APVR: Hour-Level Long Video Understanding with Adaptive Pivot Visual Information Retrieval

Hong Gaoe*[1,2], Yiming Bao*[2], Xuezhen Tu[2], Bin Zhong[2], Linan Yue[1], Min-Ling Zhang[1†]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]ZTE Corporation, Nanjing 210012, China
230239517@seu.edu.cn, dr.ymbao@gmail.com, zhangml@seu.edu.cn

Abstract

Current multimodal large language models (MLLMs) struggle with hour-level video understanding, facing significant challenges not only in modeling the substantial information volume of long videos but also in overcoming the memory wall and resource constraints during both training and inference. Although recent training-free approaches have alleviated resource demands by compressing visual features, their reliance on incomplete visual information limits the performance potential. To address these limitations, we propose **A**daptive **P**ivot **V**isual information **R**etrieval (**APVR**), a training-free framework that hierarchically retrieves and retains sufficient and important visual information. It breakthroughs the memory wall limitation via two complementary components: Pivot Frame Retrieval employs query expansion and iterative spatio-semantic confidence scoring to identify relevant video frames, and Pivot Token Retrieval performs query-aware attention-driven token selection within up to 1024 pivot frames. This dual granularity approach enables the processing of hour-long videos while maintaining semantic fidelity. Experimental validations on three different baseline MLLMs demonstrate significant performance improvements up to 9.5%, 4.6% and 9.7% on LongVideoBench, VideoMME and MLVU, respectively. APVR achieves state-of-the-art results for both training-free and training-based approaches.

Introduction

The proliferation of long-form video content necessitates robust understanding capabilities that extend beyond the current limitations of multimodal large language models (MLLMs). While existing video-based models (Lin et al. 2023; Zhang et al. 2025; Bai et al. 2025a; Team et al. 2025) demonstrate proficiency on short video sequences, their scalability for hour-level video remains challenging. Specifically, they encounter not only the memory wall but also the inherent dilution of semantic information across extensive temporal spans.

The core challenge for long video understanding with Video MLLMs lies in efficiently modeling and extracting sparse but critical semantic information across both temporal and spatial dimensions. Existing research generally falls into three paradigms: (1) uniform frame sampling, (2) sparse key-frame retrieval, and (3) sufficient frames processing while requiring computational tricks.

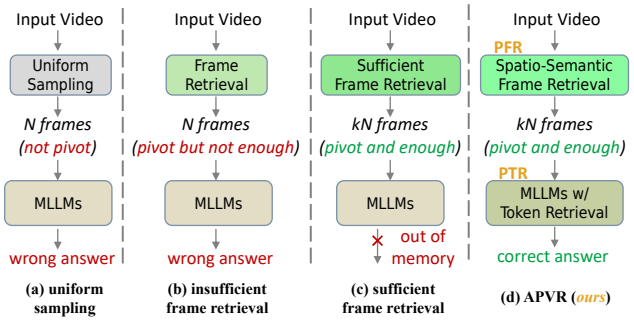

Figure 1: Given a long video, uniform sampling (a) and insufficient frame retrieval (b) yield incorrect answer, while naively increasing pivot frames (c) raises OOM. Our APVR (d) explores joint frame-token co-retrieval, concentrating computational resources on the most relevant information.

The first paradigm, uniform sampling, understands the entire video by evenly selecting frames across the sequence (Bai et al. 2025a; Zhang et al. 2025). As illustrated in Fig. 1(a), this strategy samples frames that are not relevant and pivot, leading to wrong answer generated by MLLMs. The second line of work attempts to address inefficiency by retrieving only N sparse, query-relevant key frames from long videos (Ataallah et al. 2024; Tang et al. 2025a; Guo et al. 2025; Ye et al. 2025; Cheng et al. 2025; Luo et al. 2025). This key-frame selection greatly reduces the computational load; however, as shown in Fig. 1(b), it neglects the temporal-semantic relationships between events and often fails in scenarios that require comprehensive temporal reasoning or long-range logical understanding. To improve temporal coverage, a third paradigm retrieves a denser set of kN pivot frames (Fig. 1(c)) to capture relevant information. Meanwhile, it quickly encounters the memory wall when fed to MLLMs. Recent solutions leverage training-time trick, such as multimodal sequence parallelism (Fu et al. 2025; Liu et al. 2024c) or feature compression (Shu et al. 2024; Liu et al. 2025a), to handle longer sequences. However, these

*These authors contributed equally.
†Corresponding Author.

methods require resource-intensive multi-stage re-training and are tightly coupled to specific MLLM architectures, limiting their adaptability in a rapidly evolving models.

The inherent trade-off between temporal coverage and computational feasibility has hindered existing methods from achieving comprehensive video understanding. To address this core challenge, we propose **A**daptive **P**ivot **V**isual information **R**etrieval (**APVR**) framework, a training-free approach that integrates spatio-semantic frame retrieval with token retrieval and compression, as shown in Fig. 1(d). This two-stage retrieval mechanism concentrates computational resources on the most relevant visual information for the input query, thus enabling efficient, scalable and adaptable understanding of hour-long videos without retraining or encountering memory limitations.

Specifically, at the frame level, our *Pivot Frame Retrieval (PFR)* component expands the original query into four types of semantic information: *objects, descriptions, relations,* and *semantics*. This expansion enables more comprehensive frame scoring from spatial and semantic perspectives using complementary visual models: CLIP for semantic similarity and Grounding-DINO for object detection and spatial reasoning. The framework incorporates temporal diffusion mechanisms to maintain temporal coherence and employs adaptive resampling strategies to efficiently and iteratively refine frame scoring and final selection. At the token level, our *Pivot Token Retrieval (PTR)* component extends the concept of semantic importance to fine-grained visual representations. This component leverages query-aware multi-layer attention scoring to identify the most relevant visual tokens to select and the irrelevant visual tokens to compress within selected frames, employing dynamic chunk-wise token selection and head-wise soft voting mechanisms to maintain both computational efficiency and semantic accuracy.

The main contributions of this work are as follows:

• We propose APVR, a novel training-free framework that addresses scalability challenges in long video understanding through dual-granularity information retrieval. Our approach combines frame-level pivot retrieval with token-level adaptive selection to breakthrough the memory wall.

• We develop a mechanism that leverages spatio-semantic confidence scoring and query-aware attention scoring to preserve temporal structure and semantic details, ensuring accurate understanding of complex video narratives while maintaining computational efficiency.

• We demonstrate that intelligent dual-granularity retrieval provides a sustainable alternative to parameter scaling for long video understanding. Our training-free design ensures plug-and-play integration with existing MLLM architectures while preserving foundational capabilities.

Related Work

Video MLLMs Recent years have witnessed significant advances in multimodal large language models (MLLMs) for visual perception and understanding. As a critical modality in visual data analysis, video has also been integrated into MLLM frameworks. Notable efforts include Qwen2.5-VL (Bai et al. 2025b), InternVideo2.5 (Wang et al. 2025b), VideoLlama3 (Zhang et al. 2025), and VideoChat-Flash (Li et al. 2024b), which explore video understanding through a unified architecture consisting of three key components: (a) a video encoder for spatio-temporal feature extraction, (b) a projector for aligning visual features with linguistic embeddings, and (c) a video-aware LLM for multimodal reasoning. Due to the inherent complexity of video content (e.g., multi-object interactions, long-range temporal dependencies), existing MLLMs often fail to extract semantically salient information for video-centric downstream tasks requiring precise spatiotemporal reasoning, such as video question answering and spatiotemporal video grounding. Some works(Bai et al. 2025b; Wei et al. 2025) exploit to aggregate timestamps in the RoPE(Su et al. 2024). However, these methods exhibit significant performance degradation when extrapolating to videos exceeding their pre-training duration. Moreover, current architectures also exhibit limitations of inefficient cross-modal alignment between visual information and linguistic queries, particularly in long-form videos exceeding 5 minutes (Li et al. 2024a).

Long Video Understanding Recently advances in long video understanding can be broadly categorized into training-based and training-free paradigms. In the training-based paradigm, MLLMs are directly trained using large-scale long video datasets. The whole training process is usually divided into several stages to stably improve different components of MLLM, as well as to improve the ability to model different types of input visual data (Li et al. 2024b; Zhang et al. 2025). In order to process extra-long videos, current methods explore two different strategies: sequence parallelism (Chen et al. 2024; Zhang et al. 2024b; Shen et al. 2025) or visual feature compression (Song et al. 2024; Shen et al. 2024; Liu et al. 2025a). For instance, LongVILA (Chen et al. 2024) adopts a five-stage training pipeline with a novel multi-modal sequence parallelism to efficiently learn from long videos. Video-XL-Pro (Liu et al. 2025a) builds a learnable module to generate compact and comprehensive tokens for long video understanding. In the training-free paradigm, long video understanding task is organized as an agent-like process (Liu et al. 2025b; Wang et al. 2024c), where MLLMs serve as reasoning engine within the system. These training-free methods mainly focus on retrieving the most important information in video frames (Tang et al. 2025b; Park et al. 2024; Ye et al. 2025; Guo et al. 2025; Cheng et al. 2025) or visual tokens (Wang et al. 2025a; Luo et al. 2025). For example, AKS (Tang et al. 2025b) adopts a plug-and-play module for keyframe selection and information prefiltering for video-based MLLMs. QuoTA (Luo et al. 2025) adopts an ante-hoc training-free module for query-aware visual token assignment and important information retrieval. Despite these advancements, few methods simultaneously explore frame and token optimization, which is addressed in this work through adaptive pivot information retrieval.

Method

Overview

The framework of the proposed APVR is illustrated in Fig. 2, which is a training-free system. Given an input video

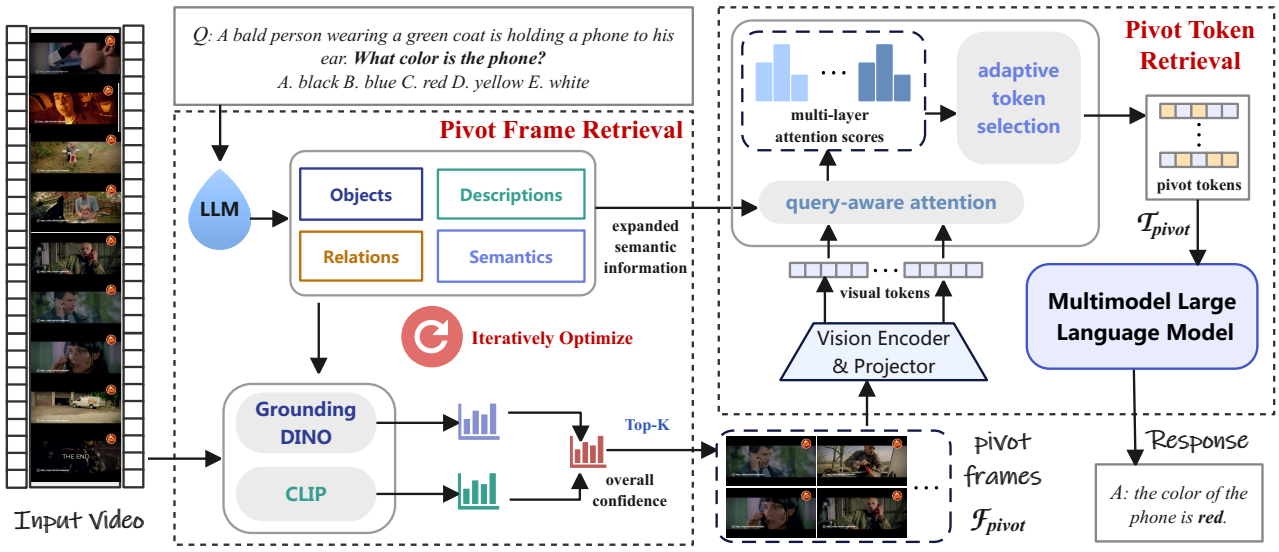

Figure 2: The overall framework of our proposed APVR. We integrate two plug-and-play components, Pivot Frame Retrieval and Pivot Token Retrieval, into MLLMs to improve the performance of long video understanding. With frame-level and token-level adaptive selection, APVR can accurately understand complex videos with computational efficiency.

$\mathcal{F} = \{f_t\}_{t=1}^{N}$ with N frames sampled under specific fps and query Q, the system aims to retrieve the most important visual information for accurate response. To achieve this, we build two main components in APVR: *Pivot Frame Retrieval* (PFR) and *Pivot Token Retrieval* (PTR). In PFR, we search pivot frames $\mathcal{F}_{pivot}$ with high confidence scores using expanded queries and basic visual models. In PTR, we explore an attention-driven token selection in MLLMs to retrieve pivot tokens $\mathcal{T}_{pivot}$ that are most relevant with expanded queries.

Pivot Frame Retrieval

The progress of PFR is developed in an iterative and efficient manner, as elaborated in Alg. 1.

Semantic Information Expansion First, We expand the given query Q to four types of semantic information as follows:

Objects: The key objects Obj_k or cue objects Obj_c that are detectable by visual models such as Grounding-DINO (e.g., *"bold person", "phone", "blue background"*). All objects are aggregated as $Obj = \{Obj_k, Obj_c\}$.

Descriptions: The entities or hypernym concepts of objects based on knowledge graph augmentation, denoted as Des. (e.g. *"pearl headband: decorative accessory worn on head; jewelry"*)

Relations: The triplet relationship $Rel \subseteq Obj \times \mathcal{R} \times Obj$ to enhance the confidence of frames with logical links between objects. In each triplet (o_i, r, o_j), the type of relation $r \in \mathcal{R}$ is exactly one of the following: (1) **spatial** which means that o_j and o_j appear in the same frame, such as *"a bold man holding a red phone"*. (2) **time** which means that o_j and o_j appear in different frames in orders such as *"A woman walks through, after a cat follows."* (3) **attribute** which means that o_i appears with an attribute describable by o_j such as *"A woman in a white skirt"*. (4) **causal** which means that o_j and o_j appear in orders with a cause-effect manner such as *"A woman opens the door, the cat is scared and runs out."*

Semantics: The semantics information of query and options based on knowledge graph, denoted as Sem (e.g., *'leash often appears with dog"*).

The prompt template for LLM-based query expansion is provided in Appendix.

Spatio-Semantic Confidence Scoring Given a video $\mathcal{F}$, and query Q with expanded semantic information $\{Obj, Des, Rel, Sem\}$, instead of exhaustively processing all N frames, we uniformly sample few frames with initial stride ∇ in the first iteration and adaptively resample frames for the subsequent iteration based on the online updated frame confidence. To ensure progressively finer sampling, we gradually decrease the stride by $\nabla^p = \max(1, \frac{\nabla}{p})$. The frames sampled in iteration p are denoted as $\mathcal{F}^p_{samp}$.

Our algorithm exploits two basic visual models, i.e., Grounding-DINO and CLIP, to comprehensively score $\mathcal{F}^p_{samp}$ as follows:

CLIP-based Similarity Scoring: We leverage the CLIP model to compute cross-modal similarity between text and visual features as the semantic confidence score based on expanded query and the visual feature by $s_t^{CLIP} = \Pi_{CLIP}(f_t, Q, Des, Sem)$. The details of the function Π_{CLIP} are provided in Appendix.

Grounding-DINO Object Detection: We leverage the Grounding-DINO model to compute the spatial confidence score based on the query-aware objects long with their relations by $s_t^{GD} = \Pi_{GD}(f_t, Obj, Rel)$. The details of the

Algorithm 1: Pivot Frame Retrieval (PFR)

Require: video frames $\mathcal{F}$, query Q, iterations P, stride ∇
Ensure: Pivot frames $\mathcal{F}_{pivot} = \{f_t\}_{t=1}^{K}$

Initialize:
$\mathcal{S} \leftarrow \mathbf{0}^N$ ▷ Confidence scores
$\mathcal{V} \leftarrow \emptyset$ ▷ Visited frames set
$\{Obj, Des, Rel, Sem\} \leftarrow \text{LLM}(Q)$ ▷ Query Expansion
for $p = 1$ **to** P **do**
  $\nabla_p \leftarrow \max\left(1, \frac{\nabla}{p}\right)$
  $\mathcal{F}^p_{samp} \leftarrow \text{AdaptiveResample}(\mathcal{F}, \mathcal{S}, \mathcal{V}, \nabla_p)$
  for $f_t \in \mathcal{F}^p_{samp}$ **do**
    $s_t^{CLIP} \leftarrow \Pi_{CLIP}(f_t, Q, Des, Sem)$
    $s_t^{GD} \leftarrow \Pi_{GD}(f_t, Obj, Rel)$
    $\mathcal{S}_t \leftarrow (1-\lambda)s_t^{CLIP} + \lambda s_t^{GD}$
    $\mathcal{S}_{[t-w,t+w]} \leftarrow \max\left(\mathcal{S}_{[t-w,t+w]}, \frac{\mathcal{S}_t}{1+|i-t|}\right)$
  end for
  $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{F}^p_{samp}$
end for
return $\mathcal{F}_{pivot} \leftarrow \text{topk}(\mathcal{S} \odot \mathcal{F}, K)$

function Π_{GD} are provided in Appendix.

The final confidence score of this iteration is computed as a weighted sum of the CLIP and Grounding-DINO scores: $\mathcal{S}_t = (1-\lambda) \cdot s_t^{CLIP} + \lambda \cdot s_t^{GD}$.

The frames scored in the current iteration are put into the visited frame set $\mathcal{V}$.

Temporal Diffusion: To account for the temporal continuity of video content and improve frame resampling, we introduce a temporal score diffusion mechanism that propagates confidence scores across neighboring frames as:

$$\mathcal{S}_i = \max(\mathcal{S}_i, \frac{\mathcal{S}_t}{1+|i-t|}), i \in [t-w, t+w]. \quad (1)$$

This diffusion process helps capture temporally coherent segments relevant to the query.

Adaptive Resampling Based on the updated $\mathcal{S}$, we further design an adaptive and hybrid candidate sampling strategy for subsequent iterations that combines two types of set:

High-confidence Set: We select frames with higher scores from all unvisited frame set $\mathcal{U} = \{f_t \mid f_t \notin \mathcal{V}\}$ based on the distribution as:

$$\mathcal{H}_s = \text{TopK}(\mathcal{S} \odot \mathcal{U}, \frac{N}{2\nabla_p}). \quad (2)$$

Uncertainty Set: We calculate Shannon entropy based on the score distribution as:

$$\mathcal{E}_\gamma(s_i) = -\sum_{j=i-\gamma}^{i+\gamma} e_j \cdot \log(e_j), e_j = \frac{s_j}{\sum_{k=j-\gamma}^{j+\gamma} s_k}, \quad (3)$$

then the frames with high entropy are identified by:

$$\mathcal{H}_e = \{f_t \mid \mathcal{E}_\gamma(s_t) > \mu + 0.5\sigma\}, \quad (4)$$

where μ and σ are the mean and standard deviation of the entropy in $\mathcal{C}$.

The union candidate set for this iteration sampling is $\mathcal{C} = \mathcal{H}_s \cup \mathcal{H}_e$. Assume that there are $N_\mathcal{C}$ frames in $\mathcal{C}$. The frames $\mathcal{F}^p_{samp}$ are sampled in the way as:

$$\mathcal{F}^p_{samp} = \mathcal{C}_{\text{multi}} \cup \mathcal{C}_{\text{rand}}, \quad (5)$$

$$\mathcal{C}_{\text{multi}} \sim \text{Multinomial}(S \odot \mathcal{C}, (1-\alpha)N_\mathcal{C}), \quad (6)$$

$$\mathcal{C}_{\text{rand}} \sim \text{Unif}(\text{Perm}(S \odot \mathcal{C}, \alpha N_\mathcal{C})), \quad (7)$$

where α balances the multinomial score-based importance sampling $\mathcal{C}_{\text{multi}}$ and the diverse random sampling $\mathcal{C}_{\text{rand}}$ to capture the full semantic content of the video and also to avoid falling into local optimal.

After P iterations, the K pivot frames are retrieved by $\mathcal{F}_{pivot} = \text{TopK}(\mathcal{S} \odot \mathcal{F}, K)$ and fed into the next module.

Pivot Token Retrieval

Pivot Token Retrieval extends the concept of semantic importance from frame-level to token-level granularity. In this module, we adaptively retrieve the pivot visual token within MLLM to maintain computational efficiency while preserving critical semantic information.

First, the retrieved pivot frames $\mathcal{F}_{pivot}$ are processed by a Vision Encoder $\mathcal{G}(\cdot)$ and a Projector $\mathcal{P}(\cdot)$ to generate dense visual token representations: $\mathcal{T}_{vis} = \mathcal{P}(\mathcal{G}(\mathcal{F}_{pivot})) = \{\mathcal{T}_t\}_{t=1}^{K}$. Our objective is to retrieve the pivot token $\mathcal{T}_{pivot}$. To this end, we introduce two components: query-aware multi-layer attention scoring and adaptive token selection.

Query-aware Multi-Layer Attention Scoring We leverage the expanded semantic information from the LLM-augmented query to guide token selection across multiple transformer layers. First, the attention scores for layer l are computed as:

$$A^l_{cross} = \text{softmax}(\frac{q^l_{text} \cdot (k^l_{vis})^T}{\sqrt{d}}), \quad (8)$$

where $A^l_{cross} \in R^{h \times d \times d_q \times d_v}$, q^l_{text} denotes the query states derived from $\{Q, Des, Sem\}$ and k^l_{vis} denotes the key state derived from the visual tokens $\mathcal{T}_{vis}$. Then the token-level attention scores of each layer are aggregated at the query dimension by:

$$a^l = \sum_{d_q} A^l_{cross}. \quad (9)$$

This formulation captures how much attention each key token receives across all query positions, providing a comprehensive measure of its importance for the current layer. The superscript l will be omitted for concision.

Adaptive Token Selection The simplest token selection strategy is to directly retain the Top-K tokens while evicting others. However, this approach ignores the continuity of visual tokens and may reduce accuracy. Another factor that matters is that attention scores across heads may exist disparities, which further result in independence between heads. Thus, considering the continuity of visual tokens as well as the disparities across heads, we introduce an adaptive token selection strategy consisting of two components: dynamic chunk-wise selection and head-wise soft voting.

Methods	Params	Training -free	LVB val	VideoMME(w/o sub.) Long	VideoMME(w/o sub.) Overall	MLVU dev
Proprietary Methods						
GPT4-V(OpenAI 2023)	-	No	59.1	53.5	59.9	-
GPT4-o(Hurst et al. 2024)	-	No	66.7	65.3	71.9	64.6
Gemini-1.5-Pro(Team et al. 2024)	-	No	64.0	67.4	75.0	-
Open-Sourse Methods						
VideoLLaVA(Lin et al. 2023)	7B	No	58.2	36.2	39.9	67.0
VITA-1.5(Fu et al. 2025)	7B	No	-	47.1	56.1	-
mPLUG-Owl3(Ye et al. 2024)	7B	No	52.1	50.1	59.3	63.7
Oryx-1.5(Liu et al. 2024b)	7B	No	56.3	-	58.3	67.5
QuoTA(Luo et al. 2025)	7B	Yes	59.0	55.7	65.9	71.9
Video-XL(Shu et al. 2024)	7B	No	49.5	-	55.5	64.9
ViLaMP(Cheng et al. 2025)	7B	No	57.8	-	67.5	72.6
AdaReTaKe(Wang et al. 2025a)	7B	Yes	62.6	58.3	67.7	75.0
AKS(Tang et al. 2025a)	7B	Yes	62.7	-	65.3	-
NVILA(Liu et al. 2024c)	8B	No	57.7	54.8	64.2	70.1
ByteVideoLLM(Wang et al. 2024a)	14B	No	-	56.4	64.6	65.0
Qwen2-VL(Wang et al. 2024b)	7B	No	55.6	53.8	63.3	66.9
Qwen2-VL w/ APVR(*ours*)	7B	Yes	60.9(+9.5%)	55.1(+2.4%)	65.2(+3.0%)	73.4(+9.7%)
Qwen2.5-VL(Bai et al. 2025a)	7B	No	59.5	55.6	65.4	70.2
Qwen2.5-VL w/ APVR(*ours*)	7B	Yes	**64.9**(+9.1%)	**59.1**(+6.3%)	**68.4**(+4.6%)	<u>76.1</u>(+8.4%)
VideoLLaMA3(Zhang et al. 2025)	7B	No	59.8	54.9	66.2	73.0
VideoLLaMA3 w/ APVR(*ours*)	7B	Yes	<u>63.5</u>(+6.2%)	<u>58.7</u>(+6.9%)	<u>68.1</u>(+2.9%)	**77.2**(+5.5%)

Table 1: Video understanding accuracy(%) on LongVideoBench (LVB), VideoMME and MLVU. We categorize the methods based on whether they are training-free. Bold denotes the best result while underline the second best.

Dynamic Chunk-wise Selection: The long video understanding reaches the memory wall when prefilling the visual token. Thus, we perform token selection for the KV cache to reduce redundancy. Given a key or value cache to select, we first divide it into W chunks at the token dimension. Then, the base selection ratio of the chunk w is computed as:

$$\eta_w = \frac{\sum_w a}{\max(\{\sum_i a \mid i = 1, ..., W\})}. \tag{10}$$

The dynamic ratio of each chunk is calculated by:

$$\rho_w = \min(1.0, \sqrt{\frac{|\{j : a_j > 0.01 \cdot \max(a)\}|}{L_w}}), \tag{11}$$

where L_w is the chunk length. The final selection ratio is determined by: $\gamma_w = \rho_w \cdot \eta_w$.

Head-wise Soft Voting: Based on the ratio γ_w, we can select the pivot token in each chunk by:

$$\mathcal{Z}_w = \mathrm{TopK}(a_w \odot \mathcal{T}_w, \gamma_w L_w). \tag{12}$$

This results may appear to be significant disparities in different attention heads. To figure out this, we introduce head-wise voting for final token selection:

$$\mathcal{Z}_w = \mathrm{TopK}((\sum_{j=1}^{h} \mathrm{softmax}(a_{w,j})) \odot \mathcal{T}_w, \gamma_w L_w), \tag{13}$$

Method	Long (10-60m)	Avg
APVR (Qwen2-VL-7B)	51.2	60.9
w/o exp. sem. info.	50.4	59.1
APVR (Qwen2.5-VL-7B)	58.5	64.5
w/o exp. sem. info.	56.6	63.2
APVR (VideoLLaMA3-7B)	56.9	63.5
w/o exp. sem. info.	55.5	62.3

Table 2: Ablations of expanded semantic information.

where the softmax function calculates the per-head scores and gives a soft voting via normalization and sum. Finally, we update the KV cache as the selected pivot token in each layer:

$$\mathcal{T}_{pivot} = \mathrm{Concat}(\{\mathcal{Z}_w\}_{w=1}^{W}). \tag{14}$$

Finally, the selected pivot tokens will be fed into the MLLM to generate the final response.

Experiments

Experimental Setup

Datasets We evaluate the performance of APVR in three popular benchmarks that contains hour-level videos:

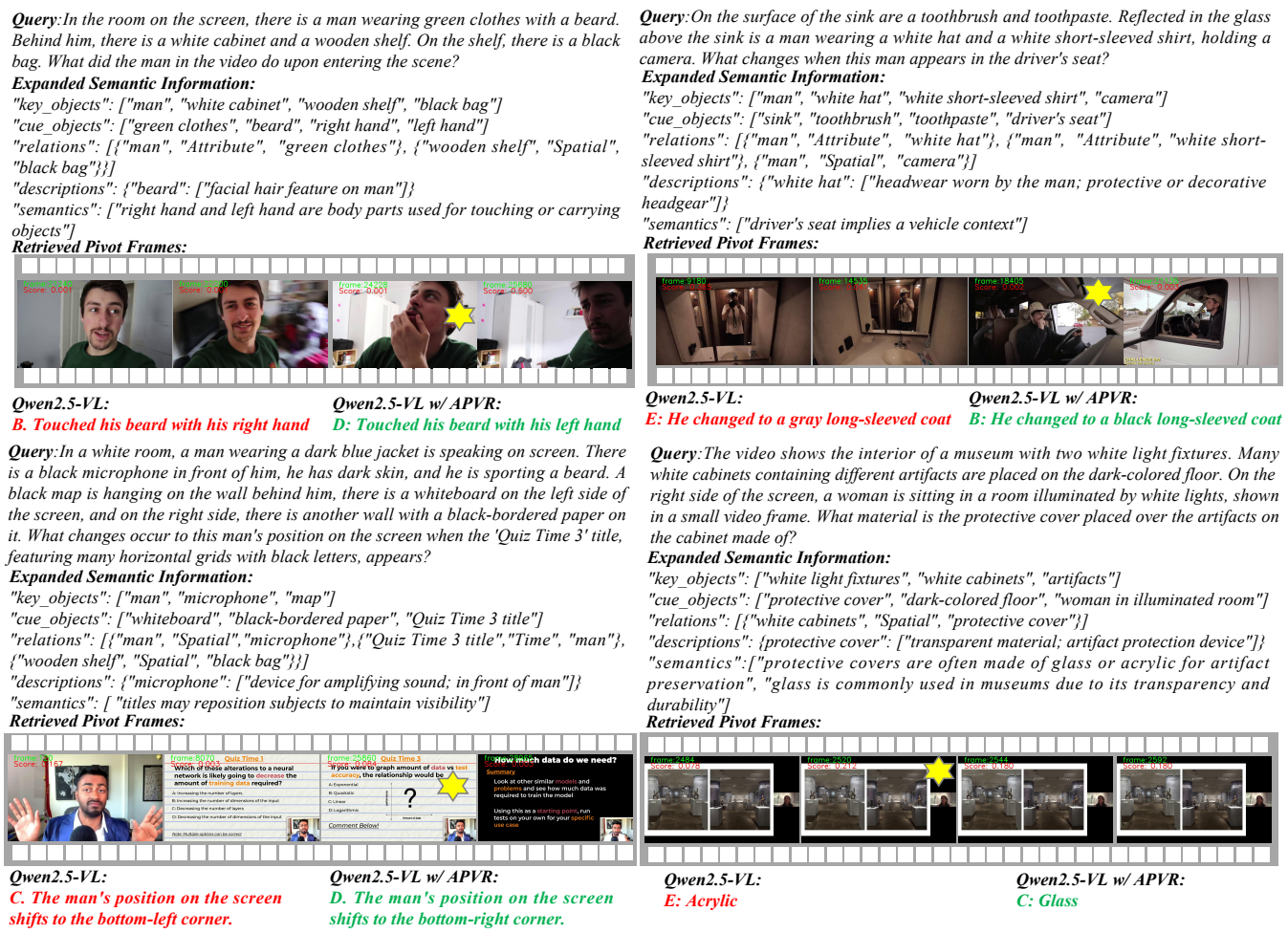

Figure 3: Qualitative Comparison of APVR with the baseline MLLM. The number and score of the pivot frame is drawn in the left-top with green and red text, respectively. Yellow stars indicate the key frame for correct response.

LongVideoBench(Wu et al. 2024): A benchmark consisting of 17 categories of referred reasoning questions. **VideoMME**(Fu et al. 2024): A multi-modal evaluation benchmark focused on fine-grained video understanding tasks. **MLVU**(Zhou et al. 2025): A large-scale long video benchmark with a large wide of 9 distinct tasks. To highlight the importance of the pivot information retrieval for visual understanding, we do not use subtitles for all benchmarks.

Implementation Details We integrate APVR with three baseline MLLMs, namely, Qwen2-VL-7B (Wang et al. 2024b) , Qwen2.5-VL-7B (Bai et al. 2025a) and VideoLLaMA3 (Zhang et al. 2025). The MLLMs process a prompt involving the question, retrieved pivot video frames, and the multi-choice answer options to generate responses. We first densely extracted the raw frames from video with $fps = 2$. Then, the sampled frames are iteratively scored using two basic visual models with $\lambda = 0.5$: CLIP (Radford et al. 2021) with ViT-B-16 (Dosovitskiy et al. 2020) backbone, and Grounding-DINO (Liu et al. 2024a) with Swin-T (Liu et al. 2021) backbone, based on the expanded semantic information generated by LLM. All evaluations are conducted on 8 NVIDIA A800 GPUs with 80GB memory, using LMMs-Eval (Zhang et al. 2024a) framework. Under the above setup, our APVR enables processing of up to $K = 1024$ frames, compared to the baseline 7B model's capacity of 256 frames under optimal MLLM configurations. Our APVR can understand an hour-level long video and generate correct answers in up to 2 minutes.

Comparison to the State-of-the-Art

We first compare the accuracy of video question answering between APVR with recent training-based or training-free methods. All results are summarized in Table 1. Upon three different baseline MLLMs, APVR improves overall accuracy on all three widely used benchmarks. Notably, APVR based on Qwen2.5-VL reaches 64.9% accuracy in LongVideoBench, which is higher than the powerful proprietary models such as GPT4-V and Gemini-1.5-Pro. The overall performance of APVR also surpasses other training-

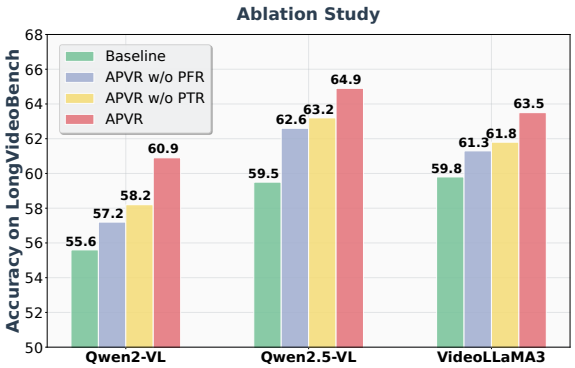

Figure 4: The result of ablation study on LongVideoBench for the two designed components: PFR and PTR.

K	Long	Average	λ	Long	Average
1024	58.3	64.9	0.1	57.4	64.2
256	56.4	63.1	0.2	57.8	64.5
128	53.0	61.1	0.5	58.3	64.9
64	49.8	58.6	0.8	58.0	64.8
32	45.4	54.3	0.9	58.9	64.5

Table 3: Ablations of selected frames K and weight λ.

based and training-free methods with larger parameter scales such as 8B and 14B. The complete verification on the benchmarks confirms the capability of APVR to preserve critical spatio-temporal information in long video sequences.

Analysis on Information Retrieval

Fig. 3 presents qualitative results, comparing our APVR framework with the baseline Qwen2.5-VL across four challenging samples. The results clearly demonstrate the advantages of APVR in both the semantic expansion and the subsequent retrieval process.

Firstly, the pivot frames selected by APVR are highly correlated with both the query and the expanded semantic information. APVR distills the most informative visual information within sufficient pivot frames. As a result, APVR empowers MLLM to efficiently comprehend long-form video content, providing accurate and detailed answers even in scenarios where baselines fail or generate hallucinations. Secondly, the semantic information expansion module in APVR significantly enriches the original query, providing the downstream retrieval process with comprehensive object, attribute, and relational cues that are otherwise missing when using only the raw query. As shown in Table 2, on LongVideoBench the performances decrease without expanded semantic information no matter the baseline MLLM.

Ablation Studies on PFR and PTR

To systematically validate the effectiveness of the proposed dual-granularity retrieval strategy, we conduct ablation studies for the two core components of APVR: Pivot Frame Retrieval (PFR) and Pivot Token Retrieval (PTR), on LongVideoBench. The results, shown in Fig. 4, consistently demonstrate that removing either PFR or PTR leads to a notable decline in accuracy across all baseline MLLMs. Without PFR, the system loses its ability to precisely localize query-relevant frames, resulting in performance degradation due to the inclusion of irrelevant content (Fig. 1(a)). Without PTR, the model is less able to compress visual tokens and focus on the most informative visual details, leading to processing maximum to only 256 frames and reduced answer quality (Fig. 1(b) and (c)). These findings strongly support the necessity and rationality of our dual-granularity pivot information retrieval design.

P	∇	Long(10-60m)	Medium(1-10m)	Average
3	2	56.4	62.1	62.9
3	4	58.3	65.8	64.9
3	8	55.1	64.6	63.1
2	4	56.0	66.0	63.9
4	4	58.0	65.8	64.7
5	4	58.5	64.5	64.5

Table 4: Ablations of iteration number P and stride ∇.

Ablation Studies on Hyper-Parameters

We further ablate the key hyper-parameters that influence the performance of APVR on LongVideoBench. First, we examine the impact of the number of retrieved pivot frames K and the frame scoring weights λ. As shown in Table 3, increasing K generally improves performance, especially on long videos, by reducing the risk of missing relevant segments. For λ, we observe that optimal results are achieved by striking a balance between spatial and semantic scoring, confirming the complementarity of these two perspectives.

Second, we analyze the effect of the number of frame retrieval iterations p and the initial sampling stride ∇. As shown in Table 4, the best performance is observed when $P = 3$ and $\nabla = 4$, indicating that a moderate number of adaptive iterations and a balanced initial sampling density allow the system to progressively refine frame selection without causing overfitting or excessive computational cost.

Conclusion

We propose APVR, a training-free framework that overcomes the memory wall in hour-level video understanding via adaptive, hierarchical visual information retrieval. By integrating Pivot Frame Retrieval and Pivot Token Retrieval, APVR enables efficient and accurate processing of long videos without sacrificing semantic detail. Extensive experiments on standard benchmarks demonstrate that APVR delivers state-of-the-art results among training-free and training-based methods. Its plug-and-play design allows seamless integration with existing MLLM systems, eliminating the need for costly retraining. Our results establish retrieval-based optimization as a practical and scalable solution for long-form video comprehension.

References

Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhuge, M.; Ding, J.; Zhu, D.; Schmidhuber, J.; and Elhoseiny, M. 2024. Goldfish: Vision-language understanding of arbitrarily long videos. In *European Conference on Computer Vision*, 251–267. Springer.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, Y.; Xue, F.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; et al. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Cheng, C.; Guan, J.; Wu, W.; and Yan, R. 2025. Scaling Video-Language Models to 10K Frames via Hierarchical Differential Distillation. *arXiv preprint arXiv:2504.02438*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Fu, C.; Lin, H.; Wang, X.; Zhang, Y.-F.; Shen, Y.; Liu, X.; Cao, H.; Long, Z.; Gao, H.; Li, K.; et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Guo, W.; Chen, Z.; Wang, S.; He, J.; Xu, Y.; Ye, J.; Sun, Y.; and Xiong, H. 2025. Logic-in-Frames: Dynamic Keyframe Search via Visual Semantic-Logical Verification for Long Video Understanding. *arXiv preprint arXiv:2503.13139*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024a. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 237–255. Springer.

Li, X.; Wang, Y.; Yu, J.; Zeng, X.; Zhu, Y.; Huang, H.; Gao, J.; Li, K.; He, Y.; Wang, C.; et al. 2024b. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Liu, X.; Shu, Y.; Liu, Z.; Li, A.; Tian, Y.; and Zhao, B. 2025a. Video-XL-Pro: Reconstructive Token Compression for Extremely Long Video Understanding. *arXiv preprint arXiv:2503.18478*.

Liu, Y.; Lin, K. Q.; Chen, C. W.; and Shou, M. Z. 2025b. VideoMind: A Chain-of-LoRA Agent for Long Video Reasoning. *arXiv preprint arXiv:2503.13444*.

Liu, Z.; Dong, Y.; Liu, Z.; Hu, W.; Lu, J.; and Rao, Y. 2024b. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. 2024c. NVILA: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Luo, Y.; Chen, W.; Zheng, X.; Huang, W.; Yin, S.; Lin, H.; Fu, C.; Huang, J.; Ji, J.; Luo, J.; et al. 2025. QuoTA: Query-oriented Token Assignment via CoT Query Decouple for Long Video Comprehension. *arXiv preprint arXiv:2503.08689*.

OpenAI. 2023. GPT-4 with vision. https://platform.openai.com/docs/guides/vision. Accessed: 2024-07-16.

Park, J.; Ranasinghe, K.; Kahatapitiya, K.; Ryu, W.; Kim, D.; and Ryoo, M. S. 2024. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shen, X.; Xiong, Y.; Zhao, C.; Wu, L.; Chen, J.; Zhu, C.; Liu, Z.; Xiao, F.; Varadarajan, B.; Bordes, F.; et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Shen, Y.; Fu, C.; Dong, S.; Wang, X.; Chen, P.; Zhang, M.; Cao, H.; Li, K.; Zheng, X.; Zhang, Y.; et al. 2025. Long-VITA: Scaling Large Multi-modal Models to 1 Million Tokens with Leading Short-Context Accuray. *arXiv preprint arXiv:2502.05177*.

Shu, Y.; Liu, Z.; Zhang, P.; Qin, M.; Zhou, J.; Liang, Z.; Huang, T.; and Zhao, B. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tang, X.; Qiu, J.; Xie, L.; Tian, Y.; Jiao, J.; and Ye, Q. 2025a. Adaptive Keyframe Sampling for Long Video Understanding. *arXiv preprint arXiv:2502.21271*.

Tang, X.; Qiu, J.; Xie, L.; Tian, Y.; Jiao, J.; and Ye, Q. 2025b. Adaptive Keyframe Sampling for Long Video Understanding. *arXiv preprint arXiv:2502.21271*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team, K. K.; Yang, B.; Wen, B.; Liu, C.; Chu, C.; Song, C.; Rao, C.; Yi, C.; Li, D.; Zang, D.; et al. 2025. Kwai Keye-VL Technical Report. *arXiv preprint arXiv:2507.01949*.

Wang, H.; Nie, Y.; Ye, Y.; GuanYu, D.; Wang, Y.; Li, S.; Yu, H.; Lu, J.; and Huang, C. 2024a. Dynamic-VLM: Simple Dynamic Visual Token Compression for VideoLLM. *arXiv preprint arXiv:2412.09530*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Si, Q.; Wu, J.; Zhu, S.; Cao, L.; and Nie, L. 2025a. AdaRETAKE: Adaptive Redundancy Reduction to Perceive Longer for Video-language Understanding. *arXiv preprint arXiv:2503.12559*.

Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2024c. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 58–76. Springer.

Wang, Y.; Li, X.; Yan, Z.; He, Y.; Yu, J.; Zeng, X.; Wang, C.; Ma, C.; Huang, H.; Gao, J.; et al. 2025b. InternVideo2. 5: Empowering Video MLLMs with Long and Rich Context Modeling. *arXiv preprint arXiv:2501.12386*.

Wei, X.; Liu, X.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Tong, J.; Duan, H.; Guo, Q.; Wang, J.; et al. 2025. VideoRoPE: What Makes for Good Video Rotary Position Embedding? *arXiv preprint arXiv:2502.05173*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Ye, J.; Wang, Z.; Sun, H.; Chandrasegaran, K.; Durante, Z.; Eyzaguirre, C.; Bisk, Y.; Niebles, J. C.; Adeli, E.; Fei-Fei, L.; et al. 2025. Re-thinking Temporal Search for Long-Form Video Understanding. *arXiv preprint arXiv:2504.02259*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; et al. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Liang, Z.; Xiao, S.; Qin, M.; Yang, X.; Xiong, Y.; Zhang, B.; et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13691–13701.