

VIDEODEEPRESEARCH: LONG VIDEO UNDERSTANDING WITH AGENTIC TOOL USING

Huaying Yuan¹, Zheng Liu^{2*}, Junjie Zhou³, Ji-Rong Wen^{1*}, Zhicheng Dou^{1*}

¹Gaoling School of Artificial Intelligence, Renmin University of China

²Beijing Academy of Artificial Intelligence, ³Beijing University of Posts and Telecommunications
<https://github.com/yhy-2000/VideoDeepResearch>

ABSTRACT

Long video understanding (LVU) presents a significant challenge for current multi-modal large language models (MLLMs) due to the task’s inherent complexity and context window constraint. It is widely assumed that addressing LVU tasks requires foundation MLLMs with extended context windows, strong visual perception capabilities, and proficient domain expertise. In this work, we challenge this common belief by introducing **VideoDeepResearch**, a novel agentic framework for long video understanding. Our approach relies solely on a text-only large reasoning model (LRM) combined with a modular multi-modal toolkit, including multimodal retrievers and visual perceivers, all of which are readily available in practice. For each LVU task, the system formulates a problem-solving strategy through reasoning, while selectively accessing and utilizing essential video content via tool using. We conduct extensive experiments on popular LVU benchmarks, including MLVU, Video-MME, and LVBench. Our results demonstrate that VideoDeepResearch achieves substantial improvements over existing MLLM baselines, surpassing the previous state-of-the-art by 9.6%, 6.6%, and 3.9% on MLVU (test), LVBench, and LongVideoBench, respectively. These findings highlight the promise of agentic systems in overcoming key challenges in LVU problems.

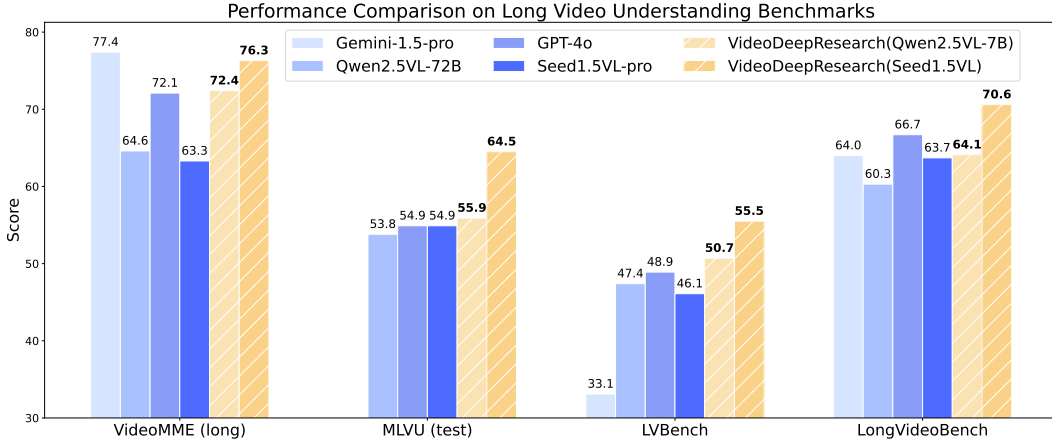


Figure 1: In our experiment, VideoDeepResearch powered by DeepSeek-R1-0528 (reasoning module) and Seed1.5VL-pro/Qwen2.5VL-7B (visual module) outperforms state-of-the-art MLLMs, including GPT-4o, Gemini-1.5-Pro and Qwen2.5-VL-72B, across popular LVU benchmarks.

1 INTRODUCTION

Long video understanding (LVU) is crucial for a wide variety of real-world applications, such as movie analysis, video surveillance, autonomous driving, and embodied AI (Fu et al., 2024; Zhou et al.,

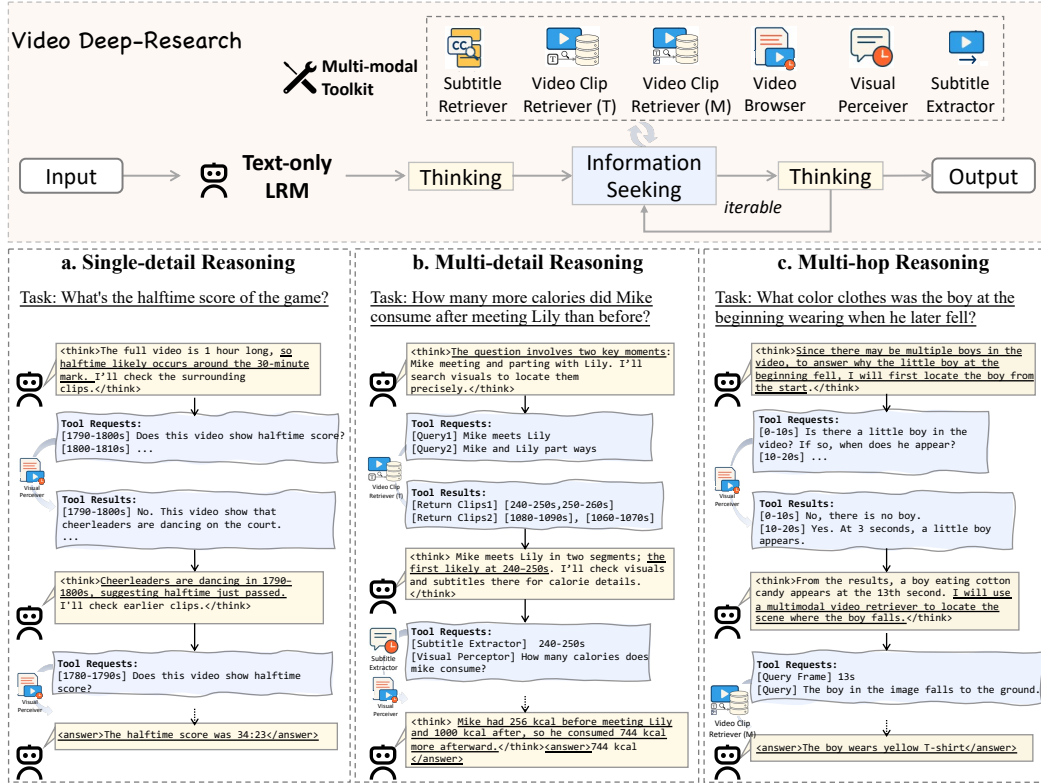


Figure 2: Framework of Video-DeepResearch.

2024; Chandrasegaran et al., 2024). While recent progress in multi-modal large language models (MLLMs) has been remarkable (OpenAI, 2024; Reid et al., 2024), significant challenges remain in addressing LVU tasks effectively. As video length increases, the complexity of understanding and reasoning over the content grows accordingly. Meanwhile, longer videos impose significantly greater computational demands and often exceed the context window limitations of current MLLMs. Although retrieval-augmented generation (RAG) (Yuan et al., 2025; Luo et al., 2024; Ataallah et al., 2024) alleviates this problem by selectively incorporating a fraction of video content, conventional approaches are typically constrained to narrow task categories, like video question-answering. These methods struggle to generalize to broader scenarios that require complex reasoning and dynamic interaction with the video content.

Three key factors are essential for effectively addressing LVU problems: fine-grained visual perception, proficient domain expertise, and the ability to handle long visual contexts. While it may seem straightforward to unify all of these capabilities into a powerful MLLM, we argue that an agentic system, equipped with specialized tools for each required functionality, provides a sufficient and often more practical alternative. With this motivation, we propose **VideoDeepResearch**, a novel framework for long video understanding built upon agentic tool using. Our framework consists of the following components.

- We employ a *text-only large reasoning model* (LRM) as the cognitive core of VideoDeepResearch. Leveraging its inherent reasoning capabilities, the LRM plans the problem-solving workflow and invokes appropriate tools to acquire relevant video information.
- We incorporate *multi-modal tools*, specifically video retrievers and visual perceivers, enabling the LRM to access arbitrary video content and convert visual information into textual form.

Despite its simplicity, VideoDeepResearch offers several advantages over the brute-force approach of directly using MLLMs. First, recent LRMs, such as OpenAI’s O1/O3 and DeepSeek’s R1, exhibit strong long-form reasoning and tool-using capabilities. When paired with well-designed

video retrievers and visual perceivers, these systems are fully capable of extracting and utilizing the necessary information from videos to solve the LVU problems. Since all these components are readily available, the agentic framework provides a more practical and comparably effective solution for most LVU tasks. Second, unlike brute-force methods that attempt to load the entire video into the MLLM’s context window, the agentic approach focuses only on the relevant segments, making it more efficient and resource-friendly. Third, the agentic framework can be seamlessly integrated with optimization strategies such as reinforcement learning and test-time scaling, which may further enhance its performance over time.

We perform comprehensive evaluations based on a variety of popular LVU benchmarks, including Video-MME (Fu et al., 2024), MLVU (Zhou et al., 2024), LVBench (Wang et al., 2024a), and LongVideoBench (Wu et al., 2024). In our experiment, VideoDeepResearch is able to outperform strong MLLMs, including GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Reid et al., 2024), and QwenVL-72B (Bai et al., 2025), and traditional RAG baselines. Meanwhile, it also demonstrates higher cost-effectiveness against the existing methods. Our prototyping system is publicly released¹ to facilitate the further research in this field.

2 METHODOLOGY

2.1 PRELIMINARY

In Long-form Video Understanding (LVU) tasks, the raw video input typically consists of a large number of frames, denoted as $\mathcal{X} = \{X_1, \dots, X_n\}$, where n represents the total number of natural frames. For hour-long videos with standard frame rates (typically 25 fps), a single video contains approximately 90,000 natural frames. However, current state-of-the-art MLLMs (OpenAI, 2024; Bai et al., 2025; Team and etc, 2024) can only process input contexts of up to 1,000 frames. Consequently, these models typically employ uniform downsampling to extract a subset $\mathcal{X}' = \{X_1, \dots, X_m\}$ from the original input frames for processing, where $m \ll n$. This approach inevitably leads to information loss. When the content relevant to a query is not captured within the sampled frames, the model becomes nearly incapable of providing correct answers. Conversely, even if models could densely process the entire video, the extensive computational overhead of processing all frames would be prohibitive.

Retrieval-augmented generation (RAG) (Luo et al., 2024; Yuan et al., 2025; Ataallah et al., 2024) is an efficient strategy that leverages a retriever to retrieve relevant frames $\mathcal{X}'' = \{X_1, \dots, X_k\}$ according to the question before passing them to the MLLMs. This strategy can effectively solve simple tasks that explicitly define the information need in the question. However, for more general or complex questions, this strategy becomes inadequate.

Different from the above strategies, we propose to solve LVU tasks by progressively reasoning over informative video segments with agentic tool using. Rather than identifying all relevant frames at once, our approach progressively expands a subset of frames $\tilde{\mathcal{X}} = \{X_{i_1}, \dots, X_{i_s}\}$ through iterative reasoning steps $\mathcal{R} = \{R_1, R_2, \dots, R_k\}$, where each reasoning step R_j guides the selection of additional frames based on the current understanding. This reasoning process ensures that $s \ll n$ while dynamically adapting frame selection strategies to different query requirements, achieving superior performance in both effectiveness and efficiency.

2.2 OVERVIEW

Long videos with numerous input frames present both computational and complexity challenges. In this paper, we propose an agentic framework that enables progressive reasoning over long videos through iterative, selective analysis with multi-modal tools. The overall framework is shown in Figure 2. Given a task input, the text-only LRM \mathcal{M} first analyzes the task to identify the necessary information and selects appropriate tools from a multi-modal toolkit. It then issues corresponding tool calls. Based on the returned results, \mathcal{M} determines whether the retrieved information suffices. If not, it continues its reasoning and retrieval process iteratively; otherwise, it produces the final answer and terminates.

¹<https://github.com/yhy-2000/VideoDeepResearch>

Algorithm 1 VideoDeepResearch Inference Process**Require:** Question q , Video V , Subtitle set \mathcal{C}_s , Task instruction \mathcal{I} **Ensure:** Answer a

```

1: Segment video into short clip set:  $\mathcal{C}_v = \text{ExtractClips}(V)$ 
2: Init context  $\mathcal{H} = \{\mathcal{I}, q\}$ ,  $a = \text{None}$ 
3: while  $a = \text{None}$  do
4:   (thought, action)  $\leftarrow \mathcal{M}(\mathcal{I})$ 
5:   if action = “answer” then
6:      $a \leftarrow \text{ParseAnswer}(\text{thought})$ ; break
7:   else
8:     Initialize tool results  $\mathcal{O}_{\text{tool}} = \emptyset$ 
9:     Parse tool calls:  $\{\text{tool}_i\}_{i=1}^n \leftarrow \text{ParseTools}(\text{thought})$ 
10:    for  $i = 1$  to  $n$  do
11:      Run  $\text{tool}_i$  and get result  $\mathcal{T}_i$ 
12:      Merge tool result:  $\mathcal{O}_{\text{tool}} \leftarrow \mathcal{T}_i \cup \mathcal{O}_{\text{tool}}$ 
13:    end for
14:    Update context:  $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{O}_{\text{tool}}$ 
15:  end if
16: end while
17: return  $a$ 

```

2.3 TOOL DEFINITION

We present a straightforward yet comprehensive toolkit comprising five specialized tools.

a). Video Clip Retriever. Long videos often contain complex content with substantial irrelevant or redundant information, making it challenging to identify segments relevant to a given query. To address this, we design a video clip retriever to efficiently localize the most informative portions. Specifically, we pre-segment the long video into fixed-length video clips, forming a candidate set $\mathcal{C}_v = c_1, c_2, \dots, cn/d$, where d denotes the duration of each clip. If necessary, the LRM formulates a search query q and invokes the video clip retriever \mathcal{R}_v to identify the top- k most relevant clips:

$$\mathcal{V}_{\text{ret}} = \mathcal{R}_v(q \mid \mathcal{C}_v). \quad (1)$$

The query q can be either a text-only input (e.g., “a boy reading a book”), suitable for standard temporal localization tasks, or a multi-modal input (e.g., Image 1 + “the boy in Image 1 is reading a book”), which supports grounding based on a reference scene. By retrieving only the most relevant segments, this module enables efficient and focused video analysis, significantly reducing computational overhead compared to full-video processing or brute-force down-sampling strategies.

b). Subtitle Retriever. Similar to the video clip retriever, the subtitle retriever identifies relevant video clips by leveraging subtitles. It targets audio-centric queries, such as “What did the man say in the car?”. Given the subtitle set \mathcal{C}_s , the subtitle retriever \mathcal{R}_s selects the top- k relevant subtitles given a textual query q :

$$\mathcal{S}_{\text{ret}} = \mathcal{R}_s(q \mid \mathcal{C}_s). \quad (2)$$

Each retrieved subtitle \mathcal{S}_{ret} is inherently associated with timestamps, enabling seamless localization of the corresponding video clips.

c). Visual Perceiver. Since LRM itself lacks visual perception capabilities, we need to equip it with an “eye” to help it access and interpret visual information. That is why we introduce the visual perceiver. Given a short video segment with timestamp $\mathcal{C}_{[t_0, t_1]}$ and a question q , the visual perceiver \mathcal{P}_c is expected to produce an answer T_A :

$$T_A = \mathcal{P}_c(q \mid \mathcal{C}_{[t_0, t_1]}). \quad (3)$$

Importantly, our visual perceiver differs from conventional applications of multimodal large language models (MLLMs) in general visual question answering tasks. Here, the visual perceiver operates on local context, using a small number of precisely located frames. This design is suitable for fine-grained understanding of short video clips once the relevant segment has been identified.

d). Subtitle Extractor. This tool facilitates subtitle localization based on explicit temporal references (e.g., “What did the host mention in the first minute?”). The LRM can navigate to subtitles within a

specified timestamp range $[t_0, t_1]$ as follows:

$$\mathcal{S}_t = \mathcal{E}_s([t_0, t_1]|\mathcal{C}_s), \quad (4)$$

where \mathcal{E}_s is a simple mapping function that associates the time interval $[t_0, t_1]$ with the corresponding entries in the subtitle set \mathcal{C}_s .

e). Video Browser. For holistic tasks that are not tied to specific details (e.g., “What is the theme of this video?”), we design a video browser that simulates human-style rapid browsing to gain a general understanding of the video. The video browser \mathcal{P}_b operates on a coarsely sampled set of frames from the entire video and directly answers general or high-level questions:

$$T_A = \mathcal{P}_b(q) \quad (5)$$

The video browser is implemented using a MLLM. Since MLLMs can also be directly applied to LVU tasks, we separately evaluate the performance of using only the MLLM with the same settings to isolate its impact on our experiments. Please refer to Table 1 for details.

Together, these tools support both visual and textual modalities, forming a comprehensive toolkit that spans from local to global understanding and from single-hop to multi-hop reasoning. By properly leveraging these tools, our framework can effectively address a wide range of general video tasks. Moreover, by incorporating a retriever to locate minimal yet valuable video segments for each task, this framework boosts computational efficiency without sacrificing performance.

2.4 VIDEODEEPRESEARCH INFERENCE PROCESS

The VideoDeepResearch inference process is illustrated in Algorithm 1. Given a question q , input video V , subtitle set \mathcal{C}_s (optional), and task instruction \mathcal{I} , the algorithm first segments the video into manageable clips \mathcal{C}_v and initializes the context with the instruction and question. The core inference loop operates through a thought-action paradigm, where the LRM \mathcal{M} generates reasoning thoughts and determines appropriate actions. When the model decides to provide an answer, it parses the final response and terminates. Otherwise, it executes tool calls by parsing multiple tools from the thought, running each tool to gather results, and merging the outputs into the evolving context \mathcal{H} . This iterative process continues until a satisfactory answer is reached, enabling the system to dynamically gather and synthesize information from various sources to address complex video-based queries.

3 EXPERIMENTS

3.1 IMPLEMENTATION DETAILS

We pre-segment each video into non-overlap 10-second video clips. We use LanguageBind-large (428M) (Zhu et al., 2023) as video and subtitle retrievers. Our text-only LRM is DeepSeek-r1-0528 (Team, 2025). The visual perceiver of VideoDeepResearch (Qwen2.5VL-7B) is based on Qwen2.5VL (Bai et al., 2025), supporting up to 32 input frames and a maximum of 24,576 tokens. For VideoDeepResearch (Seed1.5VL), the visual perceiver is Seed1.5VL-Pro (Guo and etc, 2025), also supporting up to 32 input frames, with each frame having a resolution no larger than 748×400. In terms of input modalities, LongVideoBench and VideoMME include subtitles and relevant tools, while MLVU and LVBench do not contain subtitles and corresponding tools.

3.2 BASELINES

We compare our method against both open-source state-of-the-art MLLMs (Bai et al., 2025) and proprietary models (OpenAI, 2024; Team and etc, 2024; Guo and etc, 2025). Besides, we also compare with long-video-oriented MLLMs, including LongVILA (Xue et al., 2024) and Video-XL (Shu et al., 2024). Due to the high cost of querying GPT-4o and Gemini 1.5 Pro, we report their officially published results. For Seed1.5VL-Pro, we reproduce the results by setting the maximum number of input frames to 32. This configuration ensures a fair comparison by aligning with our visual use of Seed1.5VL and maintaining consistent visual perception models and frame constraints across all frameworks. Additionally, we compare Video-R1 (Feng et al., 2025), which is a recent work that aims to enhance video reasoning capabilities through reinforcement learning. We replicate

Table 1: Evaluation results on MLVU (test), LVBench, VideoMME (long/w.subtitle), and LongVideoBench. † indicates reproduced results. #Frame refers to the frame constraints for MLLMs and our framework’s visual perceiver and video browser.

Model	#Frame	MLVU	LVBench	VideoMME(L)	LongVideoBench	Avg
LongVILA-7B† (Xue et al., 2024)	256	<u>49.0</u>	-	52.1	57.7	-
Video-XL-7B (Shu et al., 2024)	256	45.5	-	54.9	50.7	-
Video-R1-7B† (Feng et al., 2025)	64	-	-	52.2	-	-
InternVL3-8B (Zhu and etc, 2025)	64	-	-	-	<u>58.8</u>	-
Qwen2.5VL-7B† (Bai et al., 2025)	128	47.4	44.8	57.6	58.0	52.0
+ RAG† (Bai et al., 2025)	128	48.9	<u>47.2</u>	<u>58.2</u>	58.4	<u>53.2</u>
VideoDeepResearch(Qwen2.5VL-7B)	32	55.9	50.7	72.4	64.1	60.8
GPT-4o (OpenAI, 2024)	384 ¹	54.9	48.9	72.1	66.7	<u>60.6</u>
Gemini-1.5-pro (Team and etc, 2024)	256 ²	-	33.1	77.4	64.0	-
Seed1.5VL-pro† (Guo and etc, 2025)	32	54.9	46.1	63.3	63.7	57.0
Qwen2.5VL-72B† (Bai et al., 2025)	128	53.8	47.4	64.6	60.3	56.5
+ RAG† (Bai et al., 2025)	128	<u>59.0</u>	<u>52.2</u>	65.8	62.9	60.0
VideoDeepResearch(Seed1.5VL-pro)	32	64.5	55.5	<u>76.3</u>	70.6	66.7

the results using the official model weights and code, adopting the optimal setting of 64 frames. In addition, we reproduce RAG-enhanced versions of Qwen2.5VL-7B and Qwen2.5VL-72B. The RAG framework adopts the same video clip retriever and subtitle retriever to retrieve the top- k relevant video clips and subtitles. To better implement retrieval-augmented generation, we follow the strategy proposed in (Yuan et al., 2025), allocating the retrieved key frames and globally downsampled frames (which preserve global context) according to a fixed ratio α . In our experiments, we set $\alpha = 0.4$ and $k = 4$.

3.3 MAIN RESULT

As demonstrated in Table 1, our VideoDeepResearch framework exhibits remarkable advantages across multiple dimensions.

First, VideoDeepResearch achieves superior performance compared to existing methods. When built upon Qwen2.5VL-7B, our framework attains an average score of 60.8 across four challenging benchmarks, significantly outperforming the base model, which scores 52.0. Compared to other specialized long video understanding methods, it demonstrates significant improvements. For example, VideoDeepResearch achieves 55.9 compared to LongVILA-7B’s 49.0, an improvement of 6.9 points, and Video-XL-7B’s 45.5, an improvement of 10.4 points on MLVU. On VideoMME-L, it scores 72.4 compared to LongVILA-7B’s 52.1, an improvement of 20.3 points, and Video-XL-7B’s 54.9, an improvement of 17.5 points.

Second, our framework overcomes the fundamental limitation of MLLM context window constraints by processing only valuable video segments with a maximum of 32 frames. While competing methods are bounded by their context windows, typically 128 to 256 frames, VideoDeepResearch can theoretically handle arbitrarily long videos by selectively analyzing short but informative clips, enabling comprehensive video understanding without sacrificing detail. This architectural advantage becomes particularly evident when applying more accurate visual perceivers. VideoDeepResearch with Seed1.5VL-pro achieves an average score of 66.7, outperforming GPT-4o, which scores 60.6, despite GPT-4o’s substantially larger 384-frame context window. The consistent performance gains across diverse video understanding benchmarks demonstrate the generalizability and robustness of our approach, establishing VideoDeepResearch as a highly effective and efficient solution for long video understanding.

Task Analysis. To understand how VideoDeepResearch performs, we conduct a detailed analysis of various models and frameworks on the MLVU test set. The results are illustrated in Figure 3. Specifically, we compare our proposed framework VideoDeepResearch with the strong MLLM Qwen2.5VL-72B, as well as a basic RAG framework (Qwen2.5VL-72B + RAG). In addition,

¹We report the official evaluation settings of GPT-4o on each benchmark, which use 384 / 60 / 256 / 384 maximum frames, respectively.

²We report the official evaluation settings of Gemini 1.5 Pro on each benchmark, which use - / 3600 / 1 fps / 256 maximum frames, respectively.

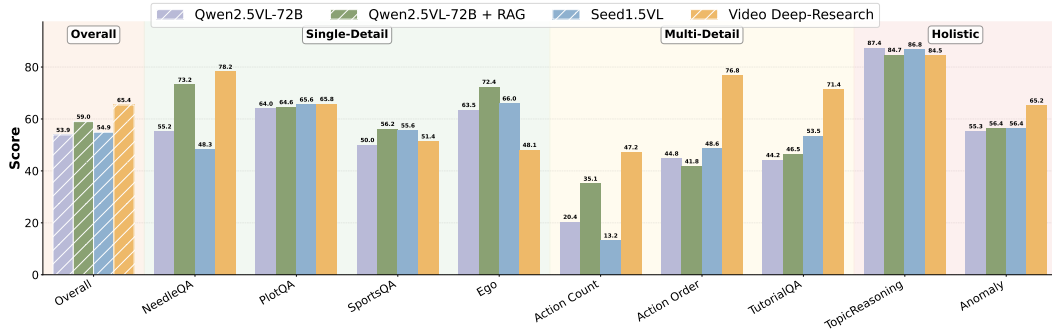


Figure 3: Task performance analysis on MLVU test set.

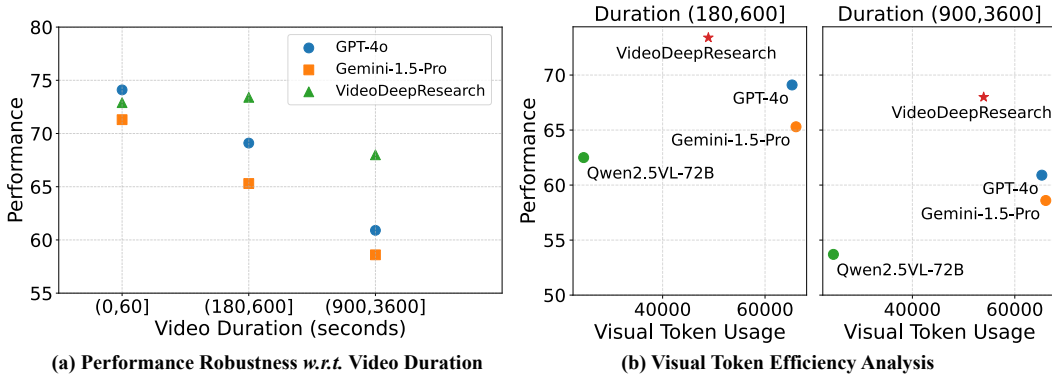


Figure 4: (a) Performance robustness w.r.t video duration and (b) Visual token efficiency analysis on LongVideoBench.

we present the performance of the visual perceiver Seed1.5VL when used independently for the question-answering task, in order to isolate the impact of the visual perceiver module. Based on the experimental results, our VideoDeepResearch framework achieves superior performance across most tasks, especially those requiring fine-grained retrieval, such as NeedleQA, action count, action order, and TutorialQA. It outperforms the best baseline models by 5.0%, 12.2%, 28.2%, and 17.9%, respectively. This strong performance is attributed to our effective decomposition of complex questions and accurate localization of relevant video segments, which avoids the significant information loss caused by brute-force downsampling and overcomes the limitations of naive RAG in handling complex tasks. In addition, our framework also excels in the Anomaly task, surpassing the best baseline by 8.8%. However, for EgoQA and SportsQA, our model underperforms. Upon inspection, we found that the retrieval module used in our framework struggles to accurately localize relevant segments for these tasks, making it difficult for the reasoning LLM to generate correct answers. In future work, we plan to improve both the retriever and retrieval strategy to achieve more robust performance across all task types.

Video Duration Analysis. As shown in Figure 4(a), we evaluate model performance across varying video durations. The results reveal a consistent decline in efficiency as video length increases: GPT-4o and Gemini-1.5 Pro exhibit performance drops of 13.2 and 12.7 points, respectively, when comparing durations of (900,3600] seconds to (0,60] seconds. In contrast, VideoDeepResearch framework demonstrates greater robustness, with only a 4.9-point reduction under the same conditions. This resilience stems from our model’s selective processing strategy. Instead of globally downsampling frames (a brute-force approach that incurs significant information loss in existing MLLMs), our method dynamically identifies and focuses on high-information segments critical to the task. These regions receive fine-grained analysis, while irrelevant segments are efficiently skipped. As video duration grows, this targeted approach increasingly outperforms conventional methods by a larger margin, highlighting its scalability advantage.

3.4 EFFICIENCY ANALYSIS

Existing long-video models address the challenges of long videos by extending the input context length, which incurs significant computational overhead and redundant calculations. To compare the visual token efficiency of different models, we analyze the number of visual tokens used in total and their corresponding effectiveness across different video duration categories on LongVideoBench. For MLLMs, the visual tokens refer to the input visual tokens. In our model, the visual tokens include all tokens passed into the visual perceiver and video browser.

As shown in Figure 4 (b), our VideoDeepResearch model demonstrates several significant advantages:

1). Superior Performance: VideoDeepResearch consistently outperforms all baseline models across both duration categories. For medium-long videos (180-600 seconds), our model achieves 73.4% performance, representing a 4.3 percentage point improvement over the best baseline (GPT-4o at 69.1%). For longer videos (900-3600 seconds), VideoDeepResearch maintains strong performance at 68.0%, significantly outperforming the best baseline by 7.1 percentage points. **2). Computational Efficiency:** While maintaining superior performance, VideoDeepResearch demonstrates remarkable computational efficiency. Our model requires only 48,932 tokens for shorter videos and 53,920 tokens for longer videos, representing a 25.0% and 17.4% reduction in token usage compared to GPT-4o and Gemini-1.5-Pro respectively, while still achieving the highest performance.

The experimental results validate VideoDeepResearch’s ability to maintain high performance while utilizing fewer computational resources. These results demonstrate that our approach effectively captures essential video understanding patterns without redundant processing.

4 RELATED WORK

4.1 LARGE REASONING MODELS

Recent advancements have significantly enhanced LLM reasoning capabilities. Traditional techniques improve reasoning through prompting methods like chain-of-thought (Wei et al., 2023) or structured reasoning processes such as tree-of-thought (Yao et al., 2023), guiding LLMs to generate intermediate reasoning steps before final answers. Recent research explores trainable methods to further boost reasoning. OpenAI’s O1 (OpenAI, 2025a) discovered test-time scaling effects, enabling free-form reasoning before question responses and achieve significant improvement. DeepSeek-R1 (Team, 2025) incentivizes reasoning capabilities through an effective reinforcement learning algorithm called Group Robust Preference Optimization (GRPO), which automates high-quality reasoning trajectory generation via trial-and-error search without extensive human oversight.

Concurrently, new frameworks leverage strong LLMs for deep research (OpenAI, 2025b; Perplexity, 2025). These systems perform multi-step internet research for complex tasks by integrating tools like search engines, enabling iterative reasoning and tool invocation. Inspired by this, we propose to address long-video understanding through iterative task analysis and multimodal tool invocation, which solves diverse and complex tasks in the long video understanding field.

4.2 MULTI-MODAL LARGE LANGUAGE MODELS

Advancements in large language models (LLMs) (Brown et al., 2020) have significantly advanced visual understanding by incorporating visual features, leading to the emergence of Multi-modal Large Language Models (MLLMs) (Wang et al., 2024b; Chen et al., 2024a; Liu et al., 2023; OpenAI, 2024; Bai et al., 2025; Zhu and etc, 2025; Yuanhan Zhang et al., 2024; Lin et al., 2023; Zhang et al., 2024; Chen et al., 2024b). These models are designed for general-purpose vision tasks and have also demonstrated strong performance in video understanding. However, most existing MLLMs primarily focus on short-context comprehension, typically handling only a few images or short video clips at a time, which limits their effectiveness on long video understanding tasks (Fu et al., 2024; Zhou et al., 2024). To address this limitation, recent efforts have explored more efficient architectures to extend the input context (Xue et al., 2024; Shu et al., 2024; Liu et al., 2025), enabling the processing of longer videos. Nonetheless, the length of input context that these models can handle remains restricted, and understanding videos of hour-long durations remains a significant challenge. In this work, we propose an agentic framework for reasoning over long videos by integrating retrieval-based tool use into the reasoning process. By decomposing complex queries and progressively identifying

key segments needed for solving the task, our approach effectively reduces the reliance on long model contexts, thereby enabling efficient and scalable long video understanding.

4.3 AGENTIC FRAMEWORK FOR LONG VIDEO UNDERSTANDING

Leveraging the powerful capabilities of large language models (LLMs) (OpenAI, 2023; Brown et al., 2020), recent studies have explored agentic approaches to tackle the challenges of long video understanding (LVU) (Wang et al., 2024c; Zhi et al., 2025; Kugo et al., 2025; Cao et al., 2025). However, many of these methods are designed for limited scenarios such as egocentric or short videos (Wang et al., 2024c; Zhi et al., 2025; Kugo et al., 2025), which restricts their applicability to more general real-world settings. In addition, these approaches often rely on dense processing of all video frames with captioning, leading to prohibitively high computational costs (Zhi et al., 2025). In this paper, we propose a general agentic framework that enables progressive reasoning over long videos through iterative, selective analysis with multi-modal tools, which support a wide range of scenarios such as tutorial, sports, vlog, ego... and a wide range of long video tasks including single-detail reasoning, multi-detail reasoning, multi-hop reasoning and etc. Additionally, our framework keeps the cost well below that of GPT-4o, which enables a favorable trade-off between effectiveness and efficiency.

5 CONCLUSION

In this work, we introduce a paradigm shift in approaching long video understanding tasks by demonstrating that sophisticated foundation models with extended context windows are not the only viable solution. Through VideoDeepResearch, we have shown that a strategic combination of text-only reasoning models and modular multi-modal toolkit can effectively tackle the computational and complexity challenges inherent in LVU tasks. Our experimental results across MLVU, Video-MME, and LVBench benchmarks provide compelling evidence that agentic frameworks can achieve superior performance while maintaining computational efficiency. The key insight lies in leveraging strategic reasoning to selectively access and process relevant video content, rather than attempting to process entire long videos through monolithic models.

Our framework challenges the prevailing assumption that LVU necessarily requires resource-intensive foundation MLLMs with extensive context capabilities. Instead, our approach demonstrates that intelligent orchestration of existing, readily available components can yield both effectiveness and efficiency gains. This finding has significant implications for the practical deployment of video understanding systems, particularly in resource-constrained environments.

Looking forward, the agentic approach to multi-modal understanding opens new research directions that emphasize strategic reasoning and selective processing over brute-force computational power. Our work establishes a foundation for future investigations into how intelligent agent frameworks can address complex multi-modal tasks across various domains, potentially reshaping how we approach video analysis and understanding at scale.

REFERENCES

- Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.
- Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.
- OpenAI. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>, May 2024.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Huaying Yuan, Zheng Liu, Minhao Qin, Hongjin Qian, Y Shu, Zhicheng Dou, and Ji-Rong Wen. Memory-enhanced retrieval augmentation for long video understanding, 2025. URL <https://arxiv.org/abs/2503.09149>.
- 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. URL <https://arxiv.org/abs/2411.13093>.
- 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.
- 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, 2024a. URL <https://arxiv.org/abs/2406.08035>.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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. URL <https://arxiv.org/abs/2502.13923>.
- Gemini Team and etc. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL <https://arxiv.org/abs/2403.05530>.
- Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi 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.
- DeepSeek-AI Team. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Dong Guo and etc. Seed1.5-vl technical report, 2025. URL <https://arxiv.org/abs/2505.07062>.
- Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024.

- Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. URL <https://arxiv.org/abs/2503.21776>.
- Jinguo Zhu and etc. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- OpenAI. Learning to reason with LLMs, 2025a. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- OpenAI. Introducing deep research, 2025b. URL <https://openai.com/index/introducing-deep-research/>.
- Perplexity. Introducing perplexity deep research, 2025. URL <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 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*, 2024b.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next-video: Advancing video understanding with llava-next. April 2024. URL <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>. Accessed: 2024-05-15.
- Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. URL <https://arxiv.org/abs/2410.02713>.
- Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024b.
- Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134, 2025.
- OpenAI. Gpt-4 technical report, 2023.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent, 2024c. URL <https://arxiv.org/abs/2403.10517>.

Zhuo Zhi, Qiangqiang Wu, Minghe shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot, 2025. URL <https://arxiv.org/abs/2504.04471>.

Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, Kazuki Kozuka, and Ehsan Adeli. Videomultiagents: A multi-agent framework for video question answering, 2025. URL <https://arxiv.org/abs/2504.20091>.

Shiwen Cao, Zhaoxing Zhang, Junming Jiao, Juyi Qiao, Guowen Song, Rong Shen, and Xiangbing Meng. Masr: Self-reflective reasoning through multimodal hierarchical attention focusing for agent-based video understanding, 2025. URL <https://arxiv.org/abs/2504.17213>.