MomentSeeker: A Task-Oriented Benchmark For Long-Video Moment Retrieval

Huaying Yuan[1,*], **Jian Ni**[3*], **Zheng Liu**[2*], **Yueze Wang**[2], **Junjie Zhou**[4], **Zhengyang Liang**[2], **Bo Zhao**[5], **Zhao Cao**[1†], **Zhicheng Dou**[1†], **Ji-Rong Wen**[1†]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Academy of Artificial Intelligence, [3]School of Smart Governance, Renmin University of China
[4]Beijing University of Posts and Telecommunications, [5]Shanghai Jiao Tong University

Abstract

Accurately locating key moments within long videos is crucial for solving long video understanding (LVU) tasks. However, existing benchmarks are either severely limited in terms of video length and task diversity, or they focus solely on the end-to-end LVU performance, making them inappropriate for evaluating whether key moments can be accurately accessed. To address this challenge, we propose **MomentSeeker**, a novel benchmark for long-video moment retrieval (LVMR), distinguished by the following features. First, it is created based on long and diverse videos, averaging over 1,200 seconds in duration, and collected from various domains, e.g., movie, anomaly, egocentric, and sports. Second, it covers a variety of real-world scenarios in three levels: global-level, event-level, and object-level, covering common tasks like action recognition, object localization, causal reasoning, etc. Third, it incorporates rich forms of queries, including text-only queries, image-conditioned queries, and video-conditioned queries. On top of MomentSeeker, we conduct comprehensive experiments for both generation-based approaches (directly using MLLMs) and retrieval-based approaches (leveraging video retrievers). Our results reveal the significant challenges in long-video moment retrieval in terms of accuracy and efficiency, despite improvements from the latest long-video MLLMs and task-specific fine-tuning. We have publicly released MomentSeeker[1] to facilitate future research in this area.

1 Introduction

Long-video understanding (LVU) attracts tremendous attention from the community given its importance to a variety of real-world applications, such as video analysis, embodied AI, and autonomous driving. In recent years, growing efforts have been made to advance the progress of LVU, especially those from multi-modal large language models (MLLMs). Nevertheless, it remains an unsolved problem, primarily due to challenges in both response quality and running efficiency. Long-video understanding requires a diverse set of skills from MLLMs, among which the ability to accurately identify and access key moments is crucial [29]. Although long videos contain rich contextual information, many tasks require reasoning over only a few critical moments. Solving these tasks effectively depends on the accurate retrieval of such moments. In this context, long-video moment retrieval (LVMR) serves a dual purpose: it not only reflects MLLMs' performance on long-video understanding, but also facilitates the development of retrieval methods to address LVU tasks.

[*]Co-first authors. [†] Corresponding authors
[1]Code is available on this repository: `https://yhy-2000.github.io/MomentSeeker/`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

https://doi.org/10.52202/085713-0934

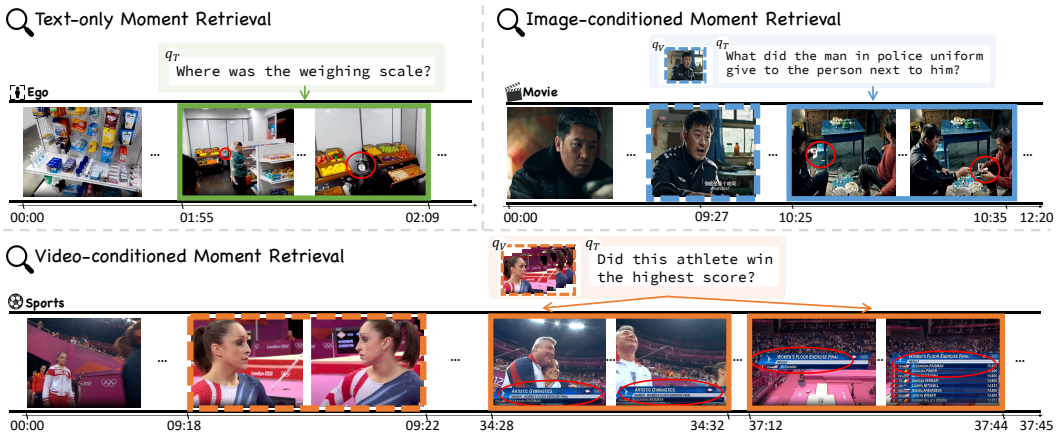


Figure 1: Demonstrative examples of the MomentSeeker benchmark. Dashed boxes denote the sources of image q_I and video q_V in the multi-modal queries, while solid boxes indicate the ground truth moment(s). Red circles mark key queried information.

However, the advancement of LVMR is severely constrained due to the lack of appropriate evaluation benchmarks. Existing moment retrieval datasets [14, 5, 15] are insufficient for LVMR, as they are created using short videos, typically only tens of seconds in duration. Besides, many of these datasets use video captions as search queries, which are overly simplistic and differ significantly from real-world LVU tasks. The recent specialized LVU benchmarks, such as Video-MME [4], MLVU [37], and LongVideoBench [29], are not well-suited for evaluating LVMR as well. Most of these benchmarks focus on the ultimate generation quality, ignoring the assessment of whether key moments are accurately retrieved. While there have been tasks like visual needle-in-a-haystack (V-NIAH) [34], they rely on synthetic test cases that emphasize frame-level reasoning, which differs fundamentally from the objective of identifying key moments for real-world LVU tasks.

To address the above limitations, we propose **MomentSeeker**, a novel benchmark for *moment retrieval in long videos*. MomentSeeker enables *task-oriented LVMR evaluation*: given a generic LVU task, e.g., causal reasoning, it assesses whether the key moments relevant to the task can be accurately identified within a long video. MomentSeeker introduces the following key features:

• **Inclusion of long and diverse videos**. MomentSeeker employs long videos while creating its evaluation tasks, with an average duration of more than 1200 seconds. This is significantly longer than those used by existing moment retrieval datasets and comparable to popular LVU benchmarks, such as VideoMME and MLVU. Additionally, the videos are sourced from diverse domains, including movie, sports, anomaly, and egocentric, ensuring broad coverage of real-world scenarios.

• **Rich forms of tasks**. MomentSeeker incorporates LVU tasks from three typical scenarios: global-level, event-level, and object-level, which comprehensively cover 9 types of popular tasks, like causal reasoning, spatial reasoning, attribute recognition, OCR, anomaly detection, etc. This sets it apart from existing benchmarks which typically rely on video captions for moment retrieval. Furthermore, MomentSeeker supports three query modalities: text-only queries, image-conditioned queries (i.e., a textual query accompanied by a reference image), and video-conditioned queries (i.e., a textual query with a reference video). These diverse query forms increase the challenge of evaluation and better reflect the complexity of real-world applications.

• **Fine-grained and high-quality annotation**. For each evaluation task, key moments are annotated with fine-grained timestamps to ensure precision and completeness. Instead of relying on MLLMs or video metadata, which can often be noisy or unreliable, MomentSeeker employs well-informed human annotators to create the tasks, thus ensuring the quality of annotations.

We conduct comprehensive experiments on MomentSeeker, where two groups of methods are studied. The first employs traditional *retrieval-based methods* [28, 38], which identify key moments by computing the relevance between query and each video chunk. The second comprises *generation-based approaches* which leverage MLLMs to directly predict key moments in long videos. We incorporate diverse models in this group, including both powerful proprietary MLLMs (e.g., GPT-4o)

https://doi.org/10.52202/085713-0934

and lightweight open-source alternatives (e.g., Qwen-VL). Our evaluation provides a comprehensive view of existing methods' performance across varying video lengths, domains, and task types, revealing that LVMR remains a challenging problem in both accuracy and efficiency.

In summary, our contributions are threefold:

1. We introduce MomentSeeker, the first specialized benchmark for long-video moment retrieval. This new benchmark fosters the development of retrieval-based methods and provides a complementary perspective to analyze MLLMs' LVU performance.

2. MomentSeeker is constructed using long and diverse videos and covers a wide range of LVU tasks. All tasks are annotated with fine-grained, high-quality labels by well-informed human workers.

3. We perform comprehensive experiments based on MomentSeeker, whose result demonstrates the challenge of LVMR in terms of both accuracy and efficiency.

2 Related Work

2.1 Long Video Understanding Benchmarks

Long video understanding has attracted growing attention, with benchmarks such as VideoMME [4] and MLVU [37] supporting tasks on videos ranging from 3 to 60 minutes. LongVideoBench [29] highlights the need for fine-grained multi-modal retrieval and reasoning by introducing referring reasoning QA tasks. However, these benchmarks evaluate long video understanding capabilities in an end-to-end manner while overlooking in-depth evaluation of moment retrieval. To assess MLLM's ability to locate key information within long contexts, V-NIAH [34] proposes a needle-in-a-haystack benchmark by inserting synthetic frames into long videos. However, due to its synthetic nature, it fails to reflect real-world multi-modal complexity and temporal dependencies. LV-Haystack [30] is a concurrent benchmark that focuses on locating task-aware keyframes in long videos. Nevertheless, it emphasizes frame-level matching and is limited in both task variety and scenario diversity.

While MomentSeeker adopts a task taxonomy resembling existing long video understanding tasks, such as Causal Reasoning, this design aims to maintain interpretability and continuity with prior work rather than mere reformulation. In contrast to previous long video understanding benchmarks that only assess answer correctness, MomentSeeker explicitly verifies grounding by requiring models to retrieve the exact temporal moment supporting each query. This addresses a key limitation: correct answers do not necessarily imply correct grounding, as models may guess without truly leveraging visual evidence. Empirically, even strong models that perform well on long video understanding benchmarks, such as Qwen2.5-VL-72B achieving 79.1 on VideoMME, obtain only 17.2 R@1 on MomentSeeker, revealing the difficulty of true moment localization. By demanding timestamp prediction, MomentSeeker enhances interpretability and factuality, making model reasoning transparent and verifiable. Moreover, accurate moment retrieval facilitates downstream long-video reasoning and agentic tasks. Overall, MomentSeeker targets the overlooked challenge of long-video moment retrieval, requiring deeper understanding, precise localization, and transparent reasoning, thus advancing the long video understanding field.

2.2 Moment Retrieval Benchmarks

Moment retrieval has been extensively studied in short-video scenarios [14, 5, 11, 8, 15], where videos typically last under three minutes and require limited temporal reasoning. Most existing benchmarks rely on synthetic settings (e.g., subtitle-based retrieval) or domain-specific content (e.g., TV shows or human activities), which restricts their applicability to real-world situations. Recent multimodal query grounding studies [33, 17] extend this line of work but still focus on short clips and rely primarily on template-based or text-only queries. Specifically, [33] employs automatically generated textual queries, while [17] incorporates audio, subtitles, and knowledge-base information to improve multimodal understanding, yet it does not address fine-grained temporal grounding. In contrast, our **MomentSeeker** benchmark targets long-video moment retrieval, featuring expert-authored, high-quality queries that require deeper temporal and causal reasoning. Additionally, long-form video benchmarks such as Ego4D and Ego-Exo4D [6, 7] contribute valuable egocentric datasets but

 https://doi.org/10.52202/085713-0934

provide only clip-level captions centered on local physical actions, lacking broader scene diversity. MomentSeeker instead spans diverse domains (e.g., sports, movies, surveillance) and introduces multimodal queries (text, text+image, text+video), enabling a more comprehensive and challenging evaluation of long-video understanding.

3 MomentSeeker

3.1 Task Definition

The long-video moment retrieval (LVMR) task is formally defined as follows. Given the query for a LVU task $\mathbf{q}$, the model is required to predict the key moments within the long video, denoted as $\mathcal{P} = [p^{(1)}, ..., p^{(k)}]$, where p^i stands for the starting and terminating timestamp of the i-th moment, i.e., p_s^i and p_t^i. The prediction result is measured by its consistency with the ground-truth moments $\mathcal{G} = [g^{(1)}, ..., g^{(m)}]$: $\phi(\{sim(p^i, g^i)|i = 1, ..., m\})$, where $sim(\cdot)$ and $\phi(\cdot)$ are the predefined similarity function and integration function, respectively (to be introduced in Section 3.5).

3.2 Video Collection

To more comprehensively evaluate the LVMR task, we construct a video dataset that is both diverse in scenarios and sufficiently long in duration, aiming to better reflect real-world conditions. To maximize diversity, we curate videos from a wide array of domains, including real-world recordings such as egocentric videos and sports, cinematic productions like movies, and simulated content like cartoons and surveillance videos for anomaly detection. Sports videos, movies, egocentric videos, and cartoons are all high-resolution, each exceeding 1080p. The sports videos are primarily sourced from Olympic tournaments. Although surveillance videos are typically low in resolution (around 320×240) due to limitations of available sources on the Internet, we apply strict filtering to retain only clips with clearly visible abnormal events, ensuring both quality and relevance. Compared to prior benchmarks that are confined to specific domains, such as TV shows [15], sports activities [5, 11], our dataset covers a broader range of visual scenarios, reflecting a wider spectrum of content, including real-world, cinematic, and simulated environments.

Furthermore, our benchmark covers a broad range of durations (Figure 2 (b)), with the longest videos extending up to approximately one hour. As shown in Table 1, our benchmark boasts an average video length of 1201.9 seconds, significantly outpacing previous moment retrieval benchmarks and aligning with the LVU benchmark. The meticulous video collection process guarantees both domain diversity and comprehensive coverage of video durations, strengthening the benchmark's ability to support LVMR tasks for the first time and its utility for real-world applications.

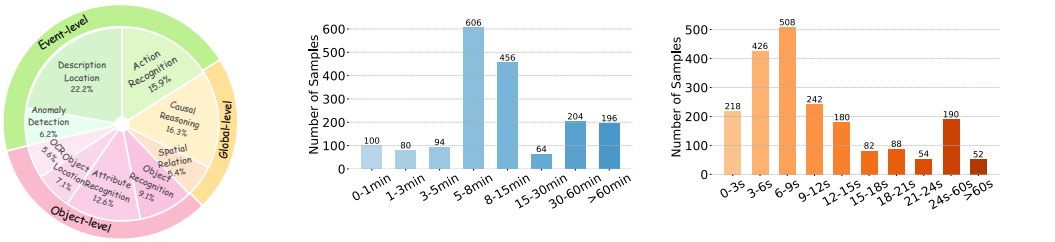
3.3 Task Creation

To evaluate video moment retrieval at different levels of semantic granularity, we construct the MomentSeeker benchmark with a hierarchical task taxonomy, as illustrated in Figure 2 (a). The benchmark encompasses a wide range of question types, each targeting distinct perception or reasoning capabilities. We categorize all tasks into three levels: global, event, and object, based on the semantic scope of the queried information. While all tasks share the common objective of grounding queries to precise temporal intervals in videos, they differ significantly in the level of reasoning, contextual integration, and perceptual detail required.

Global-Level Moment Retrieval. Global-level tasks assess the model's ability to reason over extended temporal contexts and to understand high-level semantics that span large portions of the video. These tasks often require modeling causal, temporal, or spatial relationships between multiple events or scenes. *1). Causal Reasoning* requires the model to uncover causal relationships between the question event and the answer event in order to locate the latter. For instance, answering "Why does the man need to close the bedroom window?" involves identifying a temporally distant but causally related prior event such as "It is snowing outside." Similarly, *2). Spatial Relation* requires the model to infer spatial relationships between the question event and the answer event. Questions

Benchmark	Label	Moment-targeted?	Task-oriented?	Avg. Dur (s)	#Videos	#Queries	Domain
Moment retrieval benchmarks							
TVR [15]	Auto	✓	✗	76.2	1090	5450	TV show
CharadesSTA [5]	Human	✓	✗	30.6	1334	3720	Activity
THUMOS14 [11]	Human	✓	✗	186.4	216	3457	Action
QVHighlights [14]	Human	✓	✓	150	476	1542	Vlog/News
LVU benchmarks							
VideoMME [4]	Human	✗	✓	1021.3	900	2700	YouTube
MLVU [37]	Human	✗	✓	905.8	349	502	Open
LongVideoBench [29]	Human	✗	✓	574.9	753	1337	Open
V-NIAH [34]	Auto	✗	✓	-	-	5	Open
MomentSeeker	Human	✓	✓	1201.9	268	1800	Open

Table 1: Comparison of popular moment retrieval benchmarks, LVU benchmarks (with test set statistics), and our proposed MomentSeeker benchmark. *Avg. Dur* denotes average video duration.



(a) Question Type Distribution (b) Video Duration Distribution (c) Answering Time Range Length Distribution

Figure 2: Dataset statistics. (a). Question type distribution, (b). Video duration distribution across samples, and (c) Answering time range length distribution across samples. MomentSeeker has a full spectrum of video length and covers different core abilities of the moment retrieval task.

like "How many people are there opposite the man sitting on the sofa?" demand holistic scene understanding and spatial reasoning. These tasks require integrating information across scenes, making them the most contextually demanding category.

Event-Level Moment Retrieval. Event-level tasks aim to retrieve moments corresponding to specific actions or events, requiring more localized reasoning than global tasks. The event in the query is exactly what needs to be localized. This category includes: *1) Description Location*, which matches detailed textual descriptions to video segments, focusing on visual-text alignment with minimal reasoning (e.g., "A man in a black suit and helmet is climbing up a tunnel..."). *2) Action Recognition*, which involves identifying and classifying specific actions, such as counting successful goals in a football match. *3) Anomaly Detection*, which requires detecting abnormal events without explicit textual cues (e.g., "Does the video show any activity that deviates from the normal patterns?"). This task relies on the model's ability to infer irregularities. Overall, event-level tasks target a specific event and strike a balance between the long-range semantics of global tasks and the fine-grained specificity of object-centric tasks.

Object-Level Moment Retrieval. Object-level tasks focus on identifying specific objects, their attributes, or spatio-temporal positions, emphasizing fine-grained perception and low-level visual understanding. Subtypes include: *1) Object recognition* (e.g., "What did I put on the table?"), identifying entities in the scene; *2) Object localization* (e.g., "Where was the weighing scale?"), grounding objects in space and time; *3) Attribute classification* (e.g., "What color is the ice cream in the cartoon starfish's hand?"), recognizing detailed visual properties; *4) OCR-based reasoning* (e.g., "Did this athlete win the highest score?"), detecting, reading, and analyzing the embedded text. These tasks typically span short durations and demand high spatial accuracy, contrasting with the broader semantic focus of global and event-level tasks.

https://doi.org/10.52202/085713-0934

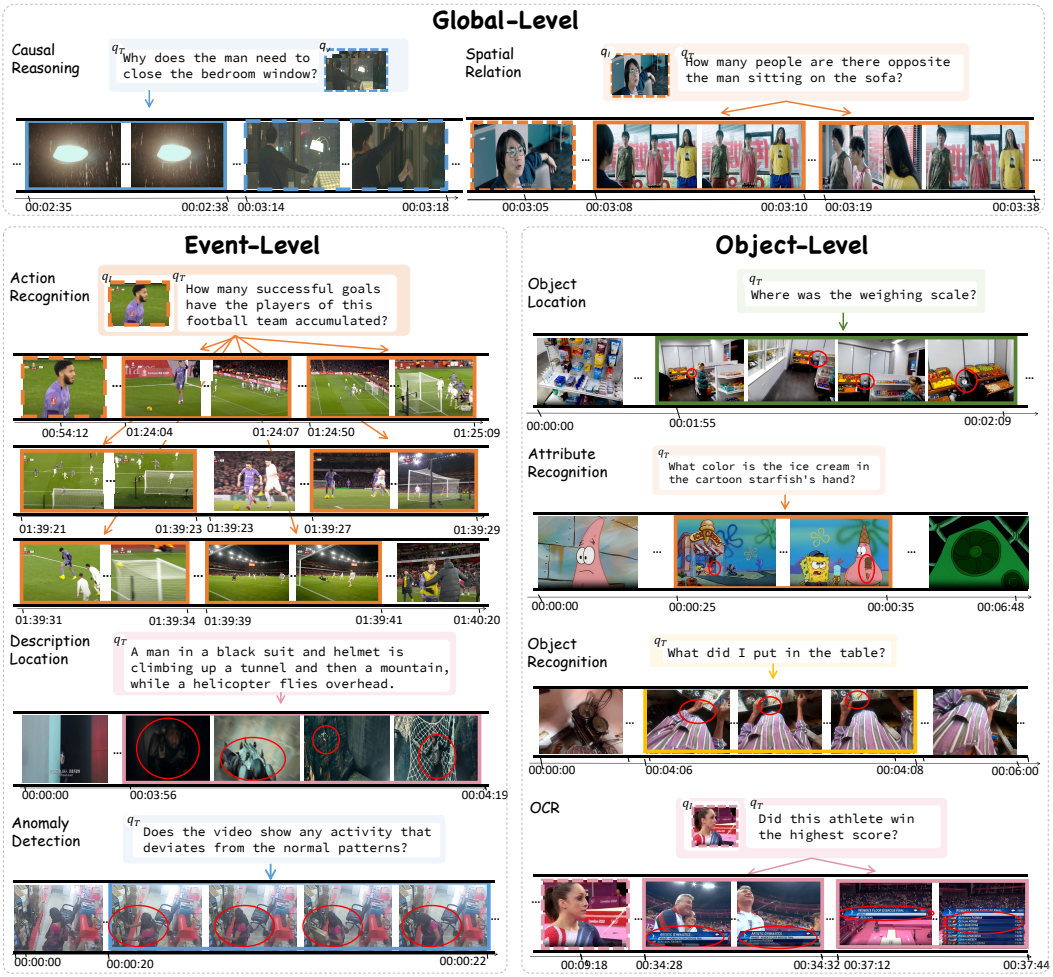


Figure 3: Examples of each task. Dashed boxes show sources of query image q_I and video q_V; solid boxes mark ground truth moments. Red circles highlight key queried information.

By organizing tasks in a hierarchical structure, the MomentSeeker benchmark provides a unified yet granular framework for evaluating the diverse capabilities of moment retrieval methods. Additionally, considering that humans often convey information using a combination of modalities (e.g., "Why does the man in this image need to close the bedroom window?"), such multi-modal expressions tend to be more natural and accurate. Therefore, in this benchmark, we incorporate three modality combinations: text-only, text with image, and text with video, with text-only serving as the primary modality and multi-modal queries acting as auxiliary. Examples of each task are shown in Figure 3.

3.4 Data Annotation

We formally define answering moments as the minimal collection of temporally localized segments that constitute both necessary and sufficient visual evidence to answer a given question. For single-detail questions (e.g., "What color is the ice cream in the cartoon starfish's hand?"), a single concise moment suffices. In contrast, multi-detail questions (e.g., "How many successful goals have the players of this football team accumulated?") span multiple, non-redundant moments, which together constitute both necessary and sufficient evidence to answer the question. As for how models identify the segments, we do not impose any constraints. For example, retrieval-based approaches may select from pre-segmented clips, while generative methods may directly predict the temporal boundaries.

Building on this foundation, we design a meticulous annotation protocol. Unlike prior datasets relying on subtitles [15] or coarse action labels [11, 5], we engage expert annotators to craft context-rich, natural questions for LVMR. Each question is paired with precise answer intervals and may be

https://doi.org/10.52202/085713-0934

text-only, image-conditioned, or video-conditioned (examples are shown in Figure 3). In the image-conditioned setting, annotators provide both a textual question and a representative frame. In the video-conditioned setting, temporal query windows serve as inputs. In terms of quantity, we ensure that text-only queries constitute the majority, with image- and video-conditioned queries serving as auxiliary, resulting in a final distribution ratio of 5:2:2. To scale annotation efficiently without compromising quality, we leverage a strong multi-modal model [25] to generate detailed descriptions of video clips for the "Description Location" task, which are then curated and refined by human annotators. After that, all samples are assigned one of nine predefined task types, including action recognition, spatial relations, causal reasoning, and others. For quality control, we first apply rule-based filtering to remove noisy samples (e.g., redundant queries, invalid intervals). Then, annotators conduct cross-checking to ensure question clarity and label validity. This two-pass process ensures consistent and reliable annotations. Figure 2 presents detailed statistics, illustrating the diverse task types, broad video durations, and varied answering times, demonstrating the benchmark's robustness and versatility for complex LVMR tasks.

3.5 Evaluation Metrics

In moment retrieval tasks, the answer to a question may involve either a single moment or multiple moments in the video. To provide a comprehensive evaluation across both single- and multi-detail scenarios, we use Recall@1 and Mean Average Precision@5 as complementary metrics.

Recall@1 (R@1) is a standard metric in moment retrieval [14, 8, 5], measuring whether the top-ranked prediction $p^{(1)}$ matches any ground-truth moment $g^j \in \mathcal{G}$ with similarity $sim(p^{(1)}, g^j)$ exceeding a predefined threshold. The similarity function sim is typically defined as Intersection over Union (IoU) [14, 8], which measures the overlap between the prediction and the ground-truth moment. Accordingly, the set $\{sim(p^{(1)}, g^j)\}$ contains the similarity scores between the prediction and all ground-truth moments. The function $\phi(\cdot)$ returns 1 if at least one of these scores exceeds the threshold, and 0 otherwise. This design accounts for the presence of multiple valid ground-truth moments, treating the prediction as correct if it sufficiently overlaps with any one of them. The final R@1 score is obtained by averaging over all queries.

Mean Average Precision@5 (mAP@5) evaluates both the accuracy and the ranking quality of the top-5 predicted temporal segments for each query. It complements Recall@1, which only considers top-1 accuracy, and is widely used in scenarios involving multiple ground-truth moment retrieval [14]. For each query, predictions are ranked by confidence, and a prediction is considered correct if its temporal IoU with any unmatched ground-truth segment exceeds a threshold. Each ground-truth can be matched at most once. The Average Precision (AP) is computed by averaging the precision at each correct prediction within the top-5 ranked results. The final mAP@5 is obtained by averaging the AP across all queries. mAP@5 reflects not only whether relevant moments are retrieved, but also how well they are ranked, rewarding systems that return correct predictions earlier in the list.

4 Experiments

4.1 Experiment Setting & Baselines

In this paper, we consider two mainstream approaches to solving the LVMR problem: retrieval-based methods and generation-based methods. Retrieval-based methods rely on chunking, where the entire video is divided into multiple independent segments. By computing the similarity between each segment and the question, these methods rank all candidate intervals and identify the most relevant ones. Such methods are often used as the retrieval component in RAG-based LVU models [31, 1, 18], where a small set of relevant video clips is first retrieved based on the question and then passed to downstream MLLMs to complete the LVU task. The second category is generation-based methods, which aim to directly test the temporal reasoning ability of MLLMs. These methods feed both the question and the video into the MLLM, prompting it to directly output a list of time intervals containing the answer in an end-to-end fashion. This evaluation of temporal grounding capability has been highlighted in several state-of-the-art MLLMs [2, 23]. In the main experiment, we set the IoU threshold to 0.3. Additionally, results with different IoU thresholds are provided in Appendix A.3. Below, we detail the settings and baseline models for each category.

 https://doi.org/10.52202/085713-0934

Table 2: Main results across different meta-tasks. #Frames indicates the number of input frames for generation-based methods and per-clip frames for retrieval-based methods.

Method	#Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
Generation-based Methods										
GPT-4o(2024-11-20) [20]	-	128	12.7	12.7	21.3	22.2	20.4	21.5	18.2	18.9
Gemini-2.5-Pro [13]	-	128	20.5	22.5	31.7	33.9	35.2	36.3	29.6	31.4
TimeChat [21]	7B	96	2.6	2.6	6.7	6.7	4.4	4.4	5.9	5.9
Lita [9]	13B	100	2.6	2.6	7.2	7.2	1.8	1.8	5.6	5.6
Qwen2.5VL [2]	7B	768	4.6	3.8	12.0	12.2	4.3	4.2	8.1	8.0
InternVL3 [39]	8B	96	3.9	3.5	7.8	8.5	4.1	4.1	5.9	6.1
Eagle2.5 [3]	8B	256	9.3	9.2	9.3	9.4	7.2	7.4	8.7	8.7
VideoChat-Flash [16]	7B	256	2.9	3.1	9.4	9.4	7.2	7.2	7.3	7.4
Video-LLaMA3 [32]	7B	256	11.1	9.9	20.9	19.0	12.8	11.7	16.4	14.9
InternVL3 [39]	38B	96	11.1	10.5	20.8	21.2	11.3	11.5	15.8	16.0
LLaVA-Video [35]	72B	96	3.6	3.5	8.6	9.8	4.6	5.6	6.3	7.2
Qwen2.5VL [2]	72B	768	13.6	13.0	21.9	21.8	12.2	11.9	17.2	16.9
Retrieval-based Methods										
E5V [10]	8.4B	1	13.1	19.5	14.5	20.7	14.9	19.8	14.3	20.1
UniIR [28]	428M	1	14.9	19.4	11.5	17.9	8.2	13.9	11.2	16.9
MM-Ret [36]	148M	1	14.2	17.9	13.6	19.4	9.7	15.4	12.4	17.7
CoVR [24]	588M	15	9.8	15.4	13.7	19.9	14.4	18.9	13.0	18.5
LanguageBind [38]	428M	8	16.2	24.6	21.4	29.4	15.5	21.0	18.2	25.4
InternVideo2 [27]	1B	8	16.8	24.5	23.5	30.9	17.0	22.7	19.7	26.6

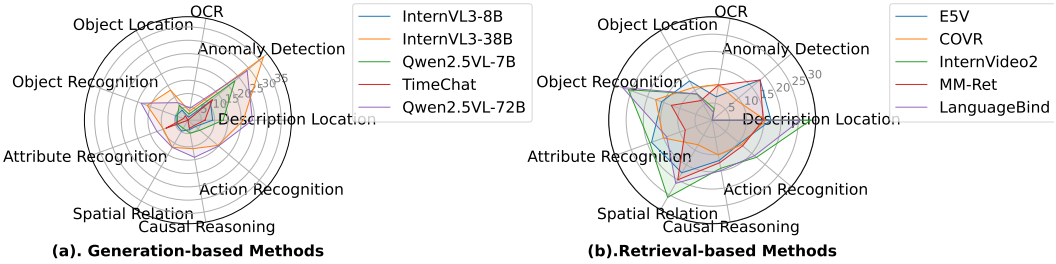


Figure 4: Sub-task performance of different retrieval-based methods and generation-based methods.

Retrieval-based Methods. We divide each video into fixed 10-second chunks without tailoring the segmentation to any specific model. The IoU threshold is set to 0.3 in the main results, with additional thresholds reported in the appendix. We compare mainstream retrieval-based methods, including dual-encoder models such as InternVideo2 [27] and LanguageBind [38]. Compositional retrieval methods such as COVR [24] and MM-RET [36] generate multi-modal embeddings by combining vision and text inputs. Recent approaches like E5V [10] and VLM2VEC [12] utilize MLLMs to embed arbitrary combinations of modalities. All these models first generate query embeddings and embeddings for pre-segmented video clips, compute similarity scores between them, rank the video clips based on these scores, and return the top-k clips as the predicted moments. Details on how each model generates embeddings are provided in the Appendix A.1.1.

Generation-based Methods. For the experimental setup, we input the video (either uniformly down-sampled frames or raw mp4) into the model and provide it with the total video duration as well as the timestamp corresponding to each frame. We then prompt the model to output a list of one or more time intervals containing the answer. The number of frames follows the optimal settings officially recommended by each model, and the time instruction format also follows the official input specifications (details for each model can be found in the Appendix A.1.2). Additionally, we provide the ablation study of different frame numbers in the Appendix A.2. We adopt the same IoU-based evaluation as used for the retrieval-based methods to enable a fair comparison. For baselines, we evaluate a wide range of state-of-the-art MLLMs, including models specifically designed for moment retrieval (e.g., TimeChat [21] and Lita [9]), leading open-source multi-modal models such as LLaVA-Video [35], InternVL3 [39], and Qwen2.5VL [2], as well as long-video-oriented MLLMs including VideoLLaMA3 [32], Eagle2.5 [3] and VideoChat-Flash [16]. Additionally, we evaluate powerful closed-source models GPT-4o (2024-11-20) [20] and Gemini-2.5-Pro [13] to provide a comprehensive assessment.

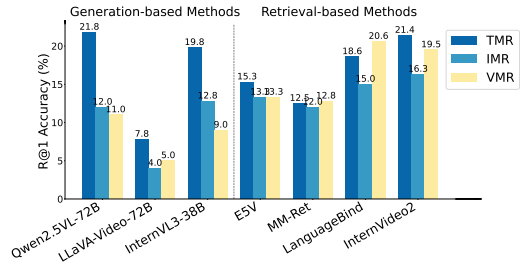


Figure 5: Evaluation results w.r.t. query modalities: TMR (text-only), IMR (image-conditioned), and VMR (video-conditioned) moment retrieval.

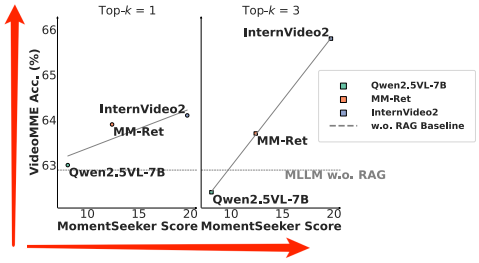


Figure 6: Comparison of MomentSeeker and LVU performance.

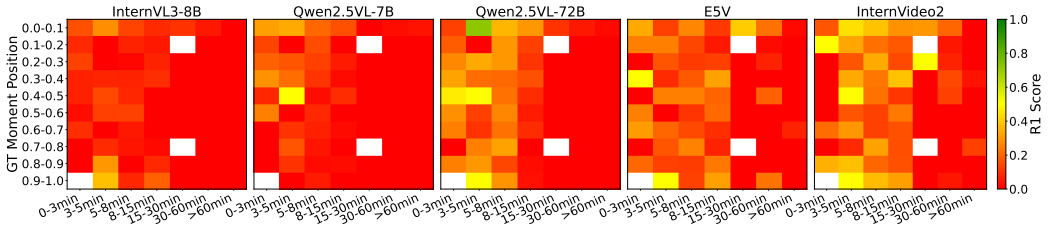


Figure 7: Accuracy of retrieval methods $w.r.t$ ground-truth interval position and video length. Generation methods better predict start/end intervals, while retrieval methods are less affected by interval position. All models decline as video duration grows.

4.2 Main Results

Table 2 and Figure 4 shows the overall results. Results indicate that despite advances in video understanding, **LVMR remains challenging for current models.** For example, InternVL3-8B achieves only 5.9 R@1 and 6.1 mAP@5, while even large-scale models like Qwen2.5VL-72B (handling up to 768 frames) reach just 17.2 R@1, underscoring current MLLMs' limitations in fine-grained temporal perception. Retrieval-based methods perform slightly better but still fall short. For example, the MLLM-based retriever E5V reaches 14.3 R@1. Even the state-of-the-art retriever InternVideo2 achieves only 19.7 R@1 in different settings. This subpar performance reflects that most retrievers are tuned for direct cross-modal alignment (e.g., caption-based retrieval), while our task requires deeper instruction understanding and complex multi-modal reasoning. Some representative visualization cases are provided in Appendix A.4.

4.3 Analysis

1. MomentSeeker remains highly challenging. Even advanced MLLMs such as GPT-4o and Gemini 2.5 Pro achieve relatively low scores on our benchmark, highlighting the difficulty of fine-grained temporal grounding. The full-set performance of GPT-4o remains largely consistent with prior results, while Gemini 2.5 Pro achieves the best reported numbers (29.6 R@1 and 31.4 mAP@5) yet still struggles on queries requiring precise temporal localization (e.g., counting the number of goals). This underscores the inherent complexity of MomentSeeker and its potential to reveal subtle temporal reasoning weaknesses in modern MLLMs.

2. Existing methods struggle with multi-modal queries. Figure 5 shows that most models perform worse on IMR (Image-conditioned Moment Retrieval) and VMR (Video-conditioned Moment Retrieval) than on TMR (Text-only Moment Retrieval). This gap is especially large for generation-based methods, indicating difficulty in integrating and reasoning across modalities. While current models can handle simpler cross-modal tasks (e.g., cross-modal retrieval) or less complex multi-modal understanding tasks (e.g., basic video understanding) reasonably well, they often fail to capture the deeper relationships required for complex multi-modal reasoning. This limitation hinders their robustness and generalizability in real-world scenarios.

https://doi.org/10.52202/085713-0934

3. Generation methods are position sensitive, while retrieval methods are not. Figure 7 illustrates model sensitivity to video duration and ground-truth interval positions. InternVL3-8B often predicts intervals near the start or end, and Qwen2.5VL-7B shows a strong bias toward the beginning. Qwen2.5VL-72B mitigates this issue with a more balanced prediction distribution. Retrieval models, treating all chunks equally regardless of position, are largely position-insensitive and consistently generate near-uniform predictions. Moreover, all models' performance generally declines as video duration increases: retrieval models face a larger candidate pool, complicating ranking, while generation models suffer greater information loss due to more aggressive downsampling, reducing prediction accuracy.

4. Generation-based models are significantly constrained by context length. As shown in Figure 7, Qwen2.5VL-72B achieves noticeably higher accuracy on short videos compared to strong retrievers such as E5V and InternVideo2. This is primarily because Qwen2.5VL-72B supports up to 768 input frames, which corresponds to approximately 6.4 minutes of video when sampled at 2 fps. For videos under 8 minutes, the information loss due to uniform frame downsampling is minimal. Moreover, generation-based methods can ingest the full context, which is especially crucial for many global-level tasks. As a result, they outperform chunking-based retrieval methods. We further confirm that longer temporal context consistently improves performance. For instance, Eagle2.5-8B (256-frame input) outperforms InternVL3-8B (96-frame input) by 2.77 overall, reinforcing the value of extended context for temporal reasoning. These results suggest that with sufficiently long context support, generation-based models have a much higher potential ceiling.

5. Generation methods lag behind retrieval methods, but scale helps. As shown in Table 2, most generation-based methods lag behind retrieval-based methods in moment retrieval performance. For example, the lightweight retriever MM-Ret (148M) achieves an overall score of 12.4 R@1 and 17.7 mAP@5, significantly outperforming the much larger generation model InternVL3-8B, which only achieves 5.9 R@1 and 6.1 mAP@5. However, as model size increases (e.g., InternVL3-38B), generation-based methods begin to approach the performance of retrieval-based ones. This indicates that while current MLLMs lack strong temporal reasoning, increasing model size helps close the gap.

6. MomentSeeker performance predicts downstream LVU task success. We use different retrievers (Qwen2.5VL, MM-Ret, InternVideo2) in a RAG pipeline to select top-k key moments, then feed them to the same MLLM (InternVL3-38B) for LVU tasks. Figure 6 shows a positive correlation between MomentSeeker retriever performance and downstream RAG-based LVU results, indicating our benchmark effectively predicts LVU task capability.

In summary, LVMR remains a challenging problem given the limitations in overall performance. Retrieval methods generally outperform generation-based methods, but both fall short in handling fine-grained temporal reasoning and complex multi-modal queries. While scaling up generation models and extending temporal context both help, even the latest video MLLMs still struggle on MomentSeeker, underscoring the benchmark's difficulty and diagnostic value. Notably, moment retrieval quality strongly correlates with LVU task performance, highlighting MomentSeeker's value as an intermediate benchmark.

5 Conclusion

We present MomentSeeker, a new benchmark designed to address the challenge of accurately locating key moments in LVU for MLLMs. MomentSeeker covers a wide range of real-world tasks and supports various query types: text-only, image-conditioned, and video-conditioned. Extensive experiments on both generation-based and retrieval-based methods reveal significant challenges in accuracy and efficiency, despite improvements brought by recent MLLMs and task-specific fine-tuning. We hope MomentSeeker will serve as a valuable resource to advance the development of more accurate and efficient LVU systems. We discuss the limitations and future work in Appendix B.

Acknowledgments

This work was supported by Beijing Natural Science Foundation No. L233008 and National Natural Science Foundation of China No. 62272467. The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

References

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

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[3] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, Tyler Poon, Max Ehrlich, Tuomas Rintamaki, Tyler Poon, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models, 2025.

[4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024.

[5] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

[7] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024.

[8] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language, 2017.

 https://doi.org/10.52202/085713-0934

[9] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant, 2024.

[10] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

[11] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014.

[12] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2025.

[13] Koray Kavukcuoglu. Gemini 2.5: Our most intelligent ai model, March 2025. Google DeepMind Blog.

[14] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[15] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.

[16] 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. Videochat-flash: Hierarchical compression for long-context video modeling, 2025.

[17] Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences, 2024.

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

[19] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023.

[20] OpenAI. Gpt-4 technical report, 2023.

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

[22] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos, 2019.

[23] Seed Team. Seed1.5-vl technical report, 2025.

[24] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5270–5279, 2024.

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

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

[27] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2025.

[28] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer, 2025.

[29] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[30] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. *arXiv preprint arXiv:2504.02259*, 2025.

[31] Huaying Yuan, Zheng Liu, Minhao Qin, Hongjin Qian, Y Shu, Zhicheng Dou, and Ji-Rong Wen. Memory-enhanced retrieval augmentation for long video understanding. *arXiv preprint arXiv:2503.09149*, 2025.

[32] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025.

[33] Gengyuan Zhang, Mang Ling Ada Fok, Jialu Ma, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. Localizing events in videos with multimodal queries, 2024.

[34] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[35] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024.

[36] Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*, 2024.

[37] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025.

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

[39] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

 https://doi.org/10.52202/085713-0934

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Yes

   Justification: We theoretically provide detailed comparison between our proposed benchmark and related benchmarks in Table 1 and explain the differences thoroughly in Section 2, which highlights the novelty of our benchmark. In addition, we conduct extensive experiments in Section 4 to evaluate existing methods on our benchmark, demonstrating its challenging nature and the diversity of evaluation perspectives.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes

   Justification: We provide a section B to discuss the limitations of the this paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: NA

Justification: NA

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: Yes

Justification: One can reproduce our result based on the code provided in the github repository (`https://github.com/yhy-2000/MomentSeeker/tree/master`).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

 https://doi.org/10.52202/085713-0934

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: Yes

   Justification: We have open-sourced all the code and data resources for this paper at the provided link (`https://yhy-2000.github.io/MomentSeeker/`).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: NA

   Justification: All the models are tested with the zero-shot setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: NA

   Justification: All the models are tested with the zero-shot setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: Yes

   Justification: See Section A.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: Yes

   Justification: Our work complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: Yes

    Justification: See Section A.4

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/085713-0934

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: NA

    Justification: See Section A.4.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: Yes

    Justification: We apply the CC BY-NC-SA 4.0 license to restrict the use of our benchmark, which is clearly stated in the benchmark download link (`https://huggingface.co/datasets/avery00/MomentSeeker`).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: NA

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: Yes

    Justification: See Section A.1.2.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: NA

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/085713-0934

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: NA

Justification: We do not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

A.1 Evaluation Setting

We conduct all the experiments on 8×A800 GPUs, each with 80 GB of memory.

A.1.1 Retrieval-based Methods

Following previous works [12, 10], we incorporated instructions (e.g., "Represent the given question in one word." and "Represent the video in one word.") to guide the model in generating informative embeddings for queries and candidates. All baselines are reproduced using their official implementations with default settings. We evaluate the video models (LanguageBind and InternVideo2) using an input of uniformly sampled 8 frames, while COVR follows its default setting of 15 frames. For image models (E5V, VLM2Vec, UniIR and MagicLens), we use the temporally middle frame as the video input. Instruction usage remains consistent with each baseline's original configuration. For image-based baselines, we use the middle frame of the video as input. CoVR [24] follows its original configuration with 15 frames at a resolution of 384, while LanguageBind, InternVideo2 [27], and MR-Embedder each use 8 frames at a resolution of 224. For image-based models (MM-Ret, MagicLens, UniIR, VLM2VEC, E5V), we use the middle frame of videos as input.

A.1.2 Generation-based Methods

All evaluation code for the generation-based methods is sourced from the official GitHub repository. The number of input frames for each model is based on the optimal frame count referenced in the original papers for long video tasks. The prompts we used are as follows:

For GPT-4o, we randomly sampled 180 test samples as the testing set. We extracted video frames at 0.5 fps, and for videos exceeding 128 frames, we uniformly sampled 128 frames. The videos were resized so that the maximum length or width did not exceed 384.

For Lita, we use the default setting: videos are extended by 100 frames without resizing. The prompt to Lita is as follows. First, we provide the timestamps of the sampled frames to Lita, allowing it to infer the temporal positions of each frame.

Time Instruction

Video 1 lasts for :.2f seconds, and {} frames are uniformly sampled from it. These frames correspond to the following timestamps: {}. Please answer the following questions based on this video.

Next, we provide task-specific prompts based on TMR, IMR, and VMR tasks.

For the TMR task, the following prompt is used. Depending on whether the input is a query or a caption, the prompt wording varies slightly.

TMR TASK PROMPT

Identify the visual event described by the following query/caption in the video, and determine its start and end times. Format: 'The event happens in the start time - end time'. Example: The event 'person turns on a light' happens in the 24.3 - 30.4 seconds. Now, here is the textual query: {}. Please return its start and end times.

For TimeChat in the default setting, videos are uniformly sampled to 96 frames. The prompts for different input types in the TMR task are as follows:

TimeChat Query-Type TMR Task

You are given a video from the QVHighlights dataset. Identify the visual event described by the given sentence, and determine its start and end times. Format: 'The event happens in the

 https://doi.org/10.52202/085713-0934

start time - end time'. Example: The event 'person turns on a light' happens in the 24.3 - 30.4 seconds. Here is the sentence: {}. Please return its start and end times.

TimeChat Caption-Type TMR Task

Detect and report the start and end timestamps of the video segment that semantically matches the given sentence. Format: 'The event happens in the start time - end time'. Example: The event 'person turns on a light' happens in the 24.3 - 30.4 seconds. Here is the sentence: {}. Please return its start and end times.

For LLaVA-Video, InternVL3, and Qwen2.5VL, we also use the default setting. Videos are sampled to 96 frames for LLaVA-Video and InternVL3, and 768 frames for Qwen2.5VL. LLaVA-Video and Qwen2.5VL share the same time instruction format as Lita. InternVL3 includes timestamps directly before each `<image>` token. All models use the same prompt template as below:

LLaVA-Video / InternVL3 / Qwen2.5VL: TMR Task

Identify the most relevant time interval(s) in the video that match the given query or caption.
Format: [[start_1, end_1], ..., [start_n, end_n]], where $1 \leq n \leq 5$.
Examples: Single interval: [[0.2, 7.8]] Multiple intervals: [[0, 10.3], [65.4, 67.3]]
Now, here is the textual query: {}
IMPORTANT: 1. Return **only** the list of relevant intervals in Video 1. 2. Do not return more than 5 intervals.

LLaVA-Video / InternVL3 / Qwen2.5VL: IMR Task

Identify one or more time intervals in the video that match the given query paired with an image. Format: [[start_1, end_1], ..., [start_n, end_n]], where $1 \leq n \leq 5$.
Examples: Single interval: [[0.2, 7.8]] Multiple intervals: [[0, 10.3], [65.4, 67.3]]
Here is the textual query: {} and the accompanying image: Image 1.
IMPORTANT: 1. Return **only** the list of relevant intervals in Video 1. 2. Do not return more than 5 intervals.

LLaVA-Video / InternVL3 / Qwen2.5VL: VMR Task

Identify one or more time intervals in the video that match the given query paired with a reference video. Format: [[start_1, end_1], ..., [start_n, end_n]], where $1 \leq n \leq 5$.
Examples: Single interval: [[0.2, 7.8]] Multiple intervals: [[0, 10.3], [65.4, 67.3]]
Here is the textual query: {} and the accompanying video: Video 2.
IMPORTANT: 1. Return **only** the list of relevant intervals in Video 1. 2. Do not return more than 5 intervals.

Annotation Guideline

```
--- Task 1: Video Search (MR) ---

Objective: Generate a question answerable by a specific video segment and annotate the
corresponding one or more answering segments.

Scope: Both the question and answer must originate from the same video.

Steps:

1. Watch the entire video to understand its content (e.g., actions, events, objects).
2. Formulate a question:
   - Ensure the question is specific and requires one or several continuous segments as
```

```
   the answer.
   - Example questions:
     * "Where was the weighing scale?"
     * "What did I put in the trashcan?"
3. Annotate the answer segment:
   - Mark the start and end timestamps (format: [[MM:SS--MM:SS], [MM:SS--MM:SS]...]).
   - Ensure the segments fully answer the question without truncation.
   - Cover all segments that answer the question—no omissions allowed.
4. Validation: Replay the annotated segment to confirm alignment with the question.

--- Task 2: Image-Conditioned Video Search (IMR) ---

Objective: Generate a question based on a static image (from the same or a different
video) and annotate the answer segments in the target video.

Steps:
1. Select an image:
   - Same video: Choose a key frame (e.g., action initiation, critical object appearance).
   - Different video: Use a frame from another video with relevant content (e.g.,
   similar objects, scenes).
2. Formulate a question:
   - The question must directly relate to the selected image, and the answer must
   exist in the target video.
   - Example questions:
     * Same video (image shows a road with street scenery): "What's the color of
     the dog that once appeared on this road?"
     * Different video (image shows the same character in different clothing):
     "What did the man in this picture give to the person next to him?"
3. Annotate the answer segment:
   - Mark the timestamps in the target video that answer the question.
   - Ensure logical consistency between the image, question, and segment.
4. Validation: Verify that the image, question, and annotated segment are coherent.

--- Quality Requirements ---

1. Consistency: Ensure IMR questions are tightly linked to the image, and
answers are precise.
2. Cross-video logic: If using an image from another video, the question must
relate to the target video's content (avoid mixing contexts).
3. Timestamp accuracy: Annotated segments must have less than 1-second error tolerance.

Report ambiguities or edge cases to the project lead for resolution.
```

A.2 Ablation of Different Video Frame Numbers

In the main experiments, we followed the optimal frame settings officially recommended for each model. Additionally, Table 2 presents the performance of different models under various frame input conditions.We observe that varying the number of input frames within a moderate range from 64 to 128 has minimal impact on performance. For example, InternVL3-38B achieves stable overall performance with R@1 values of 15.9, 15.8, and 15.6 from 64 to 128 frames, suggesting that denser sampling in this range does not significantly improve results. But increasing the number of frames from 96 to 768 for Qwen2.5VL-7B results in a marked performance gain, from 4.5 to 8.1 in R@1. This suggests that models benefit more from denser temporal coverage, likely due to their limited ability to infer information from sparsely sampled inputs.However, such improvements result in a considerable computational cost for processing 768 frames significantly increases memory and latency.

 https://doi.org/10.52202/085713-0934

Table 3: Main results across different meta-tasks. #Frames indicates the number of input frames for generation-based methods and per-clip frames for retrieval-based methods. † denotes tested on a random subset due to high cost.

Method	#Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
Generation-based Methods										
Qwen2.5VL	7B	96	1.8	1.5	7.0	7.0	2.4	2.1	4.5	4.3
InternVL3	8B	96	3.9	3.5	7.8	8.5	4.1	4.1	5.9	6.1
InternVL3	38B	64	11.1	10.5	20.9	21.3	11.3	11.5	15.9	16.0
InternVL3	38B	96	11.1	10.5	20.8	21.2	11.3	11.5	15.8	16.0
InternVL3	38B	128	11.1	10.5	20.1	20.5	11.3	11.4	15.6	15.7
Qwen2.5VL	7B	256	2.3	2.2	6.8	7.2	2.0	2.1	4.4	4.6
Qwen2.5VL	7B	768	4.6	3.8	12.0	12.2	4.3	4.2	8.1	8.0
Qwen2.5VL	7B	64	2.6	2.1	6.8	7.0	2.2	2.2	4.5	4.5
Qwen2.5VL	7B	128	2.8	2.4	7.6	7.8	2.0	1.9	4.9	4.9

Table 4: Results of IoU=0.1.

Method	#Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
Generation-based Methods										
InternVL3	8B	96	6.4	5.8	13.8	15.0	7.1	7.1	10.3	10.7
InternVL3	38B	96	23.9	22.0	31.2	31.8	20.2	19.7	26.3	26.1
Qwen2.5VL	7B	768	9.8	8.8	22.5	22.9	9.1	9.0	15.7	15.7
TimeChat	7B	96	5.3	5.3	16.0	16.0	6.6	6.6	13.1	13.1
Lita	13B	100	11.8	11.8	18.3	18.3	9.3	9.3	15.8	15.8
Qwen2.5VL	72B	768	22.6	21.2	30.7	30.7	19.9	19.0	25.7	25.1
LLaVA-Video	72B	96	8.5	8.4	16.4	18.0	8.3	9.8	12.3	13.5
Retrieval-based Methods										
E5V	8.4B	1	21.6	28.6	23.5	33.0	26.7	33.2	24.1	32.0
COVR	588M	15	14.4	21.4	23.0	31.7	21.3	28.0	20.4	28.0
InternVideo2	1B	8	29.1	38.1	38.8	48.7	27.4	35.5	32.7	41.7
UniIR	428M	1	24.5	29.6	21.6	30.9	20.0	27.2	21.7	29.3
MM-Ret	148M	1	19.3	25.8	23.3	31.1	17.5	25.5	20.4	27.9
LanguageBind	428M	8	30.4	39.3	38.1	48.0	27.2	35.0	32.6	41.5

Table 5: Results of IoU=0.2.

Method	#Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
Generation-based Methods										
IntennVL3	8B	96	4.4	4.0	10.9	12.0	5.8	5.7	8.0	8.4
IntennVL3	38B	96	18.0	16.7	25.6	26.2	15.8	15.6	21.0	20.9
Qwen2.5VL	7B	768	7.2	6.2	16.7	17.2	6.7	6.6	11.7	11.7
TimeChat	7B	96	3.9	3.9	10.5	10.5	5.3	5.3	8.8	8.8
Lita	13B	100	7.9	7.9	10.6	10.6	4.9	4.9	9.1	9.1
Qwen2.5VL	72B	768	17.5	16.4	26.5	26.4	15.8	15.2	21.3	20.9
LLaVA-Video	72B	96	4.9	4.7	12.0	13.5	6.1	7.1	8.7	9.7
Retrieval-based Methods										
E5V	8.4B	1	19.6	25.9	20.4	28.8	20.7	26.9	20.3	27.5
COVR	588M	15	13.1	19.5	19.7	27.9	17.4	23.5	17.4	24.4
UniIR	428M	1	21.6	26.3	19.2	27.3	14.4	21.3	18.2	25.1
MM-Ret	148M	1	17.5	23.0	20.1	27.3	14.0	21.3	17.4	24.3
LanguageBind	428M	8	25.0	33.6	32.3	42.6	21.1	28.2	26.8	35.6

A.3 Ablation of Different IoU Thresholds

We provide experimental results under different IoU settings (0.1,0.2,0.4,0.5) in Table 4, Table 5, Table 6, and Table 7. It can be observed that stricter IoU thresholds (e.g., 0.5) lead to a drop in the accuracy of all models. However, the experimental conclusions drawn in Section 4 still hold under different IoU settings.

Table 6: Results of IoU=0.4.

Method	#Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
Generation-based Methods										
InternVL3	8B	96	1.3	1.2	5.8	6.3	3.2	3.2	4.1	4.3
InternVL3	38B	96	7.7	7.5	16.1	16.3	8.2	8.1	11.9	12.0
Qwen2.5VL	7B	768	3.6	2.9	8.9	9.0	2.8	2.9	5.9	5.8
TimeChat	7B	96	1.3	1.3	4.9	4.9	3.5	3.5	4.3	4.3
Lita	13B	100	1.3	1.3	4.0	4.0	1.3	1.3	3.2	3.2
Qwen2.5VL	72B	768	10.0	9.4	15.2	15.2	9.1	8.6	12.3	12.0
LLaVA-Video	72B	96	2.8	2.6	5.0	6.1	4.1	4.6	4.3	4.9
Retrieval-based Methods										
E5V	8.4B	1	6.2	10.5	10.9	15.9	10.6	15.1	9.7	14.3
COVR	588M	15	6.2	9.2	10.6	15.8	10.5	14.4	9.5	13.7
InternVideo2	1B	8	10.6	15.5	18.3	23.9	14.4	18.6	15.1	20.1
UniIR	428M	1	8.8	11.0	8.3	13.6	5.6	9.9	7.5	11.8
MM-Ret	148M	1	6.4	9.4	10.6	15.0	7.6	12.2	8.6	12.7
LanguageBind	428M	8	10.3	15.3	16.7	23.4	11.8	16.4	13.5	19.1

Table 7: Results of IoU=0.5.

Method	#Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
Generation-based Methods										
InternVL3	8B	96	1.0	0.9	4.0	4.2	2.8	2.8	3.0	3.1
InternVL3	38B	96	4.6	4.4	12.6	12.8	6.3	6.1	9.0	9.0
Qwen2.5VL	7B	768	3.1	2.4	6.7	6.6	1.9	1.9	4.4	4.3
TimeChat	7B	96	1.3	1.3	2.9	2.9	1.8	1.8	2.5	2.5
Lita	13B	100	0.0	0.0	2.1	2.1	1.3	1.3	1.8	1.8
Qwen2.5VL	72B	768	6.9	6.2	11.8	11.8	5.9	5.5	9.0	8.7
LLaVA-Video	72B	96	1.3	1.1	2.6	3.5	3.2	3.6	2.5	3.0
Retrieval-based Methods										
E5V	8.4B	1	2.8	5.3	7.8	11.7	6.9	10.3	6.3	9.7
COVR	588M	15	3.9	6.1	7.7	11.6	7.3	10.1	6.6	9.8
InternVideo2	1B	8	5.7	8.5	14.2	18.2	10.8	13.0	11.0	14.1
UniIR	428M	1	4.6	5.8	6.6	10.3	3.2	6.4	5.0	7.9
MM-Ret	148M	1	3.1	4.8	6.8	10.3	5.0	7.8	5.3	8.1
LanguageBind	428M	8	4.4	8.4	12.3	17.0	7.8	10.5	8.9	12.8

A.4 Visualization Results

We provide case study of generation-based methods (GPT-4o, Qwen2.5VL-72B), and retrieval-based model InternVideo2 in Figure 8, Figure 9 and Figure 10.

Broader Impacts

All the videos in our benchmark are carefully extracted from publicly available datasets [19, 37, 26, 22], which have undergone rigorous screening by the original teams to remove harmful content. Despite our best efforts, we acknowledge that these videos may not be entirely comprehensive or free from omissions. Furthermore, we strongly discourage the use of MR-Embedder models for encoding or retrieving sensitive content.

B Limitations & Future Works

Limitations. This paper focuses on moment retrieval in long videos and presents a comprehensive benchmark evaluation. We conduct extensive experiments from multiple perspectives, including task types, query modalities, video lengths, and moment distributions. While we tested a proprietary MLLM (GPT-4o) in our study, large-scale evaluation on long videos with closed-source models remains prohibitively expensive. For instance, processing 384 frames at 512 resolution with GPT-4o costs approximately $2 per query, amounting to around $3,600 for our dataset of 1,800 queries. Due to this high cost, we did not explore other proprietary MLLMs in this work and instead focused

https://doi.org/10.52202/085713-0934

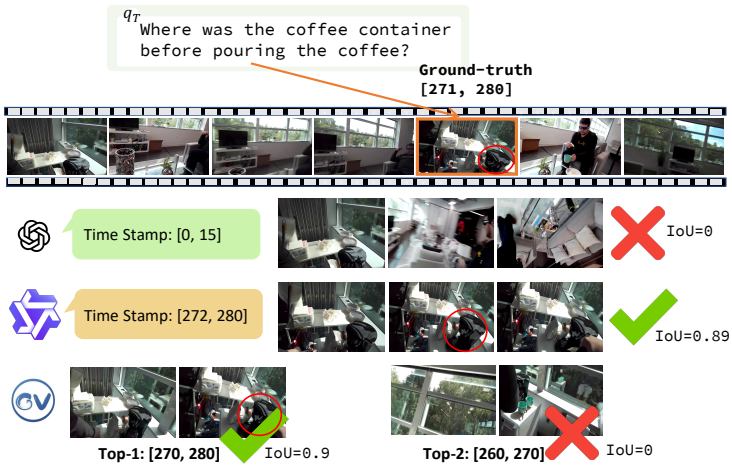


Figure 8: Case Study of generation-based methods (GPT-4o, Qwen2.5VL-72B), and retrieval-based model InternVideo2.

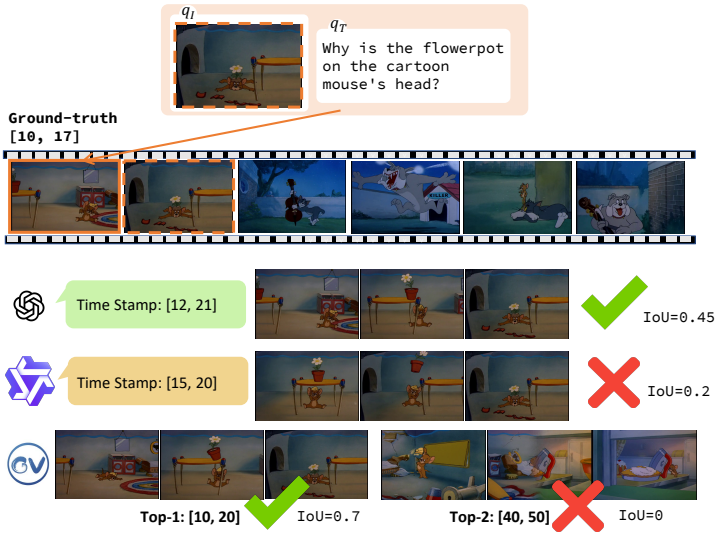


Figure 9: Case Study of generation-based methods (GPT-4o, Qwen2.5VL-72B), and retrieval-based model InternVideo2.

primarily on open-source models. In addition, although we provide extensive empirical evaluations of existing methods, we do not delve into model-specific architectural choices, such as the design of temporal embeddings or structural adaptations to enhance temporal reasoning in MLLMs.

Future work suggestions. Two promising directions emerge for future long-video understanding models. First, current MLLMs show performance drops on fine-grained temporal localization [4, 37]; thus, efforts to enhance their temporal awareness could be beneficial. Second, integrating lightweight retrievers into RAG frameworks may improve efficiency by focusing on key video segments. By advancing both directions together, we may improve the accuracy of existing LVU models on downstream tasks by enhancing their temporal awareness.

https://doi.org/10.52202/085713-0934

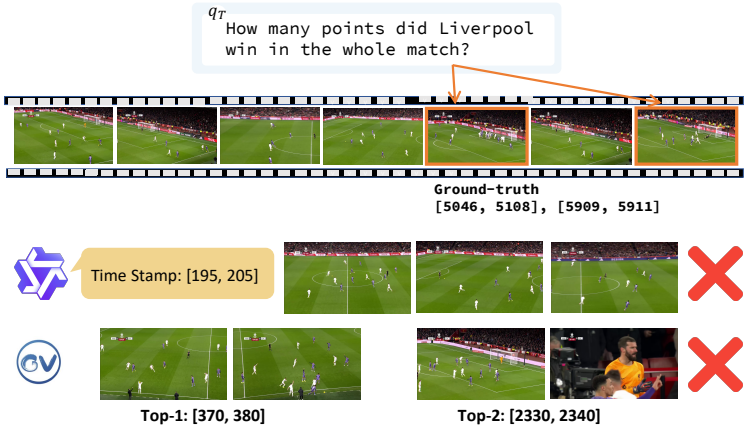


Figure 10: Case Study of generation-based methods (GPT-4o, Qwen2.5VL-7B), and retrieval-based model InternVideo2.

https://doi.org/10.52202/085713-0934