MMViR: A Multi-Modal and Multi-Granularity Representation for Long-range Video Understanding

Anonymous ACL submission

Abstract

Long videos, ranging from minutes to hours, present significant challenges for current Multi-modal Large Language Models (MLLMs) due to their complex events, diverse scenes, and long-range dependencies. Direct encoding of such videos is computationally too expensive, while simple video-to-text conversion often results in redundant or fragmented content. To address these limitations, we introduce **MMViR**, a novel multi-modal, multi-grained structured representation for long video understanding. MMViR identifies key turning points to segment the video and constructs a three-level description that couples global narratives with fine-grained visual details. This design supports efficient retrieval of relevant information and generalizes well across various scenarios. Extensive evaluations across three tasks, including QA, summarization, and retrieval, show that MMViR outperforms the prior strongest method, achieving a 19.67% improvement in hour-long video understanding while reducing processing latency to 45.4% of the original.

1 Introduction

Over the past few years, Large Language Models (LLMs) have shown remarkable progress, achieving state-of-the-art performance on various NLP tasks (Zhang et al., 2023b,c; Yao et al., 2024; Li et al., 2024; Wu et al., 2025a). Building on this success, recent research has extended LLMs to the multi-modal domain, giving rise to Multi-modal Large Language Models (MLLMs), which broaden language understanding to visual reasoning and advance the capabilities of vision-language systems (Yin et al., 2024; Zhang et al., 2024; Zhang, 2025).

Despite their strong performance on image and short-video tasks, MLLMs remain fundamentally limited when processing long-range video content spanning tens of minutes to over an hour (Liang et al., 2024). Most current MLLMs (Zhang et al., 2023a; Qwen et al., 2025) are restricted to process

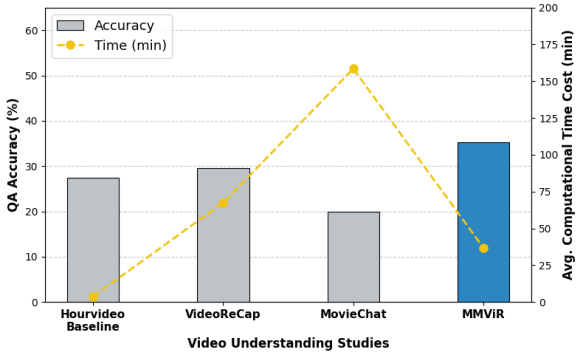

Figure 1: Performance comparison on Hourvideo QA benchmark. MMViR consistently outperforms prior methods by simultaneously optimizing for reasoning precision and inference efficiency.

a limited number of frames due to constraints in computational complexity and architecture scalability. To mitigate these issues, prior research has explored two main directions. One is to enhance MLLM architectures with specialized modules, such as long/short-term memory (Song et al., 2024; He et al., 2024) or redundancy-reduction mechanisms (Li et al., 2025; Wang et al., 2025). However, this approach still struggles to scale to extremely long videos, where the computational cost of dense frame processing becomes prohibitive. An alternative strategy is to transcribe long-range video content into textual representation, either by aggregating clip textual descriptions (Chandrasegaran et al., 2024; Pang and Wang, 2025) or multi-level reasoning chains (Sanders et al., 2024). Nevertheless, such representations often suffer from content fragmentation that undermines global narrative coherence and overlooks critical details.

This persistent trade-off leads us to rethink the fundamental design of video representations, specifically by focusing on two key dimensions: 1) **Structure:** identifying the optimal way to transform long video content without sacrificing key information; and 2) **Modality:** exploring whether incorporating additional modalities beyond textual

descriptions can represent video information more effectively. Using VideoQA as a benchmark, we evaluate various representation strategies by analyzing the trade-off between accuracy and token consumption. This analysis leads to two key findings. First, **multi-grained structures**, spanning from high-level summaries to fine-grained details, provide a more coherent way to present long videos. Second, **multi-modal representations** that fuse text with visual representation offer better compression efficiency than unimodal approaches.

Based on these findings, we introduce MMViR, a Multi-modal and Multi-grained Video Representation designed for long-video understanding. MMViR organizes long videos into a structured, searchable multi-grained format with three distinct levels: 1) a global timeline that captures the overall narrative flow; 2) clip-level coarse-grained textual descriptions from multiple perspectives to refine the timeline; 3) fine-grained visual representations paired with concise text for precise grounding. This structure ensures comprehensive information coverage and efficient retrieval, yielding a general representation that remains independent of any specific query. As a result, MMViR enables downstream tasks to access only necessary content, thereby significantly reducing computational overhead.

We evaluate MMViR on three core long-video tasks (i.e., VideoQA, summarization, and retrieval), using datasets of varying temporal scales. The results show that MMViR consistently outperform prior methods in both accuracy and computational efficiency. These findings, illustrated in Figure 1, confirm the advantages of our approach for long-range video reasoning. In summary, our main contributions are:

- We conduct an in-depth analysis of the trade-offs between efficiency and accuracy in long-video understanding, identify key factors for effective representation, and demonstrate how to implement these factors in practice.

- We introduce MMViR, a multi-modal and multi-grained framework that bridges global coherence with fine-grained detail. By decoupling representation from specific queries, MMViR enables efficient, query-agnostic retrieval across diverse long-video tasks.

- Extensive experiments on various datasets and tasks show that MMViR outperforms previous methods in both efficiency and accuracy, particularly on hour-long video benchmarks.

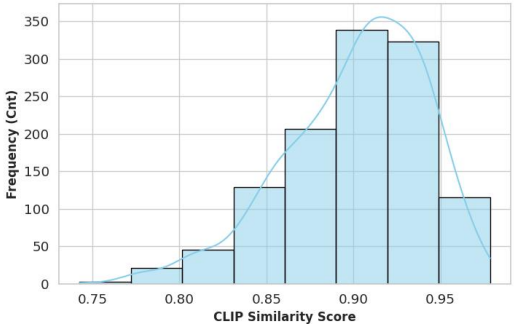

Figure 2: Overall distribution of frame-to-frame CLIP Similarity Scores of a sampled hour-long video.

2 Related Work

2.1 Long Video Understanding

Video understanding has long been a core challenge in the vision domain. Early research focused on short-term action recognition and classification (Soomro et al., 2012; Reddy and Shah, 2013; Kay et al., 2017; Carreira and Zisserman, 2017) using 2D CNNs with RNNs (e.g., LSTMs) (Donahue et al., 2015), or 3D CNNs (Taylor et al., 2010; Ji et al., 2012; Tran et al., 2015) to capture temporal information. While effective for short clips, these models could not achieve a full understanding of video content. Later, the introduction of longer video datasets (Rohrbach et al., 2015; Xu et al., 2016; Zhou et al., 2018) shifted the focus toward text–video alignment, leading to tasks like video description (Yao et al., 2015; Venugopalan et al., 2015; Pan et al., 2016) and summarization (Zhao et al., 2017; Rochan et al., 2018; Zhang et al., 2018b). However, these approaches often produced template-like outputs and missed fine-grained details (Li et al., 2019; Apostolidis et al., 2021).

The advent of cross-modal pretraining, led by CLIP (Radford et al., 2021), further advanced video understanding through models such as VideoCLIP (Xu et al., 2021). More recently, MLLMs (Zhang et al., 2023a; Qwen et al., 2025) have integrated visual modalities into LLMs, yet they often struggle with hour-long videos due to restricted context windows. Current efforts to address this, such as memory modules (Song et al., 2024), redundancy reduction (Li et al., 2025), and text-based video conversion (Chandrasegaran et al., 2024; Pang and Wang, 2025), only partially mitigate the problem

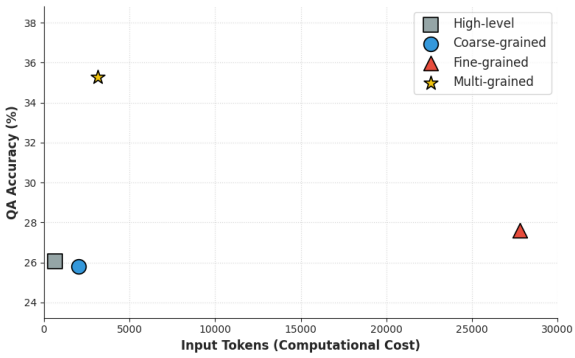

Figure 3: Token cost vs. QA accuracy for different granularities of video representations.

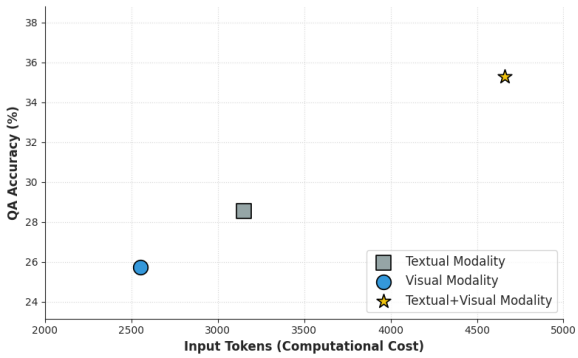

Figure 4: Token cost vs. QA accuracy for different modalities of video representations.

and often incur high computational cost or produce incoherent information. These persistent gaps highlight the need for a structured, semantically faithful representation for long-video understanding.

2.2 Video Representation

Using language models to encode videos into structured representations has become an effective strategy for video understanding. Earlier methods (Zhang et al., 2018a; Li et al., 2020; Xiao et al., 2022; Ashutosh et al., 2023; Cheng et al., 2024) typically trained language models on video–text pairs to build multi-level representations. Recently, advances in MLLMs have enabled the generation of structured representations through multi-level captioning, in which models produce descriptions that serve as organized video summaries (Zala et al., 2023; Islam et al., 2024; Sanders et al., 2024; Wu et al., 2025b). However, most current methods of this type are query-specific, meaning they build representations tailored to the single given query. This focus often limits global coherence and makes it difficult to generalize representation to other tasks.

In contrast, MMViR generates multi-grained, multi-modal representations in a query-agnostic manner, resulting in a coherent, retrieval-friendly structure that can be efficiently reused across diverse downstream tasks without being reconstructed for every new query.

3 Preliminary Investigation

In this section, we conduct a preliminary study to understand the intrinsic features of long video content. Instead of treating video as a flat sequence, we investigate its structure through three aspects: temporal coherence, information structure, and modal synergy. These observations provide the foundation and design principles for our proposed video representation.

3.1 Temporal Coherence of Long Video

Long videos are inherently non-uniform, consisting of many events with varying durations. While segmenting videos into fixed intervals is a common baseline to manage computational costs, this approach often disrupts the natural flow of information. To investigate the intrinsic structure of long-form content, we analyzed frame-to-frame visual similarity in the *Hourvideo* dataset (Chandrasegaran et al., 2024). Specifically, we analyze 50 videos from it and sample at 0.5 fps, computing frame-to-frame similarity using CLIP (Radford et al., 2021). As Figure 2 shows, frame similarities follow a "long-tailed" trend, indicating that **frame continuity is likely to be punctuated by these sparse similarity drops**, which serve as natural indicators of scene transitions (more details in Appendix A). This semantic-aware segmentation, such as using "turning points" determined by CLIP or Kernel-based Temporal Segmentation (KTS) (Potapov et al., 2014), aligns more closely with the video's natural than fixed-interval splitting.

3.2 Impact of the Data Structure

After determining *where* to segment, we investigate what information should be preserved within each segment. Drawing on insights from prior work (Zala et al., 2023; Sanders et al., 2024), long-form reasoning often requires a way to *zoom in and out*, locating an event across a duration and then analyzing its specific details. We simulate this by generating textual descriptions at three level granularities via MLLM (Zhang et al., 2023a), including global summaries (i.e., high-level), action details (i.e., coarse-grained), and frame-specific descriptions (i.e., fine-grained). By getting the response to

Figure 5: Overview of the MMViR and the downstream inference. Given a long video, MMViR leverages an MLLM to construct a multi-modal and multi-grained representation for it. The high-level timeline descriptions serve as a global semantic index, enabling efficient query-aware relevant clip localization. This design ensures a synergistic balance between global narrative coherence and fine-grained evidence, effectively addressing the computational challenges of long-form content.

1,162 queries from *Hourvideo* generated from GPT-4o (OpenAI, 2023). For multi-grained inference, we first retrieve the top-10 most relevant clips using the global summaries, then provide the action details and frame description as context. We observe that relying on single granularity cannot achieve better performance than using a combination, as shown in Figure 3.

> **Observation 1:** An effective representation must bridge multiple scales. Global summaries act as an "index" for retrieval, while local details provide the necessary "evidence" for complex reasoning.

3.3 Impact of Data Modality

Finally, we examine whether textual descriptions alone are sufficient to represent these segments. While text is efficient for indexing, it often loses subtle visual information (e.g., specific spatial relationships). We compare three configurations: text-only (i.e., multi-grained text description), vision-only (i.e., initial sampled frames), and a hybrid multi-modal approach. As shown in Figure 4, the combination representation substantially outperforms unimodal baselines with manageable computational costs.

> **Observation 2:** Textual and visual modalities are complementary. Multi-modal representations are more expressive, while maintaining a reasonable computational overhead.

4 Method

4.1 Overview of MMViR

Guided by the insights in Section 3, we propose MMViR. Unlike flat representations, MMViR explicitly models the multi-modal nature of long-form videos by organizing information into three levels of granularity: 1) *High Level Global Timeline*: concise summaries of the dominant activity in each segmented clip; 2) *Clip-level Coarse-grained Description*: multi-perspective details regarding actions, objects, and scenes content; 3) *Fine-grained Description*: localized visual-textual pairs extracted from sampled frames for grounding. As shown in Figure 5, MMViR maintains a comprehensive record of video content while providing a retrieval-friendly format that generalizes across diverse downstream tasks without requiring query or task-specific reconstruction.

4.2 Construction Pipeline of MMViR

Event-based Video Segmentation. Given a long video $\mathbb{V}$ spanning $[0, T]$, we first sample frames at a low fps (e.g., 0.5) and compute the CLIP (Radford et al., 2021) similarity between consecutive frames. Following our observations of temporal non-uniformity in Section 3, we detect the turning points $0 = t_0 < t_1 < \cdots < t_N = T$ within the similarity distribution (or via the KTS method (Potapov et al., 2014)). These points serve as indicators of potential scene changes to segment the video into N coherent clips, denoted as $C_i = \mathbb{V}[t_{i-1}, t_i]$ for $i = 1, \ldots, N$.

Hierarchical Representation Generation. For each segmented clip C_i, we utilize an MLLM (Zhang et al., 2023a) to generate a multi-modal, multi-grained representation tuple $R(C_i) = (R^{timeline}(C_i), R^{coarse}(C_i), R^{fine}(C_i))$. That is, the MLLM first generates a concise, one-sentence summary s_{C_i} that captures the dominant activity and key events of the entire clip. The clip's overall

activity and key events. This serves as the timeline-level representation:

$$R^{timeline}(C_i) = \{s_{C_i}\}, \quad (1)$$

To achieve higher semantic precision at the coarse-grained level, we divide each clip C_i into sub-segments of a fixed maximum duration. This further segmentation prevents the MLLM from overlooking critical actions in longer clips. For each sub-segment, the MLLM produces detailed descriptions $d(k)$ of the actions A_k, scenes S_k, and objects O_k:

$$R^{coarse}(C_i) = \{d(k) = (A_k, S_k, O_k)\}, \quad (2)$$

where A_k, S_k, and O_k denote the actions, scenes, and objects in the k-th sub-segment.

Finally, we sample frames within each sub-segment at a lower density. The MLLM pairs these with scene descriptions to form text-frame pairs $f_{k,m}$:

$$R^{fine}(C_i) = \{f_{k,m} \mid m = 1, \ldots, M_k\}. \quad (3)$$

where M_k is the number of sampled frames in the k-th sub-segment.

4.3 Query-based Retrieval and Reasoning

MMViR's structured design enables an *adaptive retrieval mechanism* tailored to specific task requirements. For global tasks like summarization, we leverage the full $R^{timeline}$ to capture the overall narrative. For localized reasoning tasks (e.g., VideoQA), we employ a two-stage process. In the first stage, a retrieval model (e.g., Contriever (Izacard et al., 2021)) is used to retrieve the top-k relevant clips using $R^{timeline}$ as the search index. Then, only the corresponding R^{coarse} and R^{fine} of the retrieved clips are expanded as content for the MLLM. This mechanism allows the downstream inference to "zoom in" on localized evidence while retaining global context, maintaining high accuracy while significantly reducing the computational cost of long-video reasoning.

5 Experiment

We evaluate MMViR across five diverse long-form video datasets, covering three core tasks: VideoQA, summarization, and video retrieval. The results consistently validate the robustness and generality of our representation, highlighting MMViR's unique ability in balancing semantic completeness with computational efficiency across varying temporal scales.

Dataset	#Video	#Query	Duration	Tasks
HourVideo	50	1,162	20-120m	VideoQA, Sum.
EgoSchema	500	500	∼3m	VideoQA
Video-MME	60	64	∼30m	VideoQA
MovieChat	448	448	7-8m	Sum.
Ego4D	50	36k/516	20-120m	Video Retrieval

Table 1: Dataset statistics and evaluation tasks (Ego4D's query consists of 36k narrations and 516 summaries).

5.1 Experiment Setup

Dataset. We evaluate MMViR on five benchmarks, with detailed statistics provided in Table 1.
1) Hourvideo (Chandrasegaran et al., 2024): Our primary evaluation benchmark for extreme long-range video reasoning, providing hour-long videos paired with five-way multiple-choice QA pairs and summarization ground-truth. **2) Egoschema** (Mangalam et al., 2023): Used to evaluate the understanding of egocentric activities, where each question requires global information of the associated video content. **3) Video-MME** (Fu et al., 2025): A benchmark provides a diverse set of six visual domains, which are used to evaluate the robustness across various content types of our representation in VideoQA. **4) MovieChat-1K** (Song et al., 2024): Comprising diverse movies and TV series with manual annotations, specifically used to evaluate narrative summarization capabilities. **5) Ego4D** (Grauman et al., 2022): The benchmark offers high-density temporal narrations, which are sampled and are used to measure the efficiency and precision of video retrieval in long-form sequences.
Evaluation Metrics. To assess the effectiveness and efficiency of MMViR, we categorize our evaluation into performance- and cost-based metrics:
1) VideoQA. *Accuracy*: The percentage of correctly answered questions. *Time and Token Consumption*: The average time (in minutes) and input tokens to process the hour-long video (i.e., Hourvideo dataset) before QA. **2) Summarization.** The standard evaluation metrics are adopted, including ROUGE-2, ROUGE-L (Ng and Abrecht, 2015), and METEOR (Banerjee and Lavie, 2005), to assess the quality of generated summaries. **3) Video Retrieval.** The localization precision is assessed at two temporal scales: *Frame-level PrecisionK:* Measuring the accuracy of pinpointing specific ground-truth frames. *Range-level Overlap@K:* Evaluating the temporal intersection between retrieved segments and ground-truth intervals to measure temporal alignment.
Baselines. To evaluate MMViR against current

Method	Modality	Structure	Avg. Time (min) ↓	Avg. Input Tokens ↓	Acc ↑ (Hourvideo)	Acc ↑ (Egoschema)	Acc ↑ (VideoMME)
Directly Sample Frames							
Uniform	Image	Global	–	4250.0	26.4	60.4	**54.7**
Random	Image	Global	–	4250.0	26.6	61.8	48.4
Comparable Baseline							
Hourvideo (2024)	Text	Fine-grained	3.8	4987.2	27.5	34.2	11.7
VideoRecap (2024)	Text	Multi-grained	67.5	6385.0	29.5	34.4	33.3
MovieChat (2024)	Image	Global	158.4	**377.3**	19.9	53.5	38.2
Ours (Kernel Temporal Segmentation)							
MMViR (w/o img)	Text	Multi-grained	33.1	4347.4	25.2	-	42.2
MMViR (w/o text)	Image	Fine-grained	**0.8**	2550.0	28.2	-	48.4
MMViR	Text+Image	Multi-grained	33.9	5834.5	30.5	-	46.9
Ours (Clip-based Segmentation)							
MMViR (w/o img)	Text	Multi-grained	36.0	3151.1	33.6	60.2	42.2
MMViR (w/o text)	Image	Fine-grained	**0.8**	2550.0	28.7	**65.4**	**54.7**
MMViR	Text+Image	Multi-grained	36.8	4661.7	**35.3**	61.8	51.7

Table 2: Performance comparison on VideoQA tasks. The bold in the table indicates the best performance, while the underlined indicates the second best.

state-of-the-art approaches, we select five representative baselines categorized by:
1) Caption-based Representations: *Hourvideo Baseline* (Chandrasegaran et al., 2024), generates multi-aspect textual descriptions (actions, objects, and scenes) for fixed one-minute video segments. *VideoRecap* (Islam et al., 2024), a recursive captioning framework that produces hierarchical descriptions at clip, segment, and video levels, with a focus on atomic human actions. **2) Memory-based Architectures:** *MovieChat* (Song et al., 2024), a model that utilizes a long/short-term memory management mechanism. We evaluate it in its *global mode* to process the entire video sequence end-to-end. **3) End-to-End Multi-modal Retrievers:** *CLIP4Clip* (Luo et al., 2022)., adapts CLIP features for video-text retrieval by calculating similarity between visual and textual embeddings. *MetaCLIP 2* (Chuang et al., 2025), a state-of-the-art multimodal retriever trained on web-scale multilingual data, used here as a high-performance baseline for textual-visual retrieval.

Implantation Details. All experiments are conducted on an NVIDIA RTX A6000 GPU. Videos are initially sampled at 0.5 fps. For video segmentation, "turning points" are identified using either CLIP (Radford et al., 2021) similarity with a 2nd-percentile threshold, or the KTS (Potapov et al., 2014). For each video clip, we use VideoLLaMA-7B (Zhang et al., 2023a) to generate multi-grained textual descriptions, while frames are sub-sampled at 0.05 fps for fine-grained grounding. For VideoQA task, we use Contriever (Izacard et al., 2021) to first retrieve top-10 relevant timeline descriptions. These, along with their corresponding multi-modal clip representations, are then fed into GPT-4o (OpenAI, 2023) for answering. For summarization task, we feed the complete set of timeline-level descriptions directly into GPT-4o to generate a comprehensive video summary. For video retrieval task, we use Contriever to index all timeline descriptions, returning the top-k segments most relevant to the query. More implementation details are provided in the Appendix B.

5.2 Video Question Answering

Table 2 presents a comprehensive comparison of performance between MMViR and baseline models across three long-form VideoQA benchmarks, evaluated on two primary dimensions: *predictive accuracy* and *computational efficiency* (measured by processing time). As shown, **MMViR consistently achieves superior QA accuracy across all datasets**. Notably, on the extremely-scale HourVideo benchmark, MMViR achieves the best performance while significantly reducing the processing time compared to competing approaches. For datasets with shorter durations and lower scene complexity, such as EgoSchema and Video-MME, we observe that directly retrieving high-precision

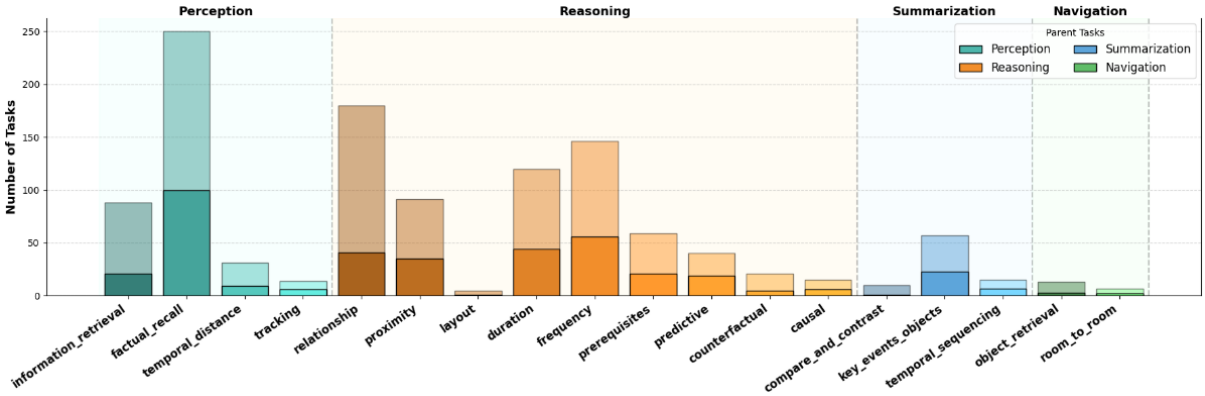

Figure 6: Distribution of Task Categories in Hourvideo Dataset and the Number of Correct Answers by MMViR

Method	Hourvideo			MovieChat-1K		
	R-2	R-L	MET.	R-2	R-L	MET.
Hourvideo	10.14	**27.33**	33.87	3.31	16.54	17.16
VideoRecap	9.15	24.99	33.54	1.12	11.93	13.63
MMViR (Ours)	**10.67**	27.15	**34.98**	**4.27**	**18.74**	**18.52**

Table 3: Summarization performance comparison on HourVideo and MovieChat-1K datasets.

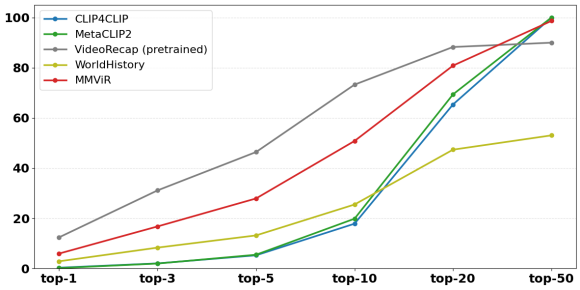

Figure 7: Performance of MMViR and baseline methods on the video retrieval task using **narration data**.

visual descriptions often yields better results than full multi-modal representations. This suggests that for simpler narratives, visual-based scene descriptors provide a more distilled, less noisy context for reasoning.

Unlike baseline methods that cause substantial computational overhead due to intensive memory

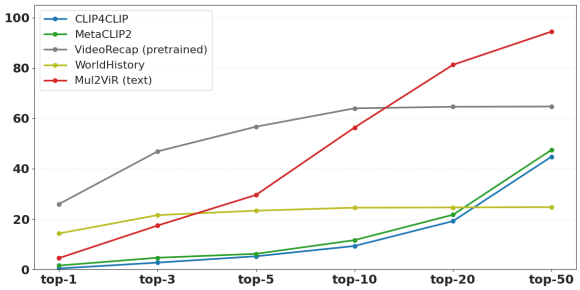

Figure 8: Performance of MMViR and baseline methods on the video retrieval task using **summary data**.

management or exhaustive multi-level captioning, MMViR effectively balances performance and cost. By leveraging an adaptive retrieval mechanism to extract only query-essential information, MMViR demonstrates superior scalability for long-video understanding, effectively bypassing typical computational bottlenecks.

5.3 Video Summarization

Table 3 evaluates the summarization capabilities of MMViR and comparative baselines. The results indicate that MMViR consistently outperforms competing methods, suggesting that our multi-granularity structure effectively preserves both global narrative coherence and fine-grained semantic details. While baselines that often struggle with information redundancy or context loss in hour-long videos, MMViR's timeline-level summaries provide a structured blueprint of the corresponding video. This allows the LLM to receive comprehensive, logically consistent video abstracts as the input background context, without being overwhelmed by irrelevant temporal noise.

5.4 Video Retrieval

Figures 7 and 8 compare the retrieval performance of MMViR against both text-based and multi-modal baselines. Notably, MMViR outperforms all non-pretrained baselines, with Top-50 accuracy even surpassing VideoReCap, a model explicitly pretrained on the Ego4D domain. The performance gap is obvious in the lower top-k ranges (e.g., $K < 10$), where end-to-end multi-modal models like MetaCLIP2 and CLIP4Clip struggle to achieve high precision. This pattern suggests that directly matching queries to raw visual frames in unconstrained long videos remains challenging due

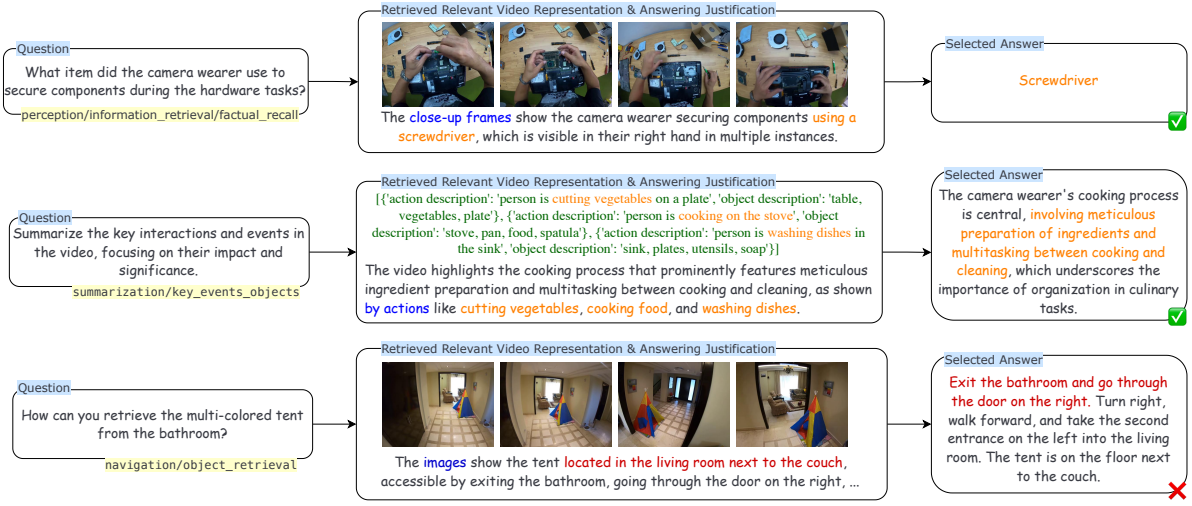

Figure 9: Case study of VideoQA reasoning on HourVideo. Text highlighted in orange/red indicates the logic supporting correct/incorrect answering. The blue-highlighted text indicate the reference modal retrieved from MMViR's representations.

to visual redundancy and semantic gaps. In contrast, MMViR's success underscores the superiority of structured, multi-level textual representations, which serve as a semantic bridge that simplifies complex cross-modal mapping into a more manageable text-to-text alignment task.

6 Analysis and Discussion

6.1 Performance Analysis Across Different Task Categories

To further investigate the specialized capabilities of MMViR, we analyze its performance across diverse task sub-types within the HourVideo dataset. As shown in Figure 6, MMViR demonstrates robust proficiency across *Perception*, *Reasoning*, and *Summarization* parent tasks, consistently achieving accuracies above 0.4. In particular, MMViR excels in *factual recall* in *Perception*, *causal* in *Reasoning*, and *key event object* in *Summarization*. This performance validates that MMViR's multi-grained representation effectively preserves global narrative coherence while maintaining the semantic depth necessary for complex multi-hop reasoning.

However, performance remains constrained in *Navigation* (avg. acc. 0.25). This discrepancy suggests that while a retrieval-centric framework is highly effective for semantic-based understanding, **it encounters a "bottleneck" in precise spatio-temporal grounding and explicit geometric relational modeling**. These findings highlight a direction for future studies to integrate a more robust spatial-aware module into long-video reasoning.

6.2 Case Analysis

Figure 9 presents representative examples of MMViR's reasoning process on the HourVideo dataset. For *Perception* task, MMViR demonstrates a precise ability to anchor its judgments on specific retrieved frames that reveal critical evidence for answering. Meanwhile, for higher-level *Summarization* tasks that require broad contextual synthesis, the relevant timeline-level descriptions function as a semantic backbone. By bridging temporal gaps that discrete frames cannot present, these textual summaries provide a continuous narrative flow, allowing the model to maintain global coherence across extreme durations. Despite these strengths, the MLLM sometimes misinterprets the background context provided by MMViR due to its intrinsic bottlenecks in cross-modal integration, leading to inaccurate responses. This highlights a future direction to optimize a simpler yet effective representation for MLLM to better understand.

7 Conclusion

In this paper, we introduce MMViR, a multi-modal, multi-grained representation designed for long-video understanding. By integrating fine-grained visual information with structured textual descriptions, MMViR effectively addresses the challenges of information redundancy and computational constraints. Extensive experiments validate its robustness across diverse tasks, offering a promising paradigm for future research into efficient, long-range video comprehension.

Limitations

While MMViR demonstrates robust performance across long-video benchmarks, several limitations remain for future investigation. First, our fine-grained visual representations rely on uniform sampling of frames. Although this ensures representation completeness, it inevitably introduces temporal redundancy, especially in retrieval tasks where many sampled frames may contain static or redundant information.

Second, our retrieval mechanism adopts a top-down approach, primarily leveraging high-level timeline summaries to localize relevant segments. Although computationally efficient, this strategy may occasionally suffer from a granularity mismatch, potentially overlooking subtle, fine-grained visual information that is not explicitly captured in the textual summaries. Future research could explore content-aware adaptive sampling to prioritize informative frames. In addition, developing more advanced cross-modal indexing structures that allow for bi-directional interaction between global narratives and local visual details, to further refine the precision and scalability of long-video comprehension.

References

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. 2023. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078.

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

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. 2024. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37.

Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. 2024. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *arXiv preprint arXiv:2410.14324*.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, and 1 others. 2025. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. 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*, pages 24108–24118.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. 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)*.

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

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and

1 others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Videochat-flash: Hierarchical compression for long-context video modeling. *Preprint*, arXiv:2501.00574.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Jun Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1925–1930.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1029–1038.

Ziqi Pang and Yu-Xiong Wang. 2025. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*.

Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Kishore K Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981.

Mrigank Rochan, Linwei Ye, and Yang Wang. 2018. Video summarization using fully convolutional sequence networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 347–363.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Kate Sanders, Nathaniel Weir, and Benjamin Van Durme. 2024. Tv-trees: Multimodal entailment trees for neuro-symbolic video reasoning. *arXiv preprint arXiv:2402.19467*.

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

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. 2010. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2025. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. *arXiv preprint arXiv:2503.12559*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025a. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Rujie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. 2025b. Longvitu: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*.

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. 2022. Hierarchical self-supervised representation learning for movie understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9727–9736.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

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

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. 2023. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065.

Bowen Zhang, Hexiang Hu, and Fei Sha. 2018a. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the european conference on computer vision (ECCV)*, pages 374–390.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Preprint*, arXiv:2306.02858.

Haopeng Zhang. 2025. Bridging multimodal and video summarization: A unified survey. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 157–171.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023c. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*.

Ke Zhang, Kristen Grauman, and Fei Sha. 2018b. Retrospective encoders for video summarization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–399.

Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

A Clip-based Video Segmentation Analysis

Beyond the analysis of the overall distribution of frame-level CLIP similarity scores in Section 3, we further examine the distribution of CLIP similarity scores over the timestamps within an entire video. As illustrated in Figure 10, the local minima (troughs) of the similarity curve are not clustered but are dispersed relatively uniformly across the video timeline. This observation provides additional empirical justification for employing a lower quantile of the CLIP similarity distribution (e.g., the 2nd-percentile) as the threshold for video segmentation.

B Additional Implementation Details

For video segmentation, to ensure semantic coherence, we enforce a minimum clip duration of 5 minutes and a sub-segment minimum of 1.67 minutes.

For long-form videos (e.g., the Hourvideo dataset), we employ a sub QA task-specific selection strategy to balance performance and inference cost. For global tasks like summarization-related in the Hourvideo dataset, we utilize the entire timeline-level context. For local tasks involving spatial or object reasoning, we extract one or more granular descriptions (fine-grained and coarse-grained sequences) specifically from the top-10 retrieved segments for the task-related aspects. For example, for Spatial-related tasks (e.g., reasoning/spatial/relationship), we focus on spatial content by utilizing the fine-grained descriptions (both textual and visual) and coarse-grained scene sequence descriptions for the top-10 retrieved segments, while omitting timeline-level and other coarse-grained details. For shorter videos (EgoSchema, VideoMME), where the context length is manageable, we provide an entire video representation input comprising all three representation levels for the top-10 retrieved segments.

In addition, for the Directly Sample method, we downsample the 0.5 FPS candidates to match GPT-4o's input constraints, feeding a maximum of 50 frames. In contrast, MMViR requires 10 relevant images when fine-grained descriptions are excluded, and approximately 30 images in total when fine-grained details are included.

C Impact of the Video Representation and Answer Generator

To further verify the effectiveness of MMViR and demonstrate that its superior performance is not s solely attributed to the specific MLLMs' capabilities, we conducted an additional experiment on the Hourvideo dataset for the downstream QA task. Specifically, we *replaced GPT-4o with an open-source LLM*, Qwen2.5-7B-Instruct, as the answer generator. By feeding the same retrieved context into this model, we compared MMViR against the Hourvideo baseline and the VideoRecap baseline. As shown in the Table 4 below, while the absolute accuracy naturally decreases when using Qwen-7B compared to GPT-4o, MMViR consistently outperforms the baselines. Furthermore, the performance margin remains stable, demonstrating that the efficacy of MMViR is model-agnostic and does not rely on the superior capabilities of a specific *stronger LLM* to achieve its gains.

Method	Qwen-7b Acc (%)	GPT-4o Acc (%)
Hourvideo	25.20	27.50
VideoRecap	27.12	29.50
MMViR (clip-based)	**34.82**	**35.30**

Table 4: MMViR's performance comparison across different MLLM answer generators.

Moreover, to address concerns regarding the upper video representation generator model's dependence, we *substituted VideoLLaMA* and the answering LLM with the Qwen2.5-7B-Instruct model to re-evaluate MMViR's performance, as shown in Table 5. The results indicate that although absolute accuracy slightly declines compared to the GPT-4o-based pipeline, MMViR consistently provides effective and efficient answers. This confirms that our proposed method remains superior regardless of the specific open-source LLM employed.

Method	Qwen-7b Acc (%)
MMViR (w/o img)	23.18
MMViR (clip-based)	28.43

Table 5: MMViR's performance on VideoQA, utilizing Qwen-7b uniformly for both video representation and answer generation.

D Impact of Retrieval Scale

Figure 11 investigates the sensitivity of VideoQA performance to the number of retrieved video clips

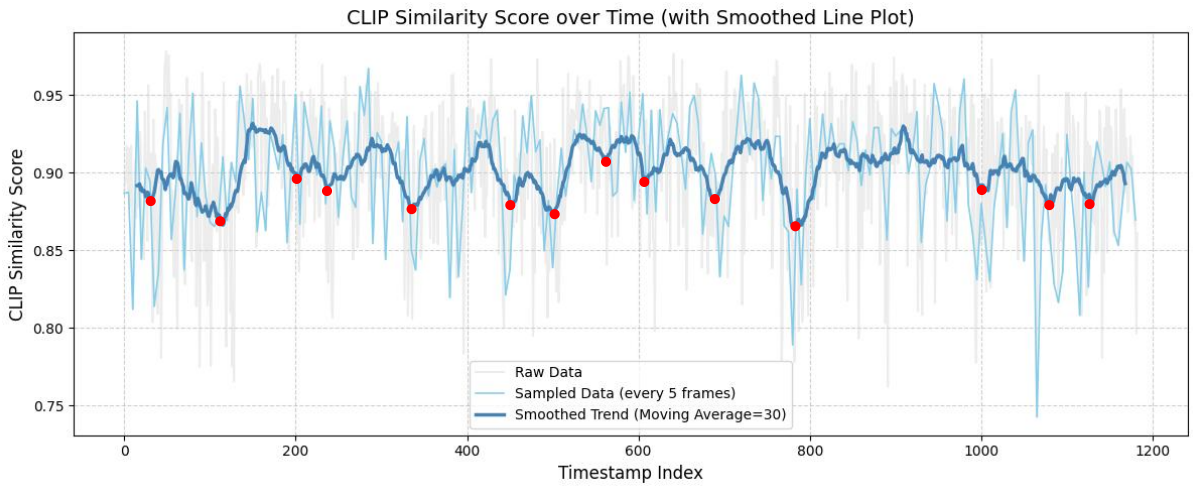

Figure 10: The distribution of CLIP similarity scores over the timestamps and its smoothed curve. The red markers indicate timestamps where noticeable turning points occur.

and their corresponding representations within MMViR. As shown, retrieving a larger set of relevant clips provides richer contextual cues, thereby consistently enhancing QA accuracy. However, this performance gain comes at the cost of increased inference latency and computational overhead, as a higher volume of retrieved descriptions substantially increases the LLM's input token size. This observation underscores a fundamental trade-off between accuracy and efficiency in long-video understanding. Consequently, the use of an adaptive retrieval strategy is essential to maintain high performance while ensuring computational scalability, allowing the model to selectively process query-critical information without incurring prohibitive token costs.

E Instruction Prompt

In this section, we provide the detailed prompts used to generate the textual representations for MMViR. Following the the descriptive dimensions defined in the HourVideo dataset (Chandrasegaran et al., 2024), we employ the following prompts to construct the textual video representations:

High-level Timeline Description

```
Your task is to analyze the given video
frame sequence extracted from a long video
for a detailed video understanding exercise,
focusing on the **motion**, to identify and
describe all of the **actions** appear in the
images, and **where** whey take place by **who**,
as well as the corresponding objectives.
YOU ARE ALLOWED TO USE A MAXIMUM OF 50 words for
this description. Please only return the
description.
```

Coarsed-grained (Action Sequence Description)

```
Your task is to analyze video frames extracted
from a video for a detailed understanding.
I will provide a frame sequence with each frame
spaced every 2 seconds. Examine these frames
closely and generate a comprehensive caption
by strictly following:
List the sequence of all **actions** and the
corresponding **objects** in the order they
occur in the given frames.
YOU ARE ALLOWED TO USE A MAXIMUM OF 200 words for
this description.
PLEASE Strictly return your results by a list of
dict in JSON format, following the example below:
[{'action description': cooking the sausages
and eggs}, {'action description': cleaning up
dishes}]
Please do not return a empty list as the result.
If there is no action appears in the given frames,
please return the string: 'no action detected'.
```

Coarse-grained (Scene Sequence Description)

```
Your task is to analyze video frames extracted
from a video for a detailed understanding.
I will provide a frame sequence with each frame
spaced every 2 seconds. Examine these frames
closely and generate a comprehensive caption by
strictly following:
List all of the actions and their corresponding
**settings** that appear in the given frames.
YOU ARE ALLOWED TO USE A MAXIMUM OF 200 words for
this description.
PLEASE Strictly return your results by a list of
dict in JSON format, following the example below:
[{'description': The man opens a cabinet in the
kitchen, 'setting': Kitchen, 'action': Opening
a cabinet}, {'description': A person is watering
plants in a garden, 'setting': Garden, 'action':
watering}]
Please do not return a empty list as the result.
If there is no action appears in the given frames,
please return the string: 'no action detected'.
```

Coarse-grained (Object Description)

```
Your task is to analyze video frames extracted
from a video for a detailed understanding.
I will provide a frame sequence with each frame
spaced every 2 seconds. Examine these frames
closely and generate a comprehensive caption
by strictly following:
List the key objects and the characters that
appear in the given frames, along with their
attributes if applicable (e.g., color, shape,
texture), each attribute is separated with a
comma.
YOU ARE ALLOWED TO USE A MAXIMUM OF 200 words for
this description.
PLEASE Strictly return your results by a list of
dict in JSON format, following the example below:
[{'object_name': man, 'number': 1},
{'object_name': dog, 'number': 2, 'attributes':
yellow}]
Please do not return a empty list as the result.
If there is no object appears in the given frames,
please return the string: 'no object detected'.
```

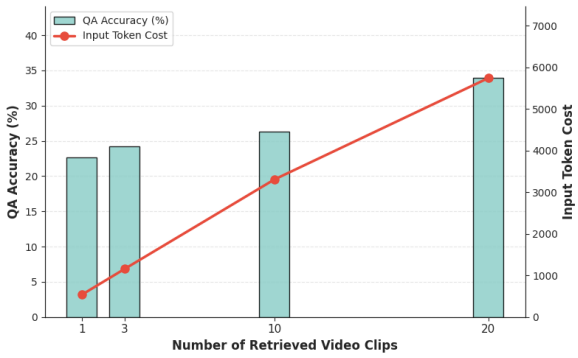

Figure 11: Impact of the Number of Retrieved Clip Representations on QA Accuracy and Inference Cost.

Fine-grained (Spatial Description)

```
Your task is to analyze the given image a
detailed understanding.
Please examine it closely and generate a
comprehensive caption by strictly following:
Observe the key objects in the image, and state
the spatial realtionships between them. List all
of key objects that appear in the image, along
with their relationships with others if
applicable, each relationship is separated with
a comma.
YOU ARE ALLOWED TO USE A MAXIMUM OF 200 words for
this description.
PLEASE Strictly return your results by a list of
dict in JSON format, following the example below:
[{'object_name': table, 'number': 1},
{'object_name': kettle, 'number': 1,
'attributes': gray, 'spatial_relationship':
[on the table, right of the kitchen]}]
Please do not return a empty list as the result.
If there is no object appears in the given frames,
please return the string: 'no object detected'.
```