoSchema [54].

Method	Params	Fullset
finetuned		
MC-ViT-L [7]	424M	44.4
LongViViT [61]	1B	33.3
zero-shot		
InternVideo [80]	478M	32.1
FrozenBiLM [91]	890M	26.9
SeViLA [93]	4B	22.7
LLoVi [94]	7B	33.5
Vamos [79]	13B	36.7
LangRepo [37]	7B	38.9
LangRepo [37]	8×7B	41.2
Ours	7B+1.3B	**44.1**

Table 4: Results on the validation set of Next-QA [86]. C, T, D denotes causal, temporal and descriptive splits.

Method	Params	C	T	D	All
finetuned					
BLIP-2 [43]	4B	70.1	65.2	80.1	70.1
LLaMA-VQA [40]	7B	72.7	69.2	75.8	72.0
Vamos [79]	7B	72.6	69.6	78.0	72.5
zero-shot					
InternVideo [80]	478M	43.4	48.0	65.1	49.1
SeViLA [93]	4B	61.3	61.5	75.6	63.6
Mistral [34]	7B	51.0	48.1	57.4	51.1
LLoVi [94]	7B	55.6	47.9	63.2	54.3
LangRepo [37]	7B	57.8	45.7	61.9	54.6
LangRepo [37]	8×7B	64.4	51.4	69.1	60.9
Ours	7B+1.3B	**65.1**	**62.2**	**78.1**	**66.2**

Table 5: Results on Next-GQA [87]. Acc@GQA is defined as the percentage of questions that are both correctly answered and visually grounded with IoP ≥ 0.5.

Method	Params	mIoP	IoP@0.5	mIoU	mIoU@0.5	Acc@GQA
w/ specialized grounding module						
TempCLIP [65, 87]	130M	25.7	25.5	12.1	8.9	16.0
SeViLA [93]	4B	29.5	22.9	21.7	13.8	16.6
w/o specialized grounding module						
LLoVi [94]	7B	20.7	20.5	8.7	6.0	11.2
LangRepo [37]	7B	20.3	20.0	8.7	6.0	11.2
LangRepo [37]	8×7B	31.3	28.7	18.5	12.2	17.1
Ours	7B+1.3B	**32.2**	**31.0**	**19.3**	**13.3**	**17.8**

4.2 Main Results

In this section, we present the results of our 8.3B model (half of Phi-2 2.7B in streaming encoder and Vicuna-7B as the LLM). We omit the comparisons to proprietary LLMs.

VideoChatGPT. Table 2 presents the results on VideoChatGPT [53] in terms of Correctness of Information (CI), Detailed Orientation (DO), Contextual Understanding (CU), Temporal Understanding (TU) and Consistency (CO). Our model outperforms LLM-based video understanding methods on all five metrics, with a significant advantage in temporal understanding. It can be attributed to the memory-propagated streaming encoding architecture that explicitly captures temporal dynamics.

EgoSchema. In Table 3, we report the *zero-shot* performance on the fullset test split of EgoSchema [54]. MC-ViT [7] consolidates a long-term memory to memorize long contexts but requires finetuning on related dataset [24]. LLM-based methods [94, 37, 79] curate answers from the captions of segmented video clips. However, these short-term captions cannot be optimized end-to-end and inevitably lose some detailed information. In contrast, we use a trainable streaming encoder to produce memory embeddings in long videos and feed them into an LLM to generate responses. Our model outperforms all zero-shot methods and is comparable to the finetuned MC-ViT, demonstrating the effectiveness of our streaming architecture for long-term temporal modeling.

Next-QA. In Table 4, we perform *zero-shot* evaluation on the validation split of Next-QA [86] covering 5K multiple-choice questions. We respectively report the accuracy on Causal (C), Temporal (T) and Descriptive (D) subsets. Our method consistently surpasses all zero-shot counterparts. Typically, compared to LangRepo [37] with Mixtral-8×7B [35], our 8.3B model improves the causal, temporal, and descriptive accuracy by 0.7%, 10.8%, 9.0% with considerably fewer model parameters.

Next-GQA. Besides the evaluation of the generated responses, we also assess the temporal grounding ability on Next-GQA [87]. We calculate the Intersection of Prediction (IoP) and Intersection of Union (IoU), and use Acc@GQA to measure the accuracy of the correctly grounded predictions. According to the comparisons in Table 5, our simple similarity score based selection achieves the highest IoP and comparable IoU to SeViLA [93] with a specialized grounding module. Moreover, the highest Acc@GQA demonstrates the comprehensive capacity for grounding and high-level understanding.

Table 6: Results on MovieChat-1K [72] global and breakpoint mode accuracy (Acc.) and score.

Method	Global		Breakpoint	
	Acc.	Score	Acc.	Score
VideoChat [45]	57.8	3.00	46.1	2.29
Video-LLaMA [95]	51.7	2.67	39.1	2.04
VideoChatGPT [53]	47.6	2.55	48.0	2.45
MovieChat [72]	62.3	3.23	48.3	2.57
MovieChat+ [73]	71.2	3.51	49.6	2.62
Ours	**90.4**	**4.42**	**54.9**	**2.80**

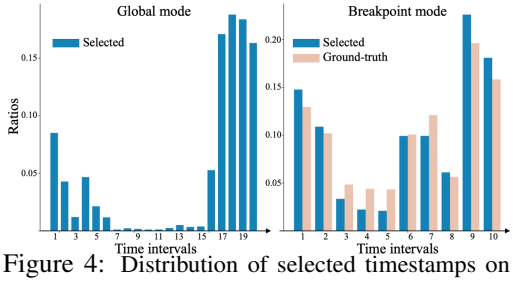


Figure 4: Distribution of selected timestamps on MovieChat-1K. We divide each video into multiple time intervals for statistical analysis.

Table 7: Results on MovieNet-QA [74]. We present the used modality, the average number of tokens input to LLM and the average inference latency per question for comprehensive comparison.

Method	Text	Vision	Tokens	Latency	Overview	Plot	Temporal
LLaMA-VID [46]	✓	✓	18430	16.03 sec	3.09	3.31	2.02
MovieLLM [74]	✓	✓	18430	16.48 sec	3.22	3.38	2.18
LLaMA-VID [46]	✗	✓	5477	10.47 sec	2.28	2.88	1.46
MovieLLM [74]	✗	✓	5477	10.43 sec	2.36	2.97	1.58
Ours	✗	✓	256	5.32 sec	**2.65**	**3.13**	**1.88**

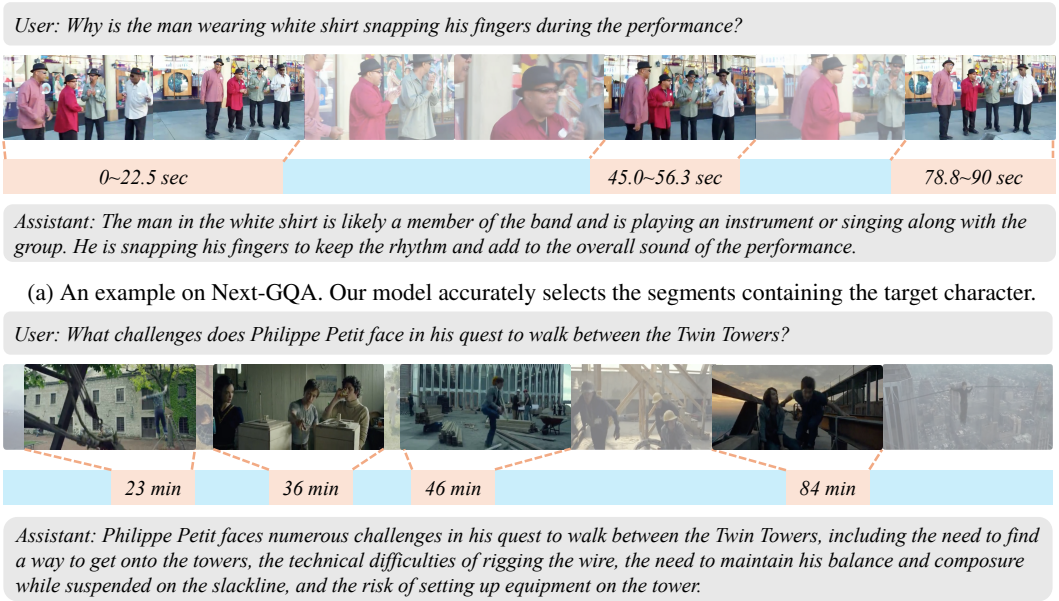
MovieChat-1K. Table 6 shows the results on MovieChat-1K [72], including a global mode for overall long-term understanding and a breakpoint mode for detailed analysis of specific moments. In breakpoint mode, [72, 73] manually extract segments according to the timestamps in questions, while our model adaptively selects the related historical memories. Fig. 4 reveals that our selected timestamps are close to the ground-truths, and the higher breakpoint accuracy validates our adaptive selection effectively gathers the desired information from long contexts. Meanwhile, we reach significantly superior results in global mode, with the model's selection concentrated at the beginning and ending parts. On the one hand, the beginning of a movie often contains hints of global information while the middle comprises redundant details. On the other hand, the condensed memories near the end of the video encapsulate the entire video, making them quite suitable for global understanding.

MovieNet-QA. Finally, we show the results on MovieNet-QA [74] consisting of 100 hour-long movies. Inspired by [74], we use GPT-3.5 to produce scores in range 0-5 to evaluate the performance in overview, plot, and temporal understanding in Table 7. Specifically, LLaMA-VID [46] compresses each frame into two tokens, which are then combined with movie subtitles as input to an LLM. MovieLLM [74] further incorporates more generated data in training. These approaches largely rely on the texts for movie understanding, and only using visual frames leads to dramatic performance drop. Moreover, its frame-wise compression is dependent on specific questions. The model has to reprocess the entire movie to extract visual features for different questions, resulting in a high inference latency of over 10 seconds per question. Conversely, our architecture requires only once streaming encoding to obtain a general condensed representation and adaptively selects significantly fewer tokens as input to LLM to answer specific questions. Therefore, we achieve a higher inference speed of 5.32 seconds per question and attain promising movie understanding without using subtitles.

Qualitative Results. We also present qualitative examples in Fig. 5. Typically, in Fig. 5a, our model accurately captures the detailed descriptions in the question, and precisely selects the relevant segments that contain the corresponding character. Moreover, in Fig. 5b, given a two-hour long movie and a high-level question on the movie plot, without relying on subtitles, VideoStreaming can comprehend the intent of the question and select relevant scenes from the lengthy video. In particular, the model selects the scenes of tightrope walk, team disputes, and equipment setup, clearly illustrating the protagonist's challenges, thereby contributing to a comprehensive answer generation.

4.3 Ablation Study

Historical Memory. We explore the influence of memory in the streaming encoding process, i.e., $\boldsymbol{H}_{k-1}$ in Eq 2. We report the fullset accuracy on EgoSchema [54] as well as global and breakpoint accuracy on MovieChat-1K [72] in Table 8. Typically, the historical memory significantly improves global understanding by 46.6%. This verifies our intuition that leveraging historical memory enables the model to produce a global representation that summarizes the entire video. Meanwhile, since



(a) An example on Next-GQA. Our model accurately selects the segments containing the target character.

(b) An example on a long movie. Our model selects typical segments that reveal the encountered challenges.

Figure 5: Examples of question answering and the selected timestamps based on specific instructions.

Table 8: Ablation studies on the effects of memory selection and historical memory in streaming encoding.

Memory	Selection	Fullset	Global Acc.	Break. Acc.
✓	✗	37.3	69.1	23.0
✗	✓	38.4	43.8	39.1
✓	✓	44.1	90.4	54.9

Table 9: Ablation studies on the number of layers used in the streaming encoder.

Layers	Params	Fullset	Acc@GQA
16	1.3B	44.1	17.8
24	2.0B	43.8	17.8
32	2.7B	41.3	15.6

we select a small portion of the encoded results from the long video as input to LLM, the lack of historical memory limits the temporal respective field and impairs the performance.

Memory Selection. We also validate the effects of our memory selection strategy. For comparison, we directly use the encoded results from the final four iterations as input to LLM and present the result in the first row of Table. 8. The historical memories in streaming encoding process enable the encoded results from the final iterations to provide coarse summarization of the entire video, thus attaining satisfactory results on global understanding. However, for questions regarding detailed analysis of specific moments, the lack of temporal selection leads to 31.9% performance drop in breakpoint mode accuracy. It demonstrates the effectiveness of our adaptive selection in gathering detailed information over the long time span, which facilitates more accurate and informative responses.

Streaming Encoder Architecture. Besides, we ablate the number of layers in Phi-2 used in memory-propagated streaming encoding. We show the results on EgoSchema [54] and Next-GQA [87] as well as the number of encoder parameters in Table 9. Interestingly, using fewer layers leads to better results. We conjecture this is because the language model is originally trained for next token prediction. Its feature space of the final Transformer decoder layer might not align with the objective of visual content condensation. Similar to [89, 49, 21], the shallower layers might produce feature embeddings that encode richer information and serve as more comprehensive condensed video representations.

Temporal Grounding Supervision. First, we present the studies on the use of temporal grounding supervision. As mentioned in Section 3.4, we employ around one-tenth of long video QA pairs to provide pseudo temporal labels. We compare four training strategies: (1) Fully weakly-supervised manner without any pseudo labels. (2) Using pseudo labels to train a *warm-up* model, then expanding to large-scale QA pairs. (3) *Mixing* all long video QA data, where the model uniformly receives temporal supervision in training. (4) Training on mixed data after warm-up initialization. The results on EgoSchema [54] and Next-GQA [87] in Table 10 indicate three key points: First, warm-up training contributes to more powerful grounding ability. The sparse temporal label supervision in mixed mode is overcome by the powerful initialization from warm-up training, which can generalize to large-scale data. Second, reusing the temporal labels after warm-up offers no additional benefits,

Table 10: Ablation studies on the use of temporal grounding supervision. Acc denotes the ratio of correctly answered questions on Next-GQA [87] regardless of grounding accuracy.

Warm-up	Mixed	Fullset	mIoP	mIoU	Acc@GQA	Acc
✗	✗	43.7	24.1	9.8	11.1	54.9
✓	✗	44.1	32.2	19.3	17.8	55.7
✗	✓	43.9	28.5	14.6	15.4	55.3
✓	✓	44.0	32.1	20.0	17.7	54.8

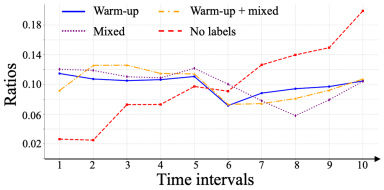


Figure 6: Distribution of selected timestamps on Next-GQA.

so we adopt warm-up as the default setting. Third, without using temporal labels, the grounding performance drops, but the QA accuracy remains stable. Fig. 6 reveals that compared to those trained with temporal labels, the weakly-supervised model selects relatively later segments that preserve previous contexts with the help of historical memory, thus maintaining comparable QA capacity.

More ablation studies on the number of summarization tokens and selected timestamps, the time prompts, and the similarity measurement are included in Appendix C.

5 Conclusion

In this paper, we introduce a novel approach to tackle the complexities of long video understanding with large language models (LLMs). Our proposed memory-propagated streaming encoding architecture segments long videos into short clips and iteratively encodes each clip in sequence. By leveraging historical memory from preceding clips, we incorporate temporal dynamics into the encoding process and produce a fixed-length memory to encapsulate arbitrarily long videos. To further augment the detailed information for handling specific questions, we develop adaptive memory selection that selects relevant timestamps based on given instructions. This approach ensures that the most pertinent historical memories are utilized for question answering, thereby facilitating detailed and informative responses. Our model achieves superior performance with substantially fewer tokens and higher efficiency on extensive long video benchmarks. We demonstrate that memories from the streaming encoding significantly enhance global video understanding, while adaptive selection results in accurate temporal grounding with respect to specific questions.

References

[1] Sharegpt. https://sharegpt.com/, 2023.

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

[3] 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.

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

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

[6] 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.

[7] Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*, 2024.

[8] 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.

[9] 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.

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[11] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.

[12] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[13] 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.

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

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Shuangrui Ding, Rui Qian, and Hongkai Xiong. Dual contrastive learning for spatio-temporal representation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5649–5658, 2022.

[18] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. *arXiv preprint arXiv:2311.17893*, 2023.

[19] Shuangrui Ding, Weidi Xie, Yabo Chen, Rui Qian, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Motion-inductive self-supervised object discovery in videos. *arXiv preprint arXiv:2210.00221*, 2022.

[20] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[22] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[25] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[26] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *arXiv preprint arXiv:2404.05726*, 2024.

[27] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

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

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

[30] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[31] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. *arXiv preprint arXiv:2402.13250*, 2024.

[32] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

[33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

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

[35] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[36] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

[37] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*, 2024.

[38] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[39] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[40] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*, 2023.

[41] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[42] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[44] 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.

[45] 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.

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

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

[48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

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

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[51] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[52] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

[53] 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.

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

[55] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[56] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[57] 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.

[58] OpenAI. Introducing chatgpt, 2022.

[59] OpenAI. Gpt4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[60] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[61] Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *arXiv preprint arXiv:2312.07395*, 2023.

[62] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Static and dynamic concepts for self-supervised video representation learning. In *European Conference on Computer Vision*, pages 145–164. Springer, 2022.

[63] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16675–16687, 2023.

[64] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7990–8001, 2021.

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

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

[67] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[68] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023.

[69] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

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

[71] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[72] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.

[73] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024.

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

[75] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`, 2023.

[76] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[77] 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.

[78] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13824–13835, 2023.

[79] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*, 2023.

[80] 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.

[81] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

[83] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.

[84] 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.

[85] 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.

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

[87] Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. Can i trust your answer? visually grounded video question answering. *arXiv preprint arXiv:2309.01327*, 2023.

[88] 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.

[89] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10075–10085, 2021.

[90] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

[91] 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.

[92] 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.

[93] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36, 2024.

[94] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

[95] 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.

[96] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[97] 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.

Limitations

One potential limitation is that we simply uniformly sample frames to form a set of short clips for memory-propagated streaming encoding. However, in a long video, different segments possess different amounts of information. The uniform sampling may result in using redundant tokens for clips with bland content. Meanwhile, the number of tokens used to represent clips with abundant visual contents and intensive temporal dynamics may be insufficient, leading to information loss. To address this limitation, we plan to explore adaptive segmentation techniques that dynamically adjust the segmented clip lengths based on the complexity and content of the video.

Impact Statements

Our proposed VideoStreaming, a streaming long video understanding architecture with large language models has various potential impacts for society. On the positive aspect, VideoStreaming contributes to improved intelligent video understanding, especially for long videos. This could be beneficial in education, entertainment, and information retrieval, where users often need to navigate and understand complex video materials. Besides, our technique could lead to advancements in multimedia analytics with applications in areas like video surveillance, market research, and content personalization.

On the negative aspect, the ability to efficiently process and retrieve information from long videos raises potential privacy and security concerns. If misused, this technology could be employed for unauthorized surveillance, personal monitoring, or other unethical purposes that infringe on individual privacy. In addition, the enhanced video understanding capabilities might be exploited for the creation of manipulated or misleading video content, leading to the spread of misinformation and the potential for social manipulation.

In conclusion, despite that VideoStreaming presents advancement in long video comprehensive, its development should be accompanied by careful consideration of ethical and societal implications.

A More Implementation Details

We use CLIP ViT-L/14 [65] to extract frame-wise features with input resolution 224×224, resulting in 256 tokens per frame. Then, we concatenate every four spatially adjacent visual tokens along channel dimension, representing each frame with 64 tokens with channel dimension 4096. The streaming encoder consists of a two-layer MLP projector (channel dimension 4096-2560-2560) with GELU activation [27] and a language model Phi-2 2.7B [25]. In the first training stage, we initially freeze Phi-2, and only tune the MLP projector on 790K caption pairs, including 558K image caption data from CC3M [70] and 232K short video caption data from WebVid 2.5M [6]. Following LLaVA [49, 48], we use AdamW optimizer [50] with global batchsize 256, initial learning rate 1×10^{-3} with cosine decay to train 1 epoch for modality alignment. Subsequently, we jointly train Phi-2 and the MLP projector on 763K QA pairs, including 625K image QA pairs [23, 30, 38, 41, 49, 55, 56, 69, 71], 40K text conversations [1] and 98K video QA pairs [10], with global batchsize 128, initial learning rate 2×10^{-5} with cosine decay.

In the second stage, we jointly train ViT, the streaming encoder and the LLM on long video data. In the memory-propagated streaming encoding process, we insert a brief prompt to indicate the explicit timestamps of the historical memory and the input clip formulated as *This contains a history of {start} to {end} seconds, and a clip sampled in {start} to {end} seconds.*. We adopt the output of the first 16 layers out of the 32 layers of Phi-2 as the condensed representation. Then, we adaptively select 4 most relevant timestamps and feed the associated 256 memory tokens into a two-layer MLP projector with channel dimension 2560-4096-4096 and an LLM, Vicuna-7B [13] to generate the final responses. We jointly train the whole architecture, including Vicuna, Phi-2, MLP projectors and ViT encoder, on long video QA data with global batchsize 128, initial learning rate 2×10^{-5} with cosine decay. In default, we first use 20K synthesized long videos and sample 10K QA pairs curated from Panda-70M with pseudeo temporal grounding labels to train memory selection as warm-up. The learning objectives contain a standard next token prediction loss and a supervised KL divergence loss that aligns the distribution of the predicted memory selection results and the pseudo temporal labels. Next, based on the warm-up model, we further train on the rest 295 long video QA pairs only with

Data provided by Panda-70M:
[1, 12]: The speaker is a female news reporter who is sitting in front of a studio camera.
[15, 23]: A group of people standing in line outside of a building.
[25, 28]: Two children standing next to each other on a tv screen.
⋮ ⋮
[153, 161]: A woman is interviewed by a reporter in a park.

Created QA pairs:
Q: Describe the video in chronological order with detailed timestamps.
A: In 1 to 12 seconds, the speaker is a female news reporter who is sitting in front of a studio camera. In 15 to 23 seconds, a group of people standing in line outside of a building. In 25 to 28 seconds, two children standing next to each other on a tv screen. ··· In 162 to 166 seconds, a woman is interviewed by a reporter in a park.
Q: Summarize the events in 15 to 23 seconds.
A: A group of people standing in line outside of a building.
Q: What is the timestamp that a man in a suit and tie is talking into a microphone at a press conference?
A: In 83 to 104 seconds.
Q: What happens in 54 to 64 seconds?
A: A picture of a person and their family from the news is featured in a video.

Figure 7: An example of the QA pairs from the captions and segmented timestamps from Panda-70M [11].

Table 11: Ablatoin study on the number of summarization tokens and selected timestamps. We report the results on EgoSchema [54] and Next-GQA [87].

P	V	Tokens	Fullset	Acc@GQA
1	4	64	32.1	9.8
1	8	128	33.4	10.5
4	1	64	41.6	15.5
4	4	256	44.1	17.8
4	8	512	44.9	18.0
16	1	256	42.5	16.3
16	4	1024	43.8	17.9

next token prediction loss. The whole training is conducted on 32 A100 (80G) GPUs for around 2.5 days.

B Long Video QA Data Creation

In addition to the existing 25K long video QA pairs on movies [72, 46], we create more QA data from two aspects. First, we leverage the existing short video QA dataset [83, 88] and synthesize short videos into minute-long videos with average duration of one minute. The original questions of each short video coarsely correspond to a temporal segment in the synthesized long video. We use this correspondence as noisy labels to supervise the memory selection. Second, recent Panda-70M [11] segments long videos into short clips and produces captions for each clips. This dataset provides the original long videos, the captions of segmented clips as well as the segmentation timestamps. Based on these cues, we produce multi-round QA conversations. Below we show an example in Fig. 7. The produced time-sensitive QA pairs are crucial to enhance the temporal awareness and guide precise memory selection in long videos.

C More Ablation Studies

We provide more ablation studies on the number of summarization tokens and selected timestamps, the effects of time prompts in the memory-propagated streaming encoding process, and the similarity measurement used in memory selection.

Table 12: Ablation study on the formulation of time prompts. We report the results on EgoSchema [54], Next-GQA [87] and MovieChat-1K [72].

Prompt	Fullset	Acc@GQA	Global Accuracy	Breakpoint Accuracy
None	38.8	12.4	66.7	19.8
Clip	40.5	15.4	70.3	44.5
Memory	42.1	16.1	83.3	43.1
Clip+Memory	44.1	17.8	90.4	54.9

Table 13: Ablation study on the similarity measurement. We report the results on EgoSchema [54], Next-GQA [87] and MovieChat-1K [72]

Similarity	Fullset	Acc@GQA	Global Accuracy	Breakpoint Accuracy
Cosine	44.1	17.8	90.4	54.9
Dot product	34.5	9.3	55.6	22.1

The Number of Summarization Tokens and Selected Timestamps. We compare using different number of summarization tokens and selected timestamps, i.e., P in Eq. 1 and V in Eq. 3. We compare the performance as well as the number of tokens input to LLM in Table 11. We conclude three observations. First, too few summarization tokens, e.g., $P = 1$, leads to substantial performance drop, since it condenses a 16-frame into only 16 tokens with significant information loss in spatial contexts. Such information loss cannot be compensated by selecting more temporal segments. Second, the performance saturates when improving P from 4 to 16. This is because the existing video benchmarks [54, 87] do not place high demands on spatial detail understanding. It is sufficient to represent each frame with 4 tokens on average. Third, increasing the number of selected timestamps only results in minor improvements, which is not proportional to the increased number of tokens. This can be attributed to the historical memory used in the streaming encoding process. The utilization of historical memory enables the condensed representation of each clip to encompass the information in preceding clips, which enlarges the temporal receptive field. Hence, increasing the number of selected timestamps does not proportionally increase the temporal receptive field, resulting in slight performance improvements.

Time Prompts. We explore three different formulations of the time prompts used in memory-propagated streaming encoding: (1) Only with the timestamps of the current clip, e.g., *This clip is sampled in {start} to {end} seconds.* (2) Only with the timestamps of the historical memory, e.g., *This contains a history of {start} to {end} seconds.* (3) Simultaneously with the timestamps of the historical memory and the current clip, e.g., *This contains a history of {start} to {end} seconds, and a clip sampled in {start} to {end} seconds.* We report the results of different time prompts in Table 12. It is obvious that the lack of time prompts leads to substantial performance drop in the MovieNet-1K breakpoint mode accuracy, which requires detailed analysis of specific moments. The reason is that the breakpoint mode requires the model to answer questions at specified timestamps, the time prompts provide the model with necessary information in adaptive selection. Meanwhile, incorporating the timestamps of historical memories results in more significant improvements in global understanding. Overall, jointly leveraging the memory and clip timestamps contributes to the best results.

Similarity Measurement. Finally, we present the study on the similarity measurement used in adaptive memory selection. We compare the default cosine similarity against simple dot product without normalization in Table 13. Empirically, we observe that dot product could result in numerical instability, leading to overflow in training. Consequently, the calculated similarity score cannot reflect the correlation between the instruction and different segments and results in poor results on questions that require accurate temporal grounding, e.g., the Acc@GQA metric on Next-GQA [87] and the breakpoint mode accuracy on MovieChat-1K [72].