Memory-enhanced Retrieval Augmentation for Long Video Understanding

Huaying Yuan[1], Zheng Liu[2], Minhao Qin[3], Hongjin Qian[2], Y Shu[4], Zhicheng Dou[1*], Ji-Rong Wen[1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Academy of Artificial Intelligence
[3]Institute of Automation, Chinese Academy of Sciences, [4]University of Trento
{hyyuan, dou}@ruc.edu.cn

Abstract

Retrieval-augmented generation (RAG) shows strong potential in addressing long-video understanding (LVU) tasks. However, traditional RAG methods remain fundamentally limited due to their dependence on explicit search queries, which are unavailable in many situations. To overcome this challenge, we introduce a novel RAG-based LVU approach inspired by the cognitive memory of human beings, which is called ***MemVid****. Our approach operates with four basics steps:* ***memorizing*** *holistic video information,* ***reasoning*** *about the task's information needs based on the memory,* ***retrieving*** *critical moments based on the information needs, and* ***focusing*** *on the retrieved moments to produce the final answer. To enhance the system's memory-grounded reasoning capabilities and achieve optimal end-to-end performance, we propose a* ***curriculum learning*** *strategy. This approach begins with supervised learning on well-annotated reasoning results, then progressively explores and reinforces more plausible reasoning outcomes through reinforcement learning. We perform extensive evaluations on popular LVU benchmarks, including MLVU, VideoMME and LVBench. In our experiment, MemVid significantly outperforms existing RAG-based methods and popular LVU models, which demonstrate the effectiveness of our approach. Our model and source code will be made publicly available upon acceptance.*

1. Introduction

Long-video understanding plays a crucial role in real-world applications such as video analysis, autonomous driving, and embodied AI. However, this task presents significant challenges for traditional multimodal large language models (MLLMs), which are primarily designed for short visual inputs, including single images, multiple images, or short videos. While recent advancements have extended the context window for MLLMs, existing techniques still struggle with suboptimal performance due to the inherent complexity of long-video understanding and the high computational costs involved in processing extended visual sequences.

Retrieval-augmented generation (RAG) has emerged as a promising approach for handling long-sequence problems. With the retrieval of useful information from long-sequence data, the model can perform its generation based on a highly simplified input, thus enabling cost-effective fulfillment of its task. While traditional RAG methods excel at addressing clearly specified queries, such as "*When did a boy chase a dog?*", they are insufficient for general long-video understanding problems which often involves implicit and complex information needs. For instance, consider the query "*What event is shared by two families as depicted in the video?*" Instead of directly resorting to a retrieval model for relevant moments, the model must first identify the two families, track their individual activities, and then determine the common event before arriving at a response. This process involves reasoning beyond straightforward retrieval, highlighting the limitations of conventional RAG techniques in handling dispersed and context-dependent information within long videos.

In contrast, humans approach long-video understanding problems far more effectively. They will first go through the entire video, forming the memorization for the holistic information about its content. When faced with a specific question, they reason about the problem, determining what information is relevant. Only then do they retrieve key moments from memory, focusing on those relevant details to arrive at a final answer. This structured process, including integrating comprehension, reasoning, and targeted retrieval, enables humans to handle complex long-video understanding tasks with remarkable proficiency.

With the above inspiration, we propose a novel RAG framework for long-video understanding, called **MemVid** (**Mem**ory-enhanced retrieval augmentation for long **Vid**eo understanding). MemRAG operates with four basic steps. First, it generates the memory for the holistic information

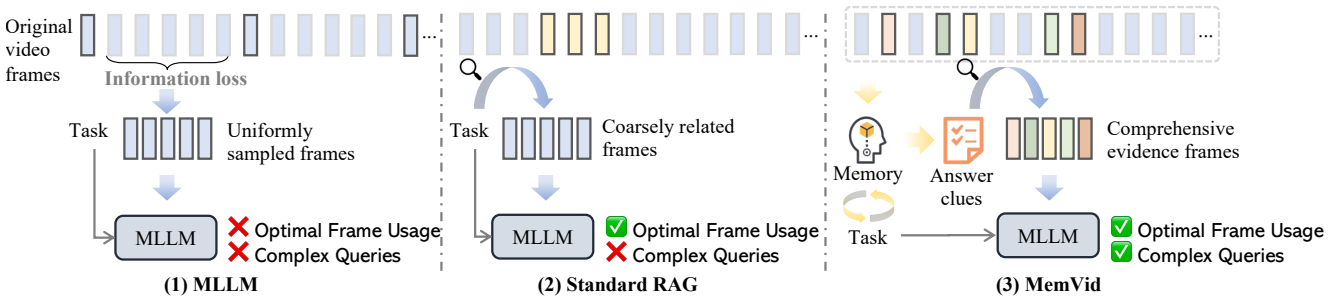


Figure 1. Comparison of different frameworks. Given a specific task, MLLMs uniformly sample frames as input, leading to significant information loss due to brute-force downsampling. Standard RAG mitigates this issue but is limited to clearly specified queries. In contrast, MemVid introduces a cognitive-inspired memorizing and reasoning process, enabling both clearly specified and complex queries thus enabling cost-effective fulfillment of its task.

of the long video. Second, it reasons about the information needs for a concrete problem based on its memory. Third, it retrieves crucial moments from the video as required by the information needs. And lastly, it generates the final answer based on the retrieval results.

The above workflow is driven by three essential modules: the memorizer, the retriever, and the generator. In our work, we focus on optimizing the memorizer while keeping other modules fixed. To achieve optimal end-to-end performance, we introduce a curriculum learning strategy. Our training process begins with supervised learning, where the memorizer is trained to generate well-structured reasoning outputs based on high-quality annotations obtained from powerful long-video MLLMs. Once this foundation, the memorizer explores various candidate reasoning trajectories, reinforcing those that lead to high-quality answers. This approach ensures a progressive refinement of reasoning capabilities, ultimately enhancing the system's overall performance.

To assess the effectiveness of MemVid, we conduct comprehensive experiments using various long-video understanding (LVU) benchmarks, including VideoMME[1], MLVU[2], and LVBench [3]. In our experiment, MemVid achieves notable advantages against existing RAG-based LVU methods, while significantly improves the cost-effectiveness in comparison to popular long-video MLLMs.

In summary, our contributions are threefold:

1. We introduce MemVid, a novel retrieval augmentation framework tailored for long-video understanding. To the best of our knowledge, this is the first approach of its kind, emphasizing the critical role of reasoning and memorization in comprehending long videos.
2. We design an effective curriculum learning strategy that enhances the memorizer's ability to improve end-to-end RAG performance by leveraging diverse training signals.
3. We conduct extensive experiments, which showcase MemVid's ability to achieve high-quality results while significantly improving the cost-effectiveness in long-video understanding.

2. Related Work

2.1. Video Understanding

Recent advances in image-based multimodal large language models (MLLMs) [4–6] have inspired extensions to video understanding. A common paradigm involves encoding video frames into visual tokens using pretrained encoders such as CLIP, followed by aligning these tokens with text through feature fusion layers. For example, methods like ST-LLM [7] and VideoChat [8] utilize Q-Former [6], a transformer-based cross-modal aggregator, to merge spatial-temporal features, while VideoLLaVA [9] and MiniGPT4-Video [10] adopt lightweight linear projections for alignment. To mitigate computational overhead, spatial-temporal pooling techniques, such as those in Video-ChatGPT [11], compress visual tokens by averaging features across frames or regions. Despite their effectiveness in capturing short-term spatial semantics (e.g., object recognition in clips), these methods struggle with long-range temporal dependencies.

2.2. Long Video Understanding

Recent advances in long video understanding have focused on two primary paradigms: compression-based methods and retrieval-augmented approaches. Compression-based techniques aim to address temporal redundancy through various mechanisms. Memory-augmented models like MovieChat [12] and MA-LMM [13] leverage memory banks to store historical visual features, though periodic consolidation risks losing fine-grained spatial details. Token reduction strategies, exemplified by LLAMA-VID [14], compress frames into two tokens via context attention modules to model inter-frame dependencies, but sacrifice spatial resolution in the process. Hierarchical alignment methods such as LongVLM [15] attempt to merge local moment features with global context through token-wise alignment, though balancing efficiency and accuracy remains challenging.

In addition to compression-based approaches, retrieval-augmented generation (RAG) methods have gained traction. While RAG has demonstrated effectiveness in language

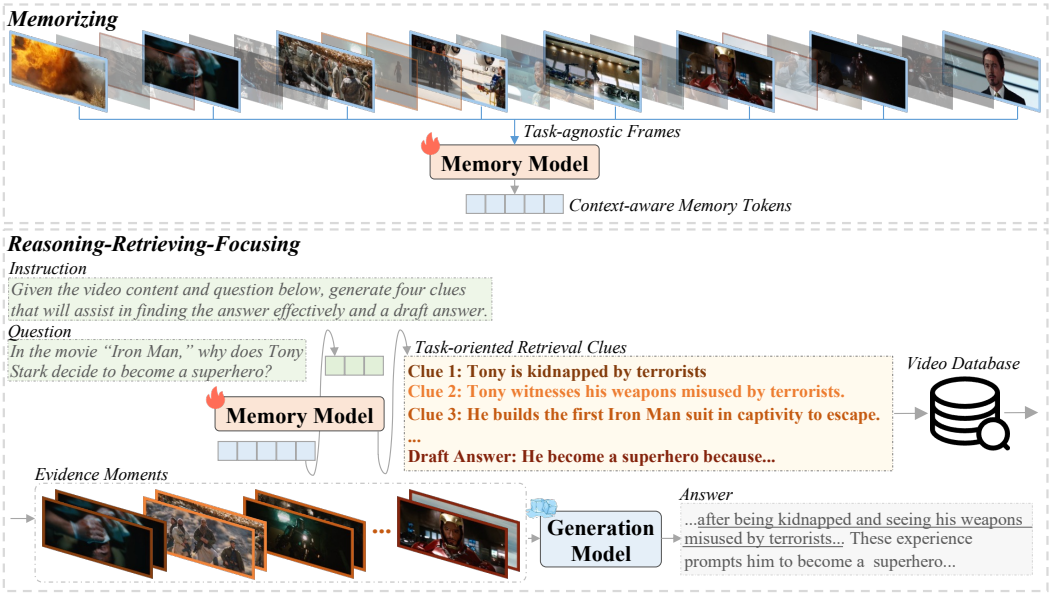


Figure 2. Overview of MemVid. Firstly, MemVid compresses long videos into a flexible, context-rich global memory. Given a question, it dynamically infers retrieval clues to resolve semantic ambiguities and structural complexities. These refined and explicit clues are used to retrieve the most relevant evidence frames, which are passed to downstream MLLMs. Under the constraint of limited context, this approach ensures that MLLMs focus on the most informative evidence frames, leading to more accurate responses.

tasks [16, 17], its adaptation to long videos introduces unique limitations. Methods like DrVideo [18] and Goldfish [19] convert moments into textual descriptions for retrieval, inevitably discarding critical visual information such as facial expressions and spatial layouts. VideoRAG [20] further compounds this issue by relying on auxiliary expert models (e.g., OCR, scene graph) for video indexing, which introduces significant latency during online question answering and struggles to generalize beyond predefined domains.

A key limitation of existing methods is their narrow focus: compression-based approaches extend the window but overlook frame redundancy, while retrieval-based systems reduce redundancy but struggle with implicit and complex information needs. Our framework bridges these gaps by unifying memory-enhanced reasoning with adaptive retrieval, mirroring human cognitive strategies.

3. Methodology

3.1. Overview of MemVid

Long video understanding aims to answer a question Q about a video $\mathcal{V}$ composed of N frames. Current MLLM methods naively downsample $\mathcal{V}$ to select a sparse subset $\mathcal{S}$ of k frames for answer generation:

$$A = \mathcal{M}(\mathcal{S}, Q \,|\, \theta), \tag{1}$$

where θ denotes the parameters of the answer generator. However, this brute-force subsampling incurs significant information loss due to $k \ll N$, especially for long videos. To mitigate this, RAG methods employ a question-guided retrieval strategy to locate relevant moments:

$$\mathcal{S}' = \text{Top-}k\,(\mathcal{V}, Q \,|\, \omega)\,, \tag{2}$$

where ω represents the retriever's parameters, and Top-$k(\cdot)$ selects k frames most relevant to Q. The answer is then generated as:

$$A' = \mathcal{M}(\mathcal{S}', Q \,|\, \theta). \tag{3}$$

Despite improvements, RAG struggles with implicit or complex questions, as relying solely on Q often leads to suboptimal retrieval. To tackle these challenges, we propose **MemVid**, a memory-enhanced framework that emulates human-like cognition by constructing a global video memory to guide context-aware retrieval. MemVid operates in four stages (Figure 2):

1. **Memorizing:** Encode the entire video into a structured memory $\mathcal{M}$ to capture long-range dependencies and contextual retrieval clues:

$$\mathcal{M} = \mathcal{R}(\mathcal{V} \,|\, \phi), \tag{4}$$

where $\mathcal{R}$ is the memory model parameterized by ϕ.

2. **Reasoning:** Leverage the memory model $\mathcal{R}$ to infer latent information needs $\mathcal{C}$ by reasoning over Q and the preprocessed global memory $\mathcal{M}$:

$$\mathcal{C} = \mathcal{R}(Q \,|\, \mathcal{M}, \phi). \tag{5}$$

This step decomposes the query intent through context-aware reasoning of Q, producing retrieval clues $\mathcal{C} = \{c_1, \ldots, c_m\}$.

3. **Retrieving:** Segment long video into moments $\mathcal{V} = \{s_1, \ldots, s_M\}$ and retrieve moments relevant to each retrieval clue $c_i \in \mathcal{C}$ and aggregate results:

$$\mathcal{S}'' = \bigcup_{c \in \mathcal{C}} \text{Top-}k\left(\mathcal{V}, c \,|\, \omega\right). \tag{6}$$

4. **Focusing:** Synthesize the final answer using retrieved informative evidence moments $\mathcal{S}''$ and the original question:

$$A'' = \mathcal{M}(\mathcal{S}'', Q \,|\, \theta). \tag{7}$$

By incorporating a memory module designed to perform both memorization and reasoning tasks, MemVid overcomes the limitations of traditional frame-level subsampling and direct question-based retrieval.

3.2. Reasoning-oriented Memory Module

The memory module achieves two objectives: (1) constructing holistic understanding of video semantics, and (2) generating task-aware clues through memory-driven reasoning. This is implemented by three-stage memory processing:

Given an input video $\mathcal{V} \in \mathbb{R}^{T \times H \times W \times 3}$, where T frames are uniformly sampled from the original video, a pretrained visual encoder E_v is used to compress raw video pixels into token-like visual features:

$$F = E_v(\mathcal{V}) \in \mathbb{R}^{(T \times K) \times d_v}, \tag{8}$$

where T denotes the frame number, K denotes the token number of each frame, and d_v the feature depth. While F captures global semantic patterns, it lacks explicit reasoning capabilities for downstream tasks.

To enable the reasoning capabilities of memory, we further convert visual features F into reasoning-oriented key-value (KV) caches using a causal transformer-based language model Θ. We convert reasoning instructions into token embeddings using the embedder E_q, obtaining $\{x_1, \ldots, x_p\}$, while the visual features are treated as token embeddings $\{x_{p+1}, \ldots, x_{p+T\times K+1}\}$, and the total input can be represented as $X = \{x_1, ..., x_{p+T\times K+1}\}$. The input X is processed by a transformer-based model, and the key-value cache $[K, V]$ is generated as follows:

$$K_1 = \mathcal{W}_k X_1, \quad V_1 = \mathcal{W}_v X_1. \tag{9}$$

For each timestep $t \in [1, p + T \times K + 1]$, we compute the new key and value as:

$$K_t = \mathcal{W}_k X_t, \quad V_t = \mathcal{W}_v X_t. \tag{10}$$

The KV cache is then updated by concatenating the new key-value pairs with the previous ones:

$$K \leftarrow \text{Concat}(K, K_t), \quad V \leftarrow \text{Concat}(V, V_t). \tag{11}$$

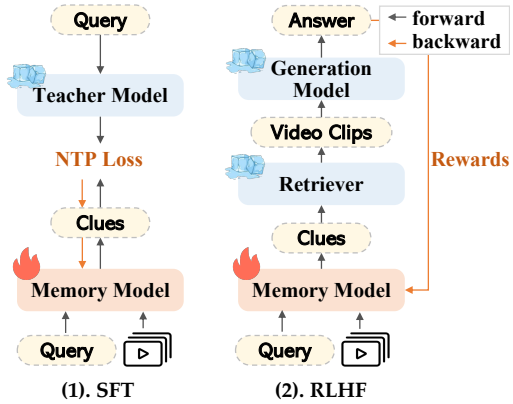


Figure 3. Illustration of the curriculum learning strategy.

The resulting memory comprises all KV states from both reasoning instructions and visual features: $M = \{K, V\}$.

When a question Q arrives, we concatenate the precomputed memory M with the question embedding $E_q(Q)$ and perform a single-pass reasoning:

$$\mathcal{C} = \Theta\left(\text{Concat}(M; E_q(Q))\right), \tag{12}$$

where $\mathcal{C}$ denotes the generated clue set. As illustrated in Figure 2, this design allows the system to dynamically contextualize Q within the video's holistic memory while maintaining computational efficiency.

3.3. Memory-Guided Retrieval

Several approaches can be used to construct a retrieval pool from a long video, such as dense captioning into text [18, 19], splitting the video into frames, or segmenting it into moments. Among these methods, dense captioning suffers from modality information loss, while frame-based approaches fail to preserve the temporal dependencies inherent in videos. Considering these limitations, we segment videos into non-overlapping fixed-duration moments to form a searchable candidate pool, maintaining both temporal and visual context. The retrieval process aggregates results from all retrieval clues in $\mathcal{C}$. For each moment s_j, we compute its similarity with each clue $c \in \mathcal{C}$, as formulated in Equation 6, and aggregate the retrieved results as input for informative moments. To construct a balanced input, we sample frames from informative moments and reorder them chronologically, forming a set of informative frames that occupy a portion α of the total context constraint. To retain global information, we supplement the remaining $1 - \alpha$ portion with uniformly sampled frames from the entire video. All frames are then organized in temporal order and fed into the downstream generation model. This approach enables the model to preserve global context while emphasizing evidence frames, ultimately enhancing accurate content understanding.

3.4. Curriculum Learning Strategy

Within the memorizer-retriever-generator framework, the current retriever and generator modules are considered sufficiently powerful. In this paper, our primary focus is on optimizing the memorizer while freezing the other two components. Since the memorizer indirectly serves the retriever and generator rather than directly generating the final answer, its optimization poses significant challenges. To achieve effective end-to-end optimization of the memorizer, we propose a curriculum learning strategy that integrates supervised fine-tuning and reinforcement learning:

Supervised Fine-Tuning. We first employ supervised fine-tuning (SFT) to establish task understanding and clue-generation capabilities. Since direct supervision data is unavailable, we generate candidate clues using a powerful multi-modal large language model (MLLM) with enhanced filtering: the MLLM uses identical prompts but undergoes stricter validation, where only clues enabling correct downstream retrieval and generation are retained. This produces high-quality training data for optimizing the memorizer through Next Token Prediction:

$$\mathcal{L}_{\text{NTP}}(\theta) = -\sum_{t=1}^{T} \log P_\theta(w_t \mid w_{<t}), \tag{13}$$

where $P_\theta(w_t \mid w_{<t})$ predicts token w_t given context $w_{<t}$. The resulting model demonstrates robust task comprehension and structured clue generation.

Reinforcement Learning. Building on SFT initialization, we refine clue quality through direct preference optimization (DPO) [21]. For each query, we sample multiple clues from the SFT model and rank them by the generator's correctness probability. Preference pairs (y_i^+, y_i^-) are constructed with a minimum score margin τ to ensure quality distinction. The optimization follows:

$$\mathcal{L}_{\text{dpo}} = -\sum_i \log \sigma \left(\beta \cdot \left[\log \frac{\pi_\theta(y_i^+)}{\pi_{\text{S}}(y_i^+)} - \log \frac{\pi_\theta(y_i^-)}{\pi_{\text{S}}(y_i^-)} \right] \right), \tag{14}$$

where π_θ is the learnable policy, π_{S} the frozen SFT reference, and β controls policy divergence. This dual-stage approach accelerates convergence while aligning clues with downstream objectives, achieving stronger generalization than single-stage methods.

4. Experiments

4.1. Experimental Settings

4.1.1. Benchmark and Metrics

We conducted comprehensive experiments on three commonly-used long video benchmarks:

VideoMME [1] consists of 2,700 expert-curated questions linked to 900 diverse videos of different lengths: short (up to 2 minutes), medium (4 to 15 minutes), and long (30 to 60 minutes). It provides two versions of questions, with subtitles and without subtitles.

MLVU [2] is a comprehensive video dataset that includes videos ranging from 3 minutes to 2 hours in duration. It encompasses a diverse set of nine tasks, including action recognition, event localization, and counting, designed to evaluate both global and local video understanding.

LVBench [3] is designed for extremely long videos, with an average duration of 4,101 seconds. It features a diverse set of tasks, including key information retrieval, event understanding, event recognition, temporal grounding and so on, all supported by high-quality human annotations.

4.1.2. Baselines

We evaluate MemVid against a wide range of baselines, which are catorized into four groups:

1. Proprietary Models: This category includes state-of-the-art closed-source models such as GPT-4V [22], GPT-4o [23], and Gemini-1.5-Pro [24], which have demonstrated strong performance in multimodal tasks. While these models achieve high scores, their closed nature limits direct architectural comparisons.

2. Open-Source MLLMs: We compare against general-purpose open-source MLLMs [8, 9, 25–27], which represent the average performance of video understanding models.

3. Open-Source Long-Context MLLMs: We further include state-of-the-art long-context MLLMs [28–33], which extend the context length of traditional MLLMs, enabling them to process longer videos effectively.

4. RAG-based MLLMs: Relevant works include Goldfish [19], SALOVA-Qwen [34], and Video-RAG [20]. Goldfish convert videos into a text corpus and retrieve key text while SALOVA-Qwen leverage a moment retrieval router to locate key moments. Video-RAG relies on audio and subtitles rather than retrieving key moments, making it orthogonal to our approach, so we do not compare against it. Additionally, we implement a $\text{RAG}_{\text{simple}}$ variant as a reference baseline, identical to our model but without the memory module.

4.1.3. Implementation Details

Our pipeline begins with momentation into 10-second moments using LanguageBind-Large [35] for cross-modal retrieval and text retrieval. The memory model generates 4 query-aware clues and a draft answer along with the original question, subsequently retrieving top-4 moments (reordered chronologically and sampled at 1 FPS), top-4 subtitle moments, and 51 globally uniform frames to preserve temporal context. Inputs are truncated to 128 frames/subtitles for computational fairness. For $\text{RAG}_{\text{simple}}$, we concatenate each question with its corresponding choice to form a query. We

Table 1. Experimental results on MLVU-test and VideoMME benchmarks. † indicates that results are reproduced using their official weights.

Model	Size	MLVU	VideoMME w/o subtitle				VideoMME w subtitle				Avg
		M-avg	Short	Medium	Long	Avg	Short	Medium	Long	Avg	
Proprietary Models											
GPT-4V [22]	-	43.3	70.5	55.8	53.5	59.9	73.2	59.7	56.9	63.3	55.5
GPT-4o [23]	-	54.9	80.0	70.3	65.3	71.9	82.8	76.6	72.1	77.2	68.0
Gemini-1.5-Pro [24]	-	-	**81.7**	**74.3**	**67.4**	**75.0**	**84.5**	**81.0**	**77.4**	**81.3**	-
Open-source VLMs											
VideoChat2 [8]	7B	35.1	48.3	37.0	33.2	39.5	52.8	39.4	39.2	43.8	39.5
VideoLLaVA [9]	7B	30.7	45.3	38.0	36.2	39.9	46.1	40.7	38.1	41.6	37.4
ShareGPT4Video [25]	7B	33.8	48.3	36.3	35.0	39.9	53.6	39.3	37.9	43.6	39.1
Qwen2VL† [26]	7B	52.9	68.0	58.4	47.9	58.1	70.7	66.2	53.4	63.4	58.1
InternVL-Chat-V1.5 [27]	20B	37.3	60.2	46.4	45.6	47.8	61.7	49.1	46.6	52.4	45.8
Open-source Long VLMs											
Kangaroo† [28]	7B	44.4	66.1	55.3	46.7	56.0	68.0	55.4	49.3	57.6	52.7
LongVA [29]	7B	41.1	61.1	50.4	46.2	52.6	61.6	53.6	47.6	54.3	49.3
LongVILA† [30]	7B	49.0	69.3	56.1	47.0	57.5	70.4	59.2	52.1	60.6	55.7
Video-CCAM [31]	14B	42.9	62.2	50.6	46.7	53.2	66.0	56.3	49.9	57.4	51.2
LongLLaVA [32]	7B	-	61.9	51.4	45.4	52.9	66.2	54.7	50.3	57.1	-
Video-XL [33]	7B	45.5	64.0	67.4	53.2	55.5	60.7	49.2	54.9	61.0	54.0
RAG-based VLMs											
Goldfish† [19]	7B	37.3	28.7	29.4	28.7	28.9	28.6	26.4	27.3	27.4	31.2
SALOVA-Qwen [34]	7B	-	52.3	50.9	46.8	50.0	-	-	-	-	-
RAG_{simple}	7B	54.9	<u>73.7</u>	59.2	52.0	61.6	74.6	63.2	53.6	63.8	60.1
MemVid	7B	**58.1**	**73.9**	**63.1**	**55.0**	**64.0**	**75.4**	**64.6**	**57.1**	**65.7**	**62.6**

then retrieve the top-2 moments for each question-choice query. For supervised fine-tuning, we generate 10,000 synthetic clues and draft answers using Qwen2VL-72B from TVQA-Long [19], NExT-QA [36], and ActivityNet-QA [37]. We filter out clues that correctly answer the questions and use them as high-quality training labels. For DPO training, we sample different answers from VICO [33], curated by VideoXL. We generate diverse clue variations for each sample and evaluate their quality using a frozen video understanding model's confidence scores on correct answers. To construct 1,000 high-quality training pairs, we retain samples where positive clues achieve a confidence score above 0.7 for the correct answer, while negative clues score below 0.3. All experiments are conducted on a single node with 8 * A800 (80GB) GPUs.

4.2. Overall Results

We evaluate MemVid against baseline models on three benchmarks (Table 1 and Table 2). Overall, MemVid establishes new state-of-the-art performance among 7B models, outperforming both conventional long video MLLMs and specialized RAG approaches. Specifically:

(1) **Comparison with General MLLMs.** Using the same generation model and input frame limits as Qwen2VL, MemVid achieves a +9.8% relative gain on MLVU and +10.2%/+3.6% improvements on VideoMME without/with subtitles. This indicates that, under identical generation conditions, MemVid more effectively locates crucial frames for answering questions compared to uniform sampling.

(2) **Comparison with Long-Context MLLMs.** MemVid outperforms the leading context-extension method, VideoXL, by +15.3% and 7.7% on VideoMME without and with subtitles, demonstrating that a carefully selected set of frames via MemVid can surpass models that rely on longer input windows (e.g., 1024 frames).

(3) **Comparison with RAG-based MLLMs.** Goldfish performs poorly due to severe information loss when converting video modalities to text. For a fair comparison, we implement a stronger baseline, RAG_{simple}, using the same retriever and downstream model. Across all tasks, MemVid outperforms RAG_{simple} by +5.8%/+3.9%/+3.0% on MLVU and VideoMME without/with subtitles, validating the effectiveness of our global memory-enhanced clue generation over conventional retrieval augmentation.

Overall, these results confirm that our memory-augmented paradigm can effectively pinpoint key frames, thereby enhancing long video understanding.

4.3. Task-Specific Performance

Table 2 shows MemVid's performance on LVBench. (1) Compared to Qwen2VL (without retrieval), it significantly improves key information retrieval (+62.3%), event understanding (+17.0%), reasoning (+10.8%), and event recognition (+14.9%), demonstrating the advantages of memory-

Table 2. Experimental results on LVBench. KIR, EU, ER, TG, Rea, and Sum represent key information retrieval, event understanding, event recognition, temporal grounding, reasoning, and summarization, respectively.

Model	LLM Params	Tasks						Overall
		KIR	EU	ER	TG	Rea	Sum	
Proprietary Models								
GPT-4o(2024-05-13) [23]	-	34.5	27.4	33.0	25.0	27.5	24.1	30.8
Gemini 1.5 Pro [24]	-	39.3	30.9	32.1	31.8	27.0	32.8	33.1
GLM-4V-Plus [38]	-	34.8	35.8	39.9	37.7	40.0	32.8	38.3
Open-source MLLMs								
TimeChat [39]	7B	25.9	21.7	21.9	22.7	25.0	24.1	22.3
MovieChat [12]	7B	25.9	23.1	21.3	22.3	24.0	17.2	22.5
LLaMA-VID [14]	13B	23.4	21.7	25.4	26.4	26.5	17.2	23.9
PLLaVA [40]	34B	26.2	24.9	25.0	21.4	30.0	25.9	26.1
CogVLM2-Video [41]	8B	31.0	27.1	28.3	25.5	25.5	38.9	28.1
LLaVA-NeXT-Video [42]	34B	34.1	31.2	30.1	31.4	35.0	27.6	32.2
Qwen2-VL† [26]	7B	32.9	34.7	40.3	34.7	39.1	28.1	37.2
RAG_{simple}	7B	38.6	38.6	46.2	33.3	43.0	**29.6**	42.0
MemVid	7B	**53.4**	**40.6**	**46.3**	**34.9**	**43.2**	28.1	**44.4**

Table 3. Ablation study on MLVU and VideoMME (long).

Model	MLVU	VideoMME	
	M-avg	w/o sub.	w sub.
MemVid	**58.1**	**55.0**	**57.1**
w/o. reasoning	54.9	51.6	53.6
w/o. memory	56.2	52.4	53.8
zero-shot	55.2	52.6	54.0
only SFT	56.4	53.7	54.9
only DPO	56.6	54.1	55.6

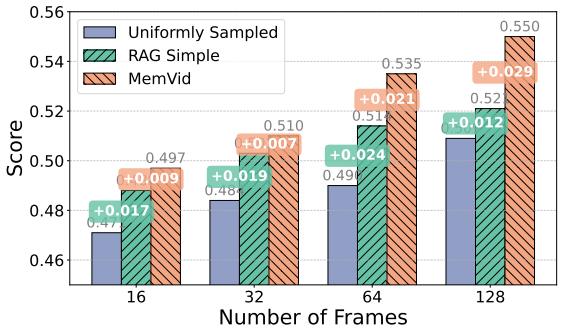


Figure 4. Performance of different frame numbers of downstream MLLM, evaluated on VideoMME (long).

Table 4. Comparison of Long-Video Models and MemVid.

Metric	VideoXL	MemVid	Gain
Frame	1024	64	↓ **93.8%**
Latency(s)	85.0	55.2	↓ **35.1%**
Memory(GB)	56.1	36.5	↓ **34.9%**
Performance	45.6	53.5	↑ **17.3%**

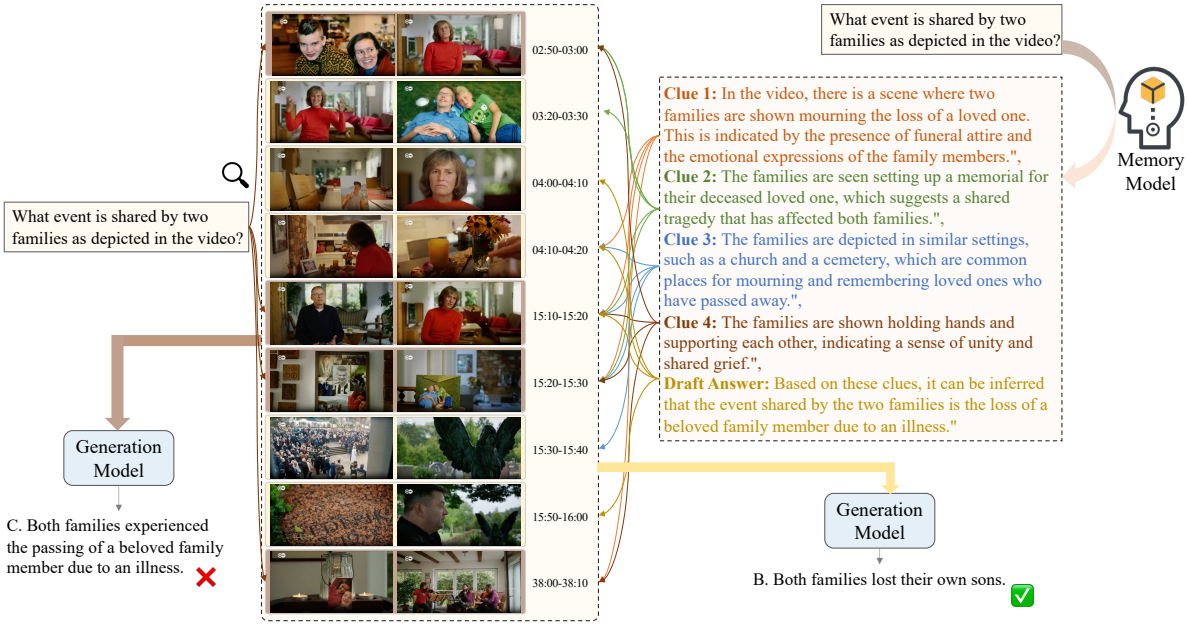
enhanced retrieval. (2) Against RAG_{simple}, MemVid excels in complex temporal modeling, notably in key information retrieval (+38.3%) and event understanding (+5.2%). However, it slightly lags in summarization, as LVBench's explicit summary questions favor direct retrieval.

4.4. Ablation Study

We evaluate MemVid's performance across training stages (Table 3): (1) Memory Mechanism: Compared to two strong RAG-based baselines, $MemVid_{w/o\ reasoning}$ (direct retrieval) and $MemVid_{w/o\ memory}$ (HyDE-inspired retrieval using generated answers [43]), MemVid achieves superior zero-shot performance. While $MemVid_{w/o\ memory}$ improves over $MemVid_{w/o\ reasoning}$ by 2.4% on MLVU and 1.6%/0.4% on VideoMME (without/with subtitles), MemVid surpasses both. DPO training further enhances performance by 3%, confirming the effectiveness of MemVid's memory mechanism. (2) Multi-stage Optimization: Fine-tuning with SFT improves performance by 1% over the zero-shot baseline, serving as a warm-up. DPO further boosts performance by 2%, demonstrating its role in generating rich, useful clues for better long-video understanding.

4.5. Frame Number Analysis

As shown in Figure 4, we compared three frame-selection strategies including uniform sampling, RAG Simple retrieval, and MemVid retrieval, with the same downstream generative model for question answering. The results indicate that MemVid consistently outperforms RAG Simple, with performance gains of about 0.9, 0.7, 2.1, and 2.9 percentage points at 16, 32, 64, and 128 frames respectively. Notably, as the number of frames increases, MemVid's advantage becomes more pronounced. This is because when the input length is short, each 10-second video may contribute only one frame per clue, limiting retrieval effectiveness; with longer inputs, each clue can yield a video clip with sufficient frames, allowing MemVid to fully leverage its strengths.



(a) RAG w/o Memory (b) MemVid

Figure 5. Visualization of MemVid on MLVU. Compared to RAG_{simple}, which retrieves only a limited amount of useful information, MemVid decomposes the problem into fine-grained and explicit clues. These clues guide the retrieval of more comprehensive supporting frames, aiding the downstream generative model in answering questions correctly.

4.6. Efficiency

We evaluate the efficiency and effectiveness of MemVid by comparing it with VideoXL in a long-video understanding task. As shown in Table 4, MemVid significantly reduces input length by selecting the most informative frames, allowing the downstream model to process only 64 frames while reducing GPU memory usage by 34.9% and inference latency by 35.1%, yet achieving 17.3% higher performance compared to VideoXL with 1024 frames. This result highlights that uniform sampling introduces substantial redundancy, whereas MemVid efficiently extracts a compact yet informative subset, enabling faster inference and superior performance.

4.7. Case Analysis

Figure 5 compares methods for the question, "What event is shared by two families?" The baseline RAG_{simple} retrieves moments directly, missing key evidence. MemVid, through context-aware reasoning, captures implicit semantics and retrieves cross-clip evidence, yielding the precise answer.

4.8. Different Downstream Backbones

We apply our context-aware RAG pipeline to various downstream MLLMs, including VILA1.5-3B, LongVA-7B-DPO, and Qwen2VL-72B. As shown in Table 5, MemVid improves performance across models, achieving 10.1% gains on VILA-1.5 (3B, 8 frames), 3.5% on LongVA (7B, 128 frames), and 2.4% on Qwen2VL (72B, 128 frames), with diminishing returns as model size increases. The largest gains on smaller models under sparse inputs suggest that MemVid effectively preserves temporal context, compensating for weaker architectures. Notably, these gains generalize across models despite MemVid being trained solely with Qwen2VL feedback, demonstrating its adaptability without model-specific tuning.

Table 5. Zero-shot application to different downstream generation models with different sizes.

Model	Size	# Frame	Performance	Gain
VILA-1.5	3B	8	32.7	-
+ MemVid	3B	8	36.0	+10.1%
LongVA	7B	128	42.4	-
+ MemVid	7B	128	43.9	+3.5%
Qwen2VL	72B	128	59.0	-
+ MemVid	72B	128	60.4	+2.4 %

5. Conclusion

In this paper, we propose MemVid, a novel RAG-based framework for LVU tasks that overcomes the need for explicit search queries. Inspired by human cognitive memory, MemVid follows four key steps: memorizing, reasoning, retrieving, and focusing. To enhance reasoning and retrieval, we introduce a curriculum learning strategy that refines performance through supervised and reinforcement learning. Comprehensive experiments on benchmarks show that MemVid significantly outperforms existing RAG-based and LVU models, demonstrating its effectiveness.

References

[1] Chaoyou Fu, Yuhan Dai, Yondong 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. *arXiv preprint arXiv:2405.21075*, 2024. 2, 5

[2] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2, 5

[3] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024. 2, 5

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

[5] 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.

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

[7] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. *arXiv preprint arXiv:2404.00308*, 2024. 2

[8] 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. 2, 5, 6

[9] 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. 2, 5, 6

[10] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 2

[11] 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. 2

[12] 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. 2, 7

[13] 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. 2

[14] 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. 2, 7

[15] 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. 2

[16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. 3

[17] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107, 2023. 3

[18] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. *arXiv preprint arXiv:2406.12846*, 2024. 3, 4

[19] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. *arXiv preprint arXiv:2407.12679*, 2024. 3, 4, 5, 6

[20] Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension, 2024. 3, 5

[21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 5

[22] OpenAI. Gpt-4 technical report, 2023. 5, 6

[23] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, May 2024. 5, 6, 7

[24] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 5, 6, 7

[25] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 5, 6

[26] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7

[27] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5, 6

[28] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie

Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 5, 6

[29] 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. *arXiv preprint arXiv:2406.16852*, 2024. 6

[30] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 6

[31] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 6

[32] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024. 6

[33] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 5, 6

[34] Junho Kim, Hyunjun Kim, Hosu Lee, and Yong Man Ro. Salova: Segment-augmented long video assistant for targeted retrieval and routing in long-form video analysis, 2024. 5, 6

[35] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 5

[36] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. 6

[37] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering, 2019. 6

[38] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 7

[39] 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*, pages 14313–14323, 2024. 7

[40] 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. 7

[41] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogvlm2: Visual language models for image and video understanding, 2024. 7

[42] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next-video: Advancing video understanding with llava-next. April 2024. Accessed: 2024-05-15. 7

[43] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels, 2022. 7