VideoLLaMB: Long Streaming Video Understanding with Recurrent Memory Bridges

Yuxuan Wang [1*] **Yiqi Song** [1,2*] **Cihang Xie** [3] **Yang Liu** [4] **Zilong Zheng** [1†]

[1] NLCo Lab, State Key Laboratory of General Artificial Intelligence, BIGAI
[2] School of Computer Science & Technology, Beijing Institute of Technology
[3] Computer Science and Engineering, University of California
[4] Wangxuan Institute of Computer Technology, Peking University

flagwyx@gmail.com, yiqis@bit.edu.cn, zlzheng@bigai.ai

https://github.com/bigai-nlco/VideoLLaMB

Abstract

Recent advancements in large-scale video-language models have shown significant potential for real-time planning and detailed interactions. However, their high computational demands and the scarcity of annotated datasets limit their practicality for academic researchers. In this work, we introduce ***VideoLLaMB****, a novel and efficient framework for long video understanding that leverages recurrent memory bridges and temporal memory tokens to enable seamless encoding of entire video sequences with preserved semantic continuity. Central to our approach is a SceneTiling algorithm that segments videos into coherent semantic units, facilitating robust understanding across tasks without requiring additional training. VideoLLaMB achieves state-of-the-art performance, surpassing existing models by* 4.2 *points on four VideoQA benchmarks and by* 2.06 *points on egocentric planning tasks. Notably, it maintains strong performance under extreme video length scaling (up to* 8×*) and excels at fine-grained frame retrieval on our proposed* ***Needle in a Video Haystack (NIAVH)*** *benchmark. With linear GPU memory scaling, VideoLLaMB processes up to 320 frames using a single Nvidia A100 GPU, despite being trained on only 16 frames—offering an unprecedented balance of accuracy, scalability, and cost-effectiveness. This makes it highly accessible and practical for the academic community.*

1. Introduction

Recent advancements in large-scale video language models, exemplified by GPT-4o and Project Astra have captivated global attention due to their potential for sophisticated interaction with real-world environments [23, 81]. These models are particularly noteworthy for their capacity to comprehend streaming video [65], which can be conceptualized as video with an unlimited context length. This capability necessitates both the observation of the current state and the ability to leverage long-term memory. Despite their promise, the training of such large-scale video-language (VidL) foundational models remains impractical for academic researchers. This impracticality arises from the substantial computational resources required by the complex, high-dimensional nature of long streaming video data, in addition to the scarcity of well-annotated, publicly available video-language datasets. These factors present significant challenges to scaling video-language models to the extent observed with large language models (LLMs).

To circumvent these challenges, the community has witnessed a growing interest in developing computationally efficient multimodal large language models (MLLMs). Traditional methods resort to *video compression* strategies, such as sampling [34, 76], aggregation [70], semantic consolidation [54], and resampling [20, 39], in order to temporally reduce the length of the video. Yet, these methods often lead to the **loss of critical visual cues**, undermining the model's ability to capture essential cues. Other approaches [47, 61] segment videos into shorter clips to mitigate the computational load of processing long videos. However, segmentation can **disrupt the semantic flow of content**, complicating the encoding process and potentially impacting the general understanding of the video narrative. Lastly, prevalent video understanding benchmarks, primarily based on linguistic question-answering pairs, exhibit **static** [26] and/or **language biases** [50, 82]. These biases favor models that rely more on static imagery or textual elements, re-

*Equal contributions
†Correspondence to Zilong Zheng ⟨zlzheng@bigai.ai⟩.

spectively, and fail to provide a comprehensive assessment of a model's capability on extended video sequences.

To address these multifaceted limitations, we introduce **VideoLLaMB**, an innovative framework that learns temporal **M**emory tokens within **B**ridge layers that recursively encode the entire video content, ensuring that no visual information is discarded deliberately (Figure 1; §2). Specifically, we devise Memory Bridge Layers, equipped with recurrent memory tokens, that function without altering the architecture of the visual encoder and LLM. Furthermore, to mitigate the backpropagation through time (BPTT) issue, we maintain long-term dependencies by preserving recurrent memory tokens in a memory cache, which is periodically refreshed through a retrieval process. To compensate for the limitations of the sliding window technique, we propose the SceneTiling algorithm that divides the video into relatively independent sequences of semantic segments. This reduces the dimensions within each semantic unit without sacrificing semantic details. By constructing our recurrent memory with a retrieval mechanism based on these semantic segments, our method strikes a balance between effective and efficient comprehension of the current state and long-term memory retention.

In §3, we highlight the empirical advantages of VideoLLaMB in comparison with prior arts as:

- **Long-Term Memory Reservation.** We demonstrate the effectiveness of VideoLLaMB in comprehensive long video understanding and enhanced frame retrieval through rigorous testing on established benchmarks. Using VideoMME [16], EgoSchema [41] and NexTQA [67], VideoLLaMB shows an average improvement of 4.2 accuracy points over PLLaVA [70], despite utilizing the same initialization and training video dataset. Notably, VideoLLaMB maintains robust performance even when video lengths extend to eight times their original duration. To further evaluate frame retrieval capabilities, we introduce the multimodal Needle in a Video Haystack (NIAVH) test, where VideoLLaMB consistently identifies the correct image within videos ranging from 1 to 320 seconds, surpassing other methods as video length increases.
- **Real-time egocentric planning.** To evaluate our model's performance in video planning tasks, we used the dataset EgoPlan [9]. Our method achieves the best performance among all 7B video-language models, showing an improvement of 2.06 accuracy points over PLLaVA.
- **Training-free streaming captioning.** By employing the SceneTiling algorithm, our method can automatically predict the end of a caption in streaming video without relying on special tokens during the training phase.

2. VideoLLaMB

VideoLLaMB is an extensible framework designed to enhance long video understanding, composed of three key modules: semantic-based segmenter (§2.1), recurrent memory layer (§2.2), and memory retriever (§2.3). Each of these components will be detailed in the subsequent sections. Figure 1 depicts the overall framework.

2.1. SceneTiling: Segmentation with Semantics

Semantic segmentation along temporal sequence has long been recognized as an important task because it preserves the non-linear structure of context and greatly aids in compressing extensive context [8, 22, 43, 48, 60]. To address the disruption of semantic flow (see §1), we introduce SceneTiling, a model-free scene segmentation algorithm inspired by TextTiling [21]. SceneTiling divides the entire video sequence into segments that are semantically distinct, ensuring intra-segment coherence.

Formally, given a sequence of n frames $\{v_1, v_2, \ldots, v_n\}$, the SceneTiling algorithm is as follows.

1. Compute the cosine similarity $S_C(\cdot,\cdot)$ between adjacent frame pairs using the `[CLS]` token from ViT, resulting in a sequence of similarity scores $\{c_1, c_2, \ldots, c_{n-1}\}$, where $c_i = S_C(\mathrm{ViT}(v_i), \mathrm{ViT}(v_{i+1}))$.
2. Calculate the depth score for each point as $d_i = (cl_i + cr_i - 2c_i)/2$, where cl_i and cr_i are the highest score to the left and right of c_i, respectively. A higher depth score indicates that the surrounding similarity is greater than at the point itself.
3. Calculate the expectation μ and variance σ of the depth scores $\{d_1, d_2, \ldots, d_{n-1}\}$. Set the segmentation threshold as $\mu + \alpha\cdot\sigma$, where α is a hyperparameter controlling the likelihood of segmenting the video. Select the $K-1$ depth scores that exceed the threshold to divide the video into K semantic segments $\{s_1, s_2, \ldots, s_K\}$. Each segment represents a relatively independent semantic unit consisting of a sequence of frames.

Aside from temporal semantic segmentation, SceneTiling enables streaming video captioning without requiring training with special tokens [7, 15, 83].

2.2. Recurrent Memory Bridge Layers

Traditional recurrent memory-based Transformers [4, 5, 25] incur significant computational costs when scaled up, *i.e.*, $\mathcal{O}(L^2/K)$, where L is the context length and K is the number of segments, primarily due to its recurrent mechanism over the whole language model. More recently, some works empirically identify that linear projection best withstands visual information within MLLMs [35, 36, 79], albeit with high space complexity, whereas the resampler has strong compressing ability on semantic information [28], though it tends to miss detailed information [69].

In this work, we devised a novel Recurrent Memory Bridge Layer, implemented as Transformer blocks, that integrates recurrent memory tokens within bridge layers to enhance the linear layer's memorization ability. For-

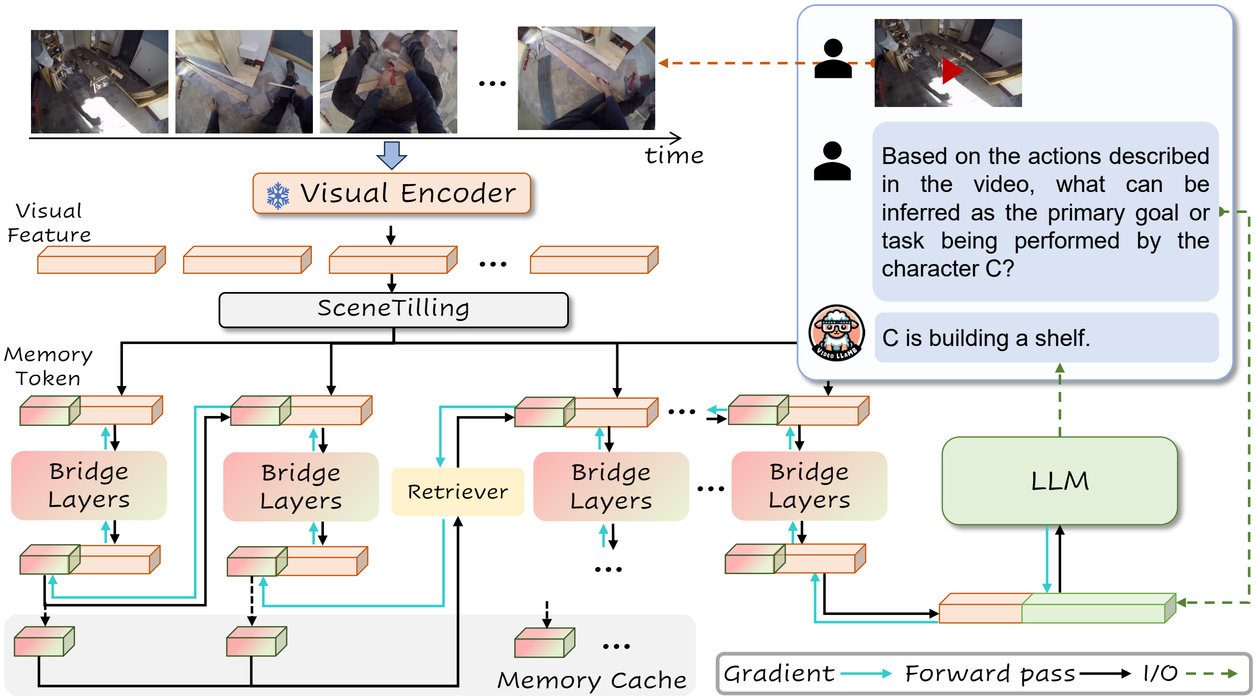

Figure 1. **Overview of VideoLLaMB.** We first extract the video features using an off-the-shelf vision encoder, then apply SceneTiling to segment the video into semantic segments (§2.1). Next, we use recurrent memory on these semantic segments to store video information within memory tokens (§2.2). We further employ a retrieval mechanism to update the memory tokens and address long-dependency issues(§2.3). Finally, we project the memory-token-augmented features from the current video segment into the LLM.

mally, for each video segment s_i, we prepend a fixed number of memory tokens, denoted as $[m_i; s_i]$, where m_i represents the memory tokens. Subsequently, we apply standard self-attention to this sequence, yielding $[m_{i+1}; o_i] = \text{BridgeLayer}([m_i; s_i])$. Here, m_{i+1} is the updated memory token, and o_i is the visual representation from the bridge layers. This process is carried out recursively, traversing the semantic video segments while updating the memory tokens. After a total of k steps, this output represents the condensed visual summary of the video sequence and will serve as the input for the LLM. As such, the Memory Bridge can compress past video into memory while preserving current video scenes through projection without losing detailed information.

2.3. Memory Cache with Retrieval

One of the primary challenges associated with recurrent memory bridge layers is the potential for gradient vanishing, which can impede the model's ability to learn long-range dependencies [29]. To mitigate this issue, we propose the incorporation of a memory cache with a retrieval strategy designed to preserve previous states of memory.

Memory Attention At each timestep i, the system stores all previous memory tokens in a memory cache, denoted as $M_i = [m_1, \ldots, m_i]$. We employ a self-retrieval mechanism to update the current memory token m_i. Specifically, we treat m_i as a query and the concatenated memory cache M_i as key and value. The model performs a standard multi-head cross-attention operation to integrate information from previous timesteps into the current memory state, yielding the updated memory token $m_{i+1} = \text{Softmax}\left(W_i^Q m_i (W_i^K M_i)^\top / \sqrt{d_k}\right) W_i^V M_i$, where W_i^Q, W_i^K, W_i^V are weight matrices for query, key and value, respectively.

Computational Complexity For bridge layers, we consider three main components for the theoretical complexity: (i) the self-attention within each segment, which scales as $\mathcal{O}((C+M)^2)$, where C is the segment length and M is the length of memory tokens; (ii) the memory retrieval, which scales as $\mathcal{O}(MK)$; and (iii) the recurrent processing. Consequently, the overall time complexity of our approach is $\mathcal{O}(K^2)$, and the space complexity is $\mathcal{O}(K)$. For the LLM, the complexity is $\mathcal{O}(M^2)$. In practice, the segment length C is a constant that depends on the constraint of LLM. K is one M-th of L, thus our segmentation approach effectively compresses semantic units to an extreme degree, thereby striking a favorable balance between computational efficiency and model efficacy. Moreover, The number of segments can be fixed to accommodate the constraints of the environment.

3. Experiments

3.1. Setup

We utilize Vicuna-7B-v1.5 as the LLM and ViT-L/14 as the visual backbone following Video-LLaVA [34]. Each frame is resized and cropped to a dimension of 224×224. The Memory Bridge Layers are based on a single-layer Transformer. Our model is trained and evaluated with 16 frames and 4 segments, following the same video data protocol as PLLaVA [70].

3.2. Long-form Video Understanding

Baselines We conduct a comparative analysis of two types of models: retrieval-based methods and generative video-language models, as elaborated in Section 4. To ensure fairness in our comparisons, we primarily focus on state-of-the-art models such as LLaVA-NeXT-Video-DPO [79] and PLLaVA [70], which utilize the same base model and video datasets as our approach. Other SoTA models, including MovieChat [54], MA-LMM [20], VideoStreaming [47], and Video-xl [52], are not consistently included in all benchmarks due to variations in training data, model configurations, and benchmark settings. Nevertheless, we evaluate the key compression settings of these baselines using the same data and model configurations as ours, as detailed in Sections 3.6 and 3.8.

Results on EgoSchema EgoSchema [41] consists of egocentric videos, each averaging **180 seconds** in length. This video QA dataset focuses on aspects such as understanding, reasoning, and long-term memory. In our experiment, we follow the precedent set by previous studies and use the public subset for evaluation. The results are presented in Table 1. Overall, our method significantly outperforms current generative video language models trained on similar data, demonstrating robust performance compared to other approaches and confirming its efficacy. Specifically, we compare our method with PLLaVA [70], which shares the same training data, LLM backbones, and input number of frames. Our method shows significant improvements over PLLaVA, indicating its superiority in understanding long egocentric videos. While our method does not yet match the performance of fine-tuned retrieval-based methods, we plan to apply our approach to larger language models to bridge this performance gap.

Length Extrapolation The model is trained on 16-frame sequences, divided into 4 segments. However, in real-world scenarios, videos can be significantly longer than this training configuration. To demonstrate VideoLLaMB's ability to extrapolate to longer videos, we conducted experiments on EgoSchema under two conditions: 1) dynamic segments, which adaptively control the number of segments based on

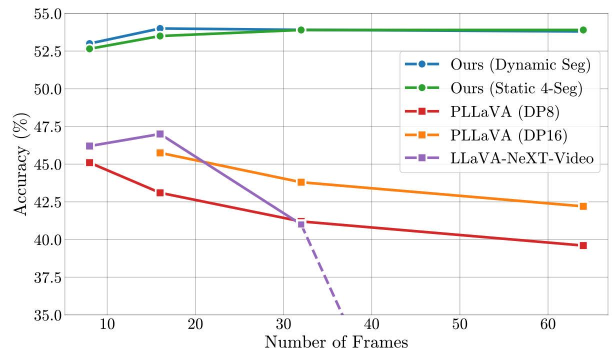

Figure 2. **Length extrapolation results** on EgoSchema dataset.

the SceneTiling threshold, and 2) static segments, fixed at 4 segments. Results in Figure 2 reveal that dynamic segments are more effective than static segments, especially for shorter videos, indicating that our method can effectively maintain an appropriate number of segments. However, as video length increases, the performance of dynamic segments declines, notably at the 32-frame mark, where both strategies use four segments. Beyond this point, increasing the number of segments results in diminishing returns, likely due to the domain gap from training on shorter videos. To address this issue, we plan to fine-tune our models on longer videos for more substantial improvements. Overall, compared to PLLaVA, our method maintains consistent performance as the input length increases. In summary, our approach effectively extracts key information from videos, outperforming the simple pooling strategies used for memory consolidation in existing methods.

Results on NExTQA NExTQA [67], featuring daily-life videos that average 45 seconds in length, is designed to test a variety of question types, specifically temporal, causal, and descriptive questions. We applied our method to NExTQA to evaluate its temporal grounding ability. To maintain consistency with established benchmarks, we used the validation set for evaluation. In Table 2, we present the comprehensive results of our analysis. For a fair comparison, our primary benchmark is against PLLaVA, which includes instruction data from the NExTQA training set. Our method surpasses PLLaVA by 2.9 points. Notably, our approach demonstrates a significant enhancement in the temporal setting, achieving a 4.6 point improvement over PLLaVA. These results indicate that our scene-segment aware method effectively improves the model's temporal grounding ability by compressing abundant information within scenes that share high semantic similarity.

Results on Long Context VideoQA We evaluate our methods on the long video benchmark, VideoMME [16], which contain videos ranging from 11 seconds to **1 hour** in length. The results are illustrated in Table 4. We compare methods using the same training data and LLM backbones. Compared to PLLaVA-B [70] and VideoChat-2 [30], our

Model	LLM	Frames	Accuracy
GPT4-o	OpenAI API	16	72.2
Retrieval-based Video-Language Models			
LongViViT* 2024	-	256	56.8
MC-ViT-L* 2024	-	128	62.5
Generative Video-Language Models			
SeViLA 2023	Flan-T5-XL	32	25.8
mPLUG-Owl 2023	LLaMA-7B	5	33.8
Video-LLaVA 2024	Vicuna-7B	8	40.2
LLaVA-NeXT-Video-DPO 2024	Vicuna-7B	32	41.6
PLLaVA 2024	Vicuna-7B	16 (16)	45.6
PLLaVA 2024	Vicuna-7B	32 (16)	43.8
VideoLLaMB (Ours)	Vicuna-7B	32 (8)	**53.8**

Table 1. **Results on Subset of EgoSchema under zero-shot setting.** *: the model has been fine-tuned using the training data from EgoSchema. (n): n frames are used in training.

Model	Temporal	Causal	Description	All
GPT4-o	70.3	78.0	80.8	76.0
Retrieval-based Video-Language Models				
AIO* 2023	48.0	48.6	63.2	50.6
VQA-T* 2021	49.6	51.5	63.2	52.3
ATP* 2022	50.2	53.1	66.8	54.3
VGT* 2022	52.3	55.1	64.1	55.0
MIST-CLIP* 2023	56.6	54.6	66.9	57.1
Generative Video-Language Models				
SeViLA 2023	61.5	61.3	75.6	63.6
LLaMA-VID 2024	53.8	60.0	73.0	59.5
Video-LLaVA 2024	56.9	61.0	75.0	61.3
LLaVA-NeXT-Video-DPO 2024	55.6	61.0	73.9	61.3
PLLaVA* 2024	62.2	68.5	**79.7**	68.2
VideoLLaMB (Ours)*	**66.8**	**71.6**	78.4	**71.1**

Table 2. **Comparison accuracy on NExT-QA.** * indicates that the instruction data includes the training data from NExTQA.

Method	Vision Encoder	LLM Size	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	Avg.
GPT-4V	GPT-4V	/	55.5	63.5	72.0	46.5	73.5	18.5	59.0	29.5	12.0	40.5	83.5	39.0	12.0	22.5	45.0	47.5	52.0	31.0	59.0	11.0	43.5
mPLUG-Owl-I 2023	ViT-L	7B	25.0	20.0	44.5	27.0	23.5	36.0	24.0	34.0	23.0	24.0	34.5	34.5	22.0	31.5	40.0	24.0	37.0	25.5	21.0	37.0	29.4
LLaMA-Adapter 2024	ViT-B	7B	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	25.0	31.5	22.5	28.0	32.0	31.7
BLIP2 2022	ViT-G	2.7B	24.5	29.0	33.5	17.0	42.0	51.5	26.0	31.0	25.5	26.0	32.5	25.5	30.0	40.0	42.0	27.0	30.0	26.0	37.0	31.0	31.4
Otter-I 2025	ViT-L	7B	34.5	32.0	39.5	30.5	38.5	48.5	44.0	29.5	19.0	25.5	55.0	20.0	32.5	28.5	39.0	28.0	27.0	32.0	29.0	36.5	33.5
MiniGPT-4 2023	ViT-G	7B	16.0	18.0	26.0	21.5	16.0	29.5	25.5	13.0	11.5	12.0	9.5	32.5	15.5	8.0	34.0	26.0	29.5	19.0	9.9	3.0	18.8
InstructBLIP 2023	ViT-G	7B	20.0	16.5	46.0	24.5	46.0	51.0	26.0	37.5	22.0	23.0	46.5	**42.5**	26.5	40.5	32.0	25.5	30.0	25.5	30.5	38.0	32.5
LLaVA 2023	ViT-L	7B	28.0	39.5	63.0	30.5	39.0	53.0	41.0	41.5	23.0	20.5	45.0	34.0	20.5	38.5	47.0	25.0	36.0	27.0	26.5	42.0	36.0
Video-LLaMA 2023	CLIP-G	7B	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0	34.1
LLaMA-Adapter 2024	ViT-B	7B	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	25.0	31.5	22.5	28.0	32.0	31.7
Video-ChatGPT 2024	ViT-L	7B	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	29.0	33.0	29.5	26.0	35.5	32.7
VideoChat 2023	CLIP-G	7B	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0	35.5
VideoChat2$^{\beta}$ 2024	UMT-L	7B	66.0	47.5	83.5	49.5	60.0	58.0	71.5	42.5	23.0	23.0	88.5	39.0	42.0	58.5	44.0	49.0	36.5	35.0	40.5	65.5	51.1
PLLaVA 7B$^{\alpha}$ 2024	ViT-L	7B	58.0	49.0	55.5	41.0	61.0	56.0	61.0	36.0	23.5	26.0	82.0	39.5	42.0	52.0	45.0	42.0	53.5	30.5	48.0	31.0	46.6
PLLaVA 13B$^{\alpha}$ 2024	ViT-L	13B	66.0	53.0	65.5	45.0	65.0	58.0	64.5	35.5	23.5	30.0	85.0	39.5	45.5	57.0	47.5	49.5	49.0	33.0	53.0	37.0	50.1
VideoLLaMB $^{\alpha}$ (Ours)	ViT-L	7B	52.0	50.5	85.5	42.5	51.0	69.5	56.0	38.5	41.0	24.0	69.5	40.0	48.0	71.5	43.5	34.5	41.5	29.5	38.0	60.0	49.33
VideoLLaMB $^{\beta}$ (Ours)	ViT-L	7B	54.5	47.0	86.5	44.5	52.0	79.0	58.5	32.0	47.0	33.0	82.5	40.5	52.0	82.0	40.5	37.5	43.0	31.0	42.5	60.0	52.5

Table 3. **Results on MVBench [32] multi-choice question answering.** We highlight top-3 results among all 7B models of each category in purple. α: training data from Xu et al. [70]. β: training with data from Li et al. [32].

Model	Data	Short	Medium	Long	All.	Comp.
LLaVA-NeXT-Vicuna 2024	LLaVA-NeXT	35.26	37.44	32.88	35.44	1.00
VideoChat2 2023	VideoChat2	-	-	-	33.7	0.49
PLLaVA 2024	PLLaVA	46.44	38.00	33.22	38.22	0.25
VideoLLaMB α (Ours)	PLLaVA	46.11	38.44	34.22	39.59	0.06
VideoLLaMB β (Ours)	VideoChat2	49.22	39.11	35.89	41.41	0.06

Table 4. **Results on VideoMME**. We list models that use the same training data and Vicuna-7B backbones for fair comparison, for existing SoTA models like Video-XL [52], VideoStreaming [47] are trained on self-constructed dataset. Comp: Compression.

method shows improvements on both benchmarks. Notably, our method demonstrates consistent improvements on long videos in Video-MME with much less compression rate.

3.3. Comprehensive Video Understanding

We evaluated our method using the comprehensive video understanding benchmark MVBench [32]. As shown in Table 3, our approach maintains strong performance across general video understanding tasks. Remarkably, with the same training data as PLLaVA, our method achieves performance comparable to a 13B-level model. Our method effectively extracts information from both short and long videos. To assess scalability, we trained our model on the VideoChat2 [30] dataset. The results, presented at the bottom of Table 3, demonstrate that our model, when trained on larger video datasets, improves accuracy on MVBench by 3.17 points and outperforms VideoChat2, which was trained on the same dataset.

3.4. Planning Tasks

Baselines The original protocol dictated the selection of a single frame corresponding to each action. To refine this approach and enhance the evaluation process, we introduce a smoother method. This involves segmenting the entire video into intervals based on predefined timesteps. This revised method is applied in the evaluation of PLLaVA, LLaVA-NeXT-Video-DPO, and VideoLLaMB.

Results on EgoPlan [9] The EgoPlan dataset [9] was developed as an egocentric question-answering benchmark tailored for embodied planning tasks, comprising 3,355 questions. The evaluation follows the framework estab-

Model	LLM	Accuracy
GPT-4V	OpenAI API	37.98
VideoLLaMA [76]	LLaMA2-Chat-7B	29.85
LLaVA-NeXT-Video [79]	Vicuna-7B	28.96
PLLaVA [70]	Vicuna-7B	30.26
VideoLLaMB (Ours)	Vicuna-7B	**32.32**

Table 5. **Results on EgoPlan under Zero-shot setting.**

lished in the original study, utilizing the probability $p(a|v,l)$ to identify the most suitable answer candidates. In Table 5, we demonstrate that our model surpasses all other video-language models in performance. This suggests that our model's use of memory significantly enhances its planning capabilities compared to methods focused on the current stage. We are confident that our method holds great promise for generalizing to practical, real-world scenarios.

3.5. Streaming Caption

Streaming dense video captions [7, 83] involves generating captions for videos in real-time, without the need to process the entire video sequence beforehand. The primary challenge in this task is determining the exact timestamps to predict event captions. Most existing methods rely on special tokens, annotated as the end of an action, for training. Our approach introduces the SceneTiling algorithm, which can automatically identify the break points in a streaming video and generate captions without requiring special training tokens. To enhance the efficiency of our method, we calculate the depth score using only the left similarity: $d_i = (cl_i - c_i)/2$. This demonstrates that our method can effectively detect scene changes and automatically generate event captions.

3.6. Stress Test: "Needle In a Video Haystack"

To address existing limitations in long-form video language understanding benchmarks, our work takes inspiration from the latest developments in the field and develops a new benchmark specifically designed for the task of identifying specific content within extensive video material, a challenge we refer to as the Needle In A Video Haystack (NIAVH). This benchmark is unique in that it supports queries in various modalities, including text, image, and video, allowing for a more comprehensive assessment of a model's video understanding capability.

Benchmark Setting In NIAVH, we utilize ego-centric videos from the Ego4D [19] dataset as the "haystack," within which we seek to locate the "needle" provided in three distinct modalities: textual, image, and video. For the textual modality, we supply a crafted description; for the image modality, we use DALL-E to create a corresponding image; and for the video modality, we use Sora [2] generated short video clip, all based on the same description. Each "needle" is set to a duration of 1 second and is inserted into the concatenated Ego4D videos at various depths and lengths. To evaluate the benchmark, a direct question about the details within the needles is set, and an LLM compares the response with the ground truth, providing a score from 1 to 10, with 10 indicating a perfect match. For quantitative results, we calculate the average scores for additional analysis.

Comparison with similar benchmarks Recent work proposes a multimodal needle-in-a-haystack benchmark MM-NIAH [57], which focuses on a mixture of images and documents as the haystack and only supports text and image needles. In contrast, NIVAH focuses on streaming video stacks and supports text, image, and video needles.

Experiment Setup Given the limitations of current methods in understanding long videos, we designed an experiment where the "haystack" is a 320-second video. The "needle" is a 1-second video clip generated by Sora, prompted by the description, "the young man seated on a cloud in the sky is reading a book". The associated question posed for the experiment is, "What is the young man seated on a cloud in the sky doing?". We divided the context into 40 intervals and set the video depth at 12 intervals.

Results and Analysis In our experiment, we evaluate our approach with four distinct methods. These include (a) adaptive pooling [70], (b) position extrapolation combined with sampling [79], (c) the integration of resampler with memory retrieval and consolidation [20], and (d) video alignment with long-context LLM without compression [77]. For each model, we standardize the video frame rate to one frame per second, aligning the number of input frames with the duration of the video in seconds. This ensures that the inputs contain the needle information and all the models are in fair comparison. The outcomes of this evaluation are depicted in Figure 3. Our analysis leads to the following key observations:

- Methods utilizing an adaptive pooling strategy risk omitting crucial information, as the length of the source material (the "haystack") is often many times greater than the target segment (the "needle").
- Pooling strategies that incorporate position extrapolation are ineffective at predicting lengths that exceed those encountered during training or fine-tuning.
- Combining a resampler with a retrieval strategy markedly improves the encoding of extended information in a video. However, the encoded length is ultimately constrained by the resampler's compression capacity.
- VideoLLaMB with retrieval is the most efficient at preserving previously encountered information. Nevertheless, it still exhibits shortcomings: it tends to forget earlier information and is prone to hallucination issues, such as misidentifying "holding book" as "holding phone".

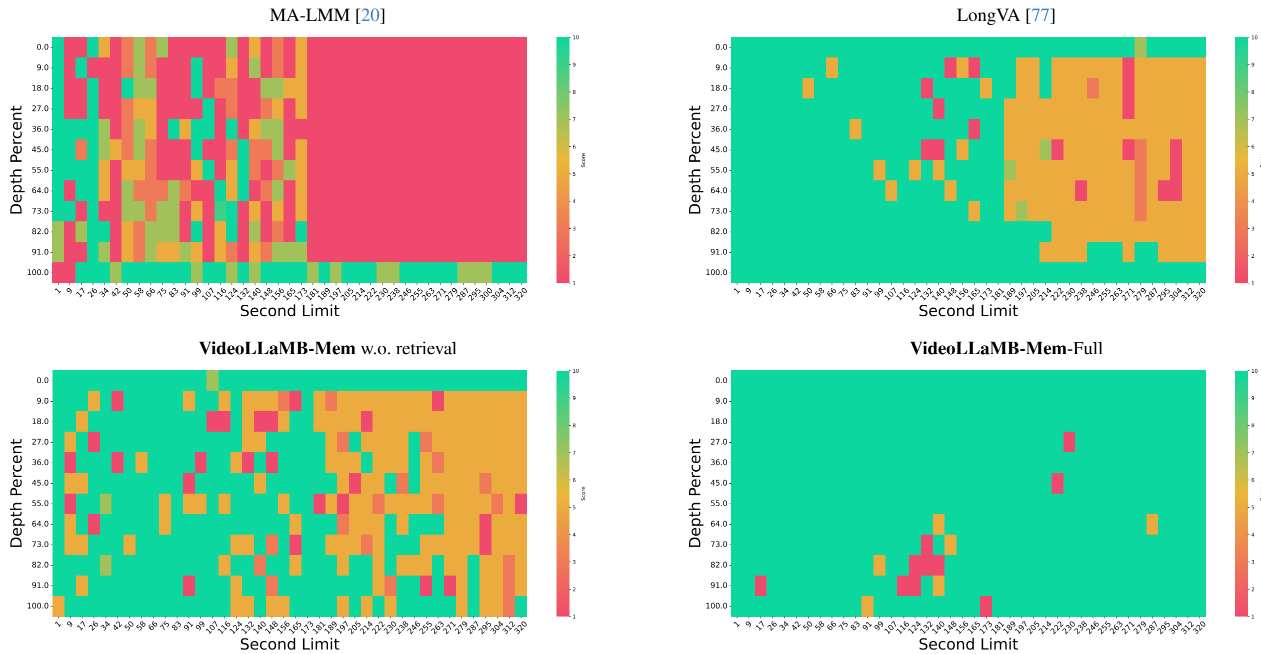

Figure 3. **Comparison of VideoLLaMB with two long video understanding models on Needle In A Video Haystack (NIAVH).** Currently, we set the context length to 320 seconds w.r.t. existing models' ability and set the frame rate to 1 fps to ensure the input contains the needle. The X-axis indicates the video length, and the Y-axis is the depth of the insertion point.

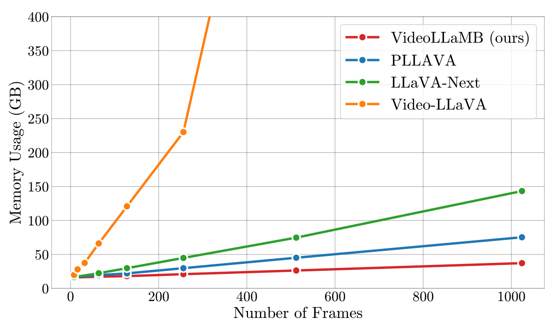

Figure 4. **GPU Memory Cost.** We apply all the experiments on a single NVIDIA A800 GPU.

3.7. Performance Analysis

Memory Cost Our model's recurrent strategy maintains a consistent visual input length to the LLM, significantly reducing GPU memory usage. While a larger memory cache theoretically requires more memory [66], the impact is minimal due to shorter memory tokens compared to visual input tokens. The recurrent memory operates on the bridge layer, minimizing intermediate costs. In our experiments on the EgoSchema [41] dataset, we compared our model against three categories: vanilla, pooling-based, and sampling-based. Results in Figure 4 show that our methods and other fixed-length input models significantly cut memory usage, with our approach compressing input length more effectively. Our design's efficiency is evident, as the memory cache incurs negligible additional memory cost.

Methods	LLM	Inference Time (s) ↓	Score ↑
MovieChat	Vicuna-7B	143.7	-
MALMM	Vicuna-7B	14.5	3.39
LLaVA-NeXT-Video-DPO	Vicuna-7B	11.1	1.72
PLLaVA	Vicuna-7B	7.4	1.82
VideoLLaMB (Ours)	Vicuna-7B	4.21	5.73

Table 6. **Average Inference Time** on the 300-second videos from NIAVH. The score is the average score on NIAVH.

Inference Time Our primary concern with our approach is the potential time expenditure associated with recurrent processes and memory retrieval. To address this, we conducted experiments to assess the efficiency of our method in comparison to others. The evaluation included all current methods capable of handling long videos. We tested each model on NIAVH with 300 second video cases to measure their performance for comparison. The results Table 6, demonstrate that our method not only outperformed the existing methods but did so even when compared to those employing a pooling strategy [70]. We attribute this improved performance to the efficient memory management mechanism integrated within the bridge layer of our method. This enables our approach to condense the visual input more effectively than others, resulting in shorter processing times of the LLM. We further analyze the composition of the inference time over videos of varying lengths, including the encoding time and generation time.

Video Duration	Feature Process	Generation	All
30	0.25	1.65	1.9
300	2.3	1.91	4.21
3000	23.4	8.1	31.5

Table 7. **Latency Analysis (sec.)**. We evaluate the inference time of different parts on different length videos.

Method	Accuracy	Δ
w.o. recurrent (mean pooling)	51.61	-2.19
w.o. recurrent (adaptive pooling)	49.4	-4.4
w.o. retrieval	52.2	-1.6
w.o. segment (uniform segment)	52	-1.8
w.o. mixture of images	49.8	-4.0
memory tokens only	50.4	-3.4
$k = 8$	52.8	-1.0
VideoLLaMB (Ours)	**53.8**	

Table 8. **Ablated results on the effects of different modules.**

3.8. Ablation Study

In this section, we present an ablation study of our method, focusing on its individual components. We analyze our method the EgoSchema dataset. The corresponding results are detailed in Table 8. Initially, we assess the effectiveness of the recurrent mechanism. To do this, we replace this mechanism with two pooling strategies: mean pooling and adaptive pooling. For comparison purposes, we configure the adaptive pooling strategies to produce a target time sequence length of 4, matching our method's settings. Our findings reveal that all pooling strategies cause a notable degradation in performance. Notably, the adaptive pooling strategy underperforms even mean pooling. We hypothesize that this discrepancy arises from differences in how training and inference are conducted; mean pooling, being more consistent, likely enhances the model's generalizability. We then evaluate the memory retrieval mechanism and observe that it is indeed capable of preserving memory to a certain degree. Lastly, we examine the impact of our semantic segmentation strategy. Compared to a uniform segmentation approach, our method is more adept at dividing videos into semantic segments. This segmentation results in a more efficient preservation of information, mitigating the information loss typically associated with sampling strategies.

4. Related Work

Long Video Language Understanding The advancement of large language models (LLMs) has significantly improved the comprehension of long videos through their interaction with human language. Current methodologies for long video analysis are categorized into scaling-up approaches, agent-based techniques, and length extrapolation strategies. Scaling-up approaches involve increasing model parameters and expanding training datasets [37], or developing more efficient architectures to replace computationally intensive transformers [6, 31], though these may not always be feasible. Agent-based techniques leverage LLMs' strategic planning by incorporating various visual experts for comprehensive understanding [10, 13, 56, 63] or converting visual inputs into textual descriptions [58, 59, 72, 75], but they can encounter efficiency issues and challenges with out-of-domain content. Length extrapolation extends image-language and short video-language modeling to longer durations using techniques such as temporal embeddings [46], prompts [49], position encodings [62, 64], frame condensation [54], visual token compression [24, 38, 39], and retrieval-based methods with visual features [20]. These often involve selective sampling, which risks information loss. Our work introduces a recurrent memory strategy to encode entire video sequences, using a memory cache to retain past memory and project the memory-augmented current semantic segment into the LLM to maintain long video understanding.

Anticipatory Video Planning Anticipatory planning, which involves predicting future actions based on past sequences and current context, has been validated as effective in language models [12, 42, 53]. This approach is analogous to video understanding, where action anticipation based on visual data is gaining prominence [14, 17, 51]. A growing research area is the intersection of action anticipation and goal-directed planning, enhancing AI capabilities in video understanding [9, 45, 80]. This challenge is especially critical in real-time streaming environments, where systems must interpret the current state and retain an extensive memory of past events to inform decision-making. Our proposed method is well-suited to address these challenges.

5. Conclusion

VideoLLaMB presents a significant advancement in video-language modeling by improving both computational efficiency and long-context understanding. Through the introduction of Memory Bridge Layers with recurrent memory tokens and the SceneTiling algorithm, our approach effectively preserves essential visual cues and maintains semantic continuity across extended video sequences. Empirical evaluations show that VideoLLaMB consistently outperforms existing methods in tasks such as long VideoQA, egocentric planning, and frame retrieval. Looking ahead, we aim to explore the integration of LLM memory with the bridge memory, with a focus on preserving the system's overall efficiency.

Acknowledgments This work was supported by the grants from the National Natural Science Foundation of China (62376031, 62372014).

References

[1] Ivana Balazevic, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J. Hénaff. Memory consolidation enables long-context video understanding. *CoRR*, abs/2402.05861, 2024. 5

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 6

[3] S. Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2907–2917, 2022. 5

[4] Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[5] Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Scaling transformer to 1m tokens and beyond with RMT. *CoRR*, abs/2304.11062, 2023. 2

[6] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *CoRR*, abs/2403.09626, 2024. 8

[7] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18407–18418, 2024. 2, 6

[8] Shixing Chen, Xiaohan Nie, David D. Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9791–9800, 2021. 2

[9] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplanbench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*, abs/2312.06722, 2023. 2, 5, 8

[10] Rohan Choudhury, Koichiro Niinuma, Kris M. Kitani, and László A. Jeni. Zero-shot video question answering with procedural programs. *CoRR*, abs/2312.00937, 2023. 8

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5

[12] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, pages 8469–8488. PMLR, 2023. 8

[13] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision (ECCV)*, pages 75–92. Springer, 2024. 8

[14] Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. Long-term anticipation of activities with cycle consistency. In *Pattern Recognition - 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28 - October 1, 2020, Proceedings*, pages 159–173. Springer, 2020. 8

[15] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm, 2024. 2

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 24108–24118, 2025. 2, 4

[17] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *J. Vis. Commun. Image Represent.*, 49:401–411, 2017. 8

[18] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. MIST : Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14773–14783. IEEE, 2023. 5

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawa-

har, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18973–18990. IEEE, 2022. 6

[20] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13504–13514, 2024. 1, 4, 6, 7, 8

[21] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguistics*, 23(1): 33–64, 1997. 2

[22] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision (ECCV)*, pages 709–727. Springer, 2020. 2

[23] Zixia Jia, Mengmeng Wang, Baichen Tong, Song-Chun Zhu, and Zilong Zheng. LangSuit·E: Controlling, planning, and interacting with large language models in embodied text environments. In *Findings of the Association for Computational Linguistics: ACL-Findings 2024*, 2024. 1

[24] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *European Conference on Computer Vision (ECCV)*, pages 271–288. Springer, 2024. 8

[25] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Y. Sorokin, and Mikhail Burtsev. In search of needles in a 11m haystack: Recurrent memory finds what llms miss. *CoRR*, abs/2402.10790, 2024. 2

[26] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 487–507. Association for Computational Linguistics, 2023. 1

[27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. 5

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2022. 2, 5

[29] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. 3

[30] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023. 4, 5

[31] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision (ECCV)*, pages 237–255. Springer, 2024. 8

[32] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 5

[33] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision (ECCV)*, pages 323–340. Springer, 2024. 5

[34] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984. Association for Computational Linguistics, 2024. 1, 4, 5

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5

[36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 2

[37] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[38] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. ST-LLM: large language models are effective temporal learners. In *European Conference on Computer Vision (ECCV)*, pages 1–18. Springer, 2024. 8

[39] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13151–13160, 2024. 1, 8

[40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 5

[41] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 4, 7

[42] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via

embodied chain of thought. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 8

[43] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Bassl: Boundary-aware self-supervised learning for video scene segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 4027–4043, 2022. 2

[44] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14386–14397, 2024. 5

[45] Dhruvesh Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, and Ruta Desai. Pretrained language models as visual planners for human assistance. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15256–15268. IEEE, 2023. 8

[46] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024. 8

[47] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024. 1, 4, 5

[48] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10143–10152, 2020. 2

[49] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2024. 8

[50] Gabriele Ruggeri and Debora Nozza. A multi-dimensional study on bias in vision-language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6445–6455. Association for Computational Linguistics, 2023. 1

[51] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *International Conference on Computer Vision (ICCV)*, pages 862–871. IEEE, 2019. 8

[52] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 26160–26169, 2025. 4, 5

[53] Chan Hee Song, Brian M. Sadler, Jiaman Wu, Wei-Lun Chao, Clayton Washington, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *International Conference on Computer Vision (ICCV)*, pages 2986–2997. IEEE, 2023. 8

[54] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, 2024. 1, 4, 8

[55] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[56] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision (ECCV)*, pages 142–160. Springer, 2024. 8

[57] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37: 20540–20565, 2024. 6

[58] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, pages 58–76. Springer, 2024. 8

[59] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in egocentric videos. *CoRR*, abs/2312.05269, 2023. 8

[60] Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5036–5048. Association for Computational Linguistics, 2023. 2

[61] Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. Vstar: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 1

[62] Yuxuan Wang, Yueqian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, Yang Liu, and Zilong Zheng. Efficient temporal extrapolation of multimodal large language models with temporal grounding bridge. In *The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 8

[63] Yuxuan Wang, Alan Yuille, Zhuowan Li, and Zheng Zilong. Exovip: Step-by-step verification and exploration with exoskeleton modules for compositional visual reasoning. In *The first Conference on Language Modeling (CoLM)*, 2024. 8

[64] Yu Wang, Zeyuan Zhang, Julian J. McAuley, and Zexue He.

LVCHAT: facilitating long video comprehension. *CoRR*, abs/2402.12079, 2024. 8

[65] Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2025. 1

[66] Tong Wu, Junzhe Shen, Zixia Jia, Yuxuan Wang, and Zilong Zheng. Tokenswift: Lossless acceleration of ultra long sequence generation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 7

[67] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786. Computer Vision Foundation / IEEE, 2021. 2, 4

[68] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision (ECCV)*, pages 39–58. Springer, 2022. 5

[69] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-vlm: Slowfast slots for video-language modeling. *ArXiv*, 2024. 2

[70] 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. *CoRR*, abs/2404.16994, 2024. 1, 2, 4, 5, 6, 7

[71] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1686–1697, 2021. 5

[72] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *CoRR*, abs/2401.08392, 2024. 8

[73] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023. 5

[74] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5

[75] 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, 2024. 8

[76] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 543–553. Association for Computational Linguistics, 2023. 1, 5, 6

[77] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision, 2024. 6, 7

[78] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[79] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2, 4, 5, 6

[80] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. AntGPT: Can large language models help long-term action anticipation from videos? In *The Twelfth International Conference on Learning Representations*, 2024. 8

[81] Xinyue Zheng, Haowei Lin, Kaichen He, Zihao Wang, Zilong Zheng, and Yitao Liang. MCU: An evaluation framework for open-ended game agents. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 1

[82] Kankan Zhou, Eason Lai, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, pages 527–538. Association for Computational Linguistics, 2022. 1

[83] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18243–18252, 2024. 2, 6

[84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592, 2023. 5