SiLVR: A Simple Language-based Video Reasoning Framework

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in test-time optimization have led to remarkable reasoning capabilities in Large Language Models (LLMs), enabling them to solve highly complex problems in math and coding. However, the reasoning capabilities of multimodal LLMs (MLLMs) still significantly lag, especially for complex video-language tasks. To address this issue, we present SiLVR, a **Si**mple **L**anguage-based **V**ideo **R**easoning framework that decomposes complex video understanding into two stages. In the first stage, SiLVR transforms raw video into language-based representations using multisensory inputs, such as short clip captions and audio/speech subtitles. In the second stage, language descriptions are fed into a powerful reasoning LLM to solve complex video-language understanding tasks. To handle long-context multisensory inputs, we use an adaptive token reduction scheme, which dynamically determines the temporal granularity with which to sample the tokens. Our simple, modular, and training-free video reasoning framework achieves the best-reported results on Video-MME (long), Video-MMMU (comprehension), Video-MMLU, CGBench, and EgoLife. Furthermore, our empirical study focused on video reasoning capabilities shows that despite not being explicitly trained on video, strong reasoning LLMs can effectively aggregate multisensory input information from video, speech, and audio for complex temporal, causal, long-context, and knowledge acquisition reasoning tasks in video. Code will be made available.

1 Introduction

Recent years have witnessed remarkable progress in general video understanding, with large multimodal models achieving strong performance on tasks such as video question-answering (VideoQA) (Team et al., 2023; Hurst et al., 2024; Bai et al., 2025; Zhang et al., 2024c), text-video retrieval (Zhao et al., 2023), and temporal localization (Huang et al., 2024; Ren et al., 2024; Wang et al., 2024c). Despite the remarkable progress, most existing methods struggle with complex video-language understanding tasks that require strong reasoning capabilities (e.g., temporal, causal, long-context, external knowledge, etc.). Following the success of reasoning LLMs (Guo et al., 2025; Jaech et al., 2024), several recent multimodal large language models (MLLMs) proposed reasoning frameworks for multimodal image/video recognition tasks (Liu et al., 2025; Fei et al., 2024; Wang et al., 2024d; Wu et al., 2025; Feng et al., 2025; Chen et al., 2025; Li et al., 2025; Wang et al., 2025). However, these methods either rely on high-quality Chain-of-Thought (CoT) data, which is expensive and time-consuming to collect, or require task-specific reward designs, leading to poor generalization. Moreover, such RL-based multimodal reasoning approaches are difficult to optimize and often require a large amount of resources for training. Lastly, many recently proposed RL post-training techniques lead to similar or sometimes even worse performance than standard Supervised Fine-tuning (SFT) approaches (Wang & Peng, 2025; Feng et al., 2025).

Motivated by the impressive reasoning abilities of recent LLMs (Guo et al., 2025; Jaech et al., 2024), we propose SiLVR, a simple, modular, and training-free language-based framework for complex video-language reasoning tasks. SiLVR decomposes video understanding into two stages:

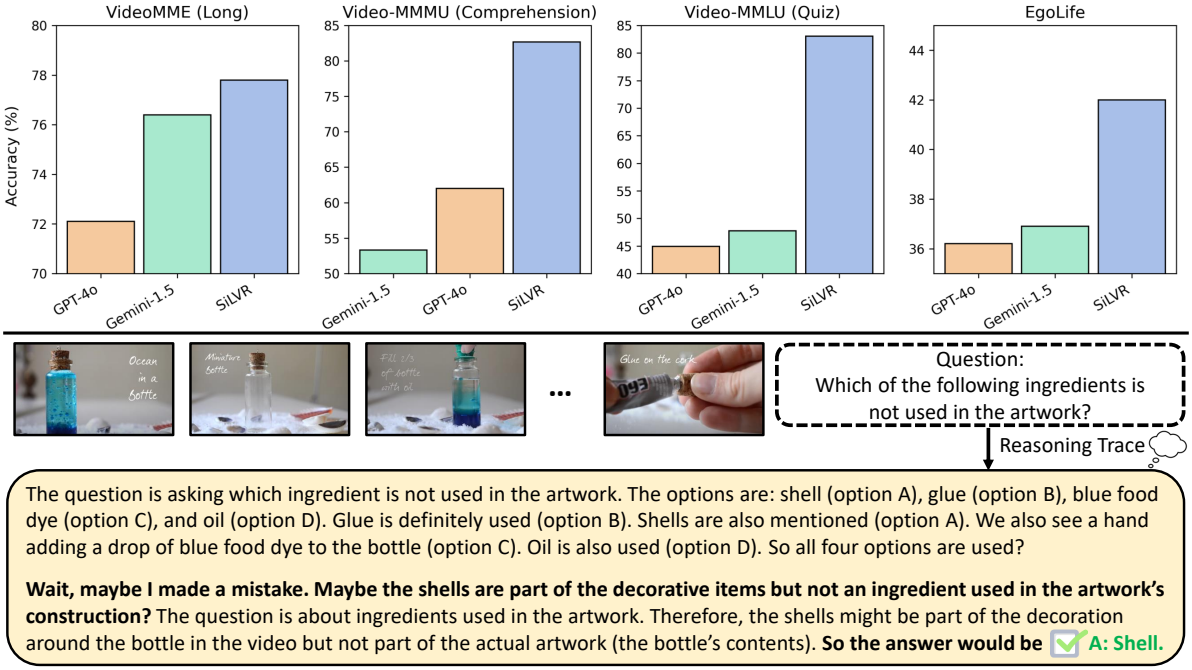

Figure 1: **Strong reasoning capabilities of SiLVR on complex video QA tasks.** SiLVR leverages recent advances in reasoning LLMs for complex video QA tasks. SiLVR achieves better performance than strong proprietary non-reasoning models (i.e., GPT-4o and Gemini-1.5) on benchmarks like Video-MME (Long), Video-MMMU (Comprehension), Video-MMLU, and EgoLife, which include temporal, causal, long-context, and external knowledge reasoning tasks. The example reasoning trace shows SiLVR's capability to perform self-correction, in which it successfully identifies that shells are decorative rather than functional.

- In the first stage, we convert raw videos into rich language-based descriptions. Specifically, we densely sample short clips from the input videos and use a pre-trained visual captioner (e.g., NVILA (Liu et al., 2024b)) to extract captions for each clip. Additionally, we use automatic speech recognition (ASR) tools to convert speech into language descriptions.
- In the second stage, we feed the rich language descriptions into a strong reasoning LLM (e.g. DeepSeek-R1 (Guo et al., 2025)) to solve complex video-language understanding tasks.

To address the issue of processing a large number of tokens in potentially hour-long videos, we propose a simple adaptive token reduction scheme, which dynamically determines the temporal granularity with which to sample the speech and video tokens. Such a token reduction scheme enables us to significantly reduce the number of input tokens to fit within the context length of an LLM, while maintaining strong reasoning performance.

Compared to prior MLLM-based video reasoning frameworks, SiLVR is simple, modular, training-free, and highly-performant. SiLVR achieves state-of-the-art results on multiple VideoQA benchmarks, including Video-MME (long), Video-MMMU (comprehension), Video-MMLU, CGBench, and EgoLife. Additionally, SiLVR demonstrates strong spatiotemporal grounding ability for video QA tasks that require localizing relevant video segments. On CGBench, a large-scale grounded VideoQA benchmark, SiLVR outperforms the previous best method by a substantial margin of **6.1%** in mIoU. Additionally, our empirical study on video reasoning capabilities of our framework suggests that despite not being trained on videos, strong reasoning LLMs can successfully aggregate information from video, speech/audio for complex temporal, causal, long-context, and external knowledge reasoning tasks on video inputs.

While SiLVR is not based on any new complex design choices, it is simple, modular, training-free, and highly performant and generalizes to multiple complex video-language understanding tasks. We believe the simple

yet effective design of SiLVR will enable the research community to build on our work and use our simple framework as a baseline to develop even more powerful video-language reasoning models.

2 Related Work

Language Reasoning Models. Recent work has significantly advanced the reasoning capabilities of LLMs through various strategies, such as Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022), Self-Consistency (Wang et al., 2023), and Monte Carlo Tree Search based methods (Wan et al., 2024; Trinh et al., 2024; Xin et al., 2024). Recently, works such as DeepSeek-R1 (Guo et al., 2025) demonstrated that applying large-scale RL with accuracy and format rewards can induce emerging reasoning capabilities in LLMs. These RL-based methods have shown strong improvements in tasks such as mathematics (Zheng et al., 2021; Azerbayev et al., 2023) and code generation (Austin et al., 2021; Hendrycks et al., 2021). Motivated by these successes, we propose to take advantage of the strong reasoning ability of LLMs for complex video-language reasoning problems.

Multimodal Reasoning Models. There have been many efforts to augment MLLMs with reasoning capabilities. One line of work focuses on decomposing the reasoning process into multiple sub-problems (Zhang et al., 2023c; Xu et al., 2024; Zhang et al., 2024b). Motivated by the success of DeepSeek-R1 (Guo et al., 2025), another line of work explores RL to elicit the reasoning ability of the MLLMs (Huang et al., 2025; Shen et al., 2025; Yang et al., 2025b; Zhang et al., 2025; Ouyang, 2025; Peng et al., 2025). In the video domain, VideoCoT (Wang et al., 2024d) and Video-of-Thought (Fei et al., 2024) propose to prompt the MLLMs with multiple reasoning steps before answering the question. In addition, multiple concurrent works propose to use GRPO to enhance video reasoning (Wu et al., 2025; Feng et al., 2025; Chen et al., 2025; Li et al., 2025; Wang et al., 2025). However, many of these RL-based methods demand substantial training computation, and achieve only marginal improvements or perform even worse than the SFT methods (Wang & Peng, 2025; Feng et al., 2025). Unlike these methods, SiLVR is simple, training-free, yet highly performant across a wide range of VideoQA benchmarks.

Complex Video-Language Understanding. A variety of benchmarks for complex video-language understanding have been proposed, with a focus on comprehensive evaluation of videos with different durations and questions spanning diverse categories (Fu et al., 2024; Li et al., 2024b; Liu et al., 2024a; Rawal et al., 2024). In parallel, several benchmarks have been introduced to assess the reasoning capabilities of large video-language models (Hu et al., 2025; Song et al., 2025; Zhao et al., 2025; He et al., 2024). On the modeling side, recent video MLLMs adapt image MLLMs by fine-tuning additional modules for temporal modeling (Lin et al., 2023; Li et al., 2023; 2024a; Zhang et al., 2023b). Several follow-up works explored spatiotemporal token compression (Islam et al., 2025; Liu et al., 2024b; Bai et al., 2025; Shen et al., 2024), building hierarchical memory (Song et al., 2023; Jin et al., 2024; Islam et al., 2024), or incorporating information from external models (Tang et al., 2025; Wang et al., 2024a) for complex and long video understanding. Another line of work explores training-free frameworks that first convert raw videos into dense visual captions, then perform reasoning with off-the-shelf LLMs (Zhang et al., 2023a; Wang et al., 2024b; Fan et al., 2024; Ma et al., 2024; Liao et al., 2024; Wang et al., 2024e; Min et al., 2024). Our method adopts a similar design. However, SiLVR focuses on solving complex video-language reasoning problems with strong reasoning LLMs and effectively integrates both visual and audio modality streams.

3 Method

Our method decomposes video-language QA into two stages: 1) extracting visual captions and transcribing speech into text, and 2) performing language-based reasoning over the extracted textual descriptions. Such a decomposed video reasoning design offers several benefits: 1) **Simplicity**: SiLVR does not require complex RL-based optimization or specialized modules for different tasks. 2) **Generalizability**: our method can be applied to a wide range of complex video-language tasks without task-specific fine-tuning. 3) **Modularity**: our method's modular design enables seamless use of powerful visual captioning models and strong reasoning LLMs. 4) **Flexibility**: SiLVR supports plug-and-play integration of different captioning models, speech recognition models, and LLMs. An overview of our method is illustrated in Figure 2.

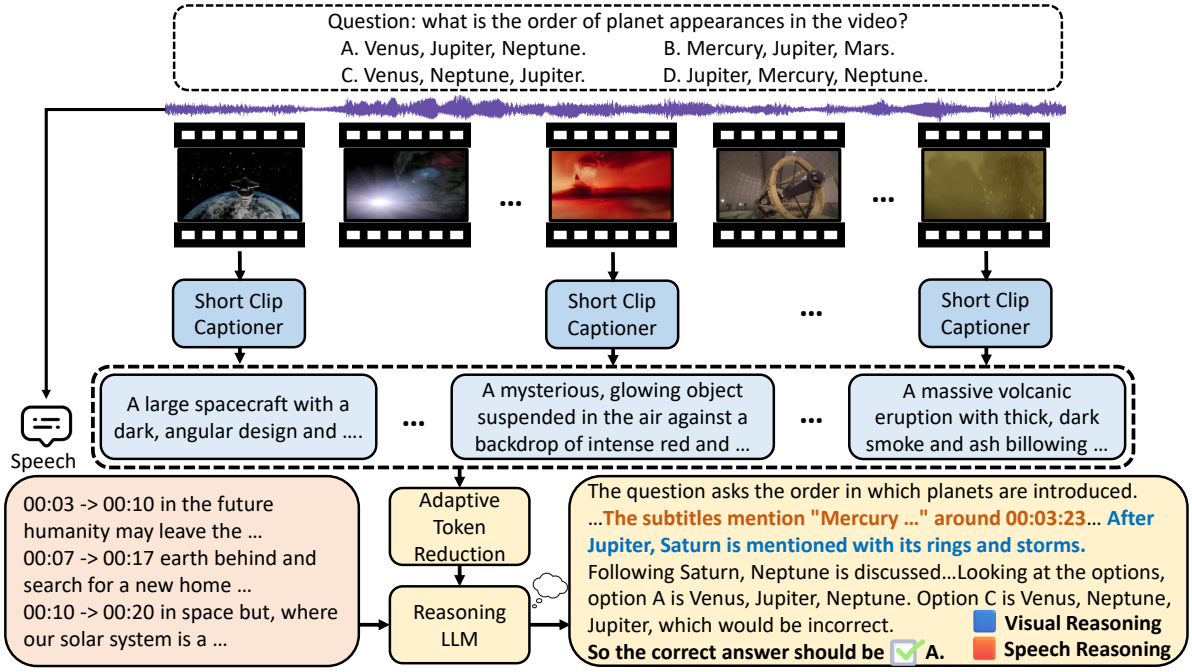

Figure 2: **Method overview.** SiLVR is a simple two-stage language-based video reasoning framework. **Top**: The video is segmented into short clips and paired with speech. A clip captioner processes each segment to generate visual descriptions. The speech is transcribed using ASR. **Bottom**: A reasoning LLM takes the question, transcribed speech, and dense visual descriptions compressed by Adaptive Token Reduction to perform complex video reasoning. In the shown example, SiLVR infers the correct order by integrating information across both visual and speech modalities. The model correctly identifies the sequence through reasoning and eliminating incorrect options.

3.1 Extracting Multimodal Descriptions

Given a video V, we first divide it into N non-overlapping short clips $v = \{v_i\}_{i=1}^{N}$, where each clip $v_i \in \mathbb{R}^{T\times H\times W\times 3}$ contains T frames of height H and width W. Each clip is passed through a pretrained visual captioning model $\mathcal{M}$ to produce a caption c_i. The sequence of captions is denoted as $C = \{c_i\}_{i=1}^{N}$, forming a temporally ordered description of the visual content.

In parallel, we apply an ASR model $\mathcal{W}$ to convert the speech into a sequence of textual descriptions $S = \{s_j\}_{j=1}^{K}$, where s_j is a timestamped transcription of a spoken segment. K denotes the total number of segments, which is determined by an ASR model. We then concatenate S and C one after the other to form a rich, language-based description of the video (including audio/speech) and feed them into a reasoning LLM as described next.

3.2 Language-Based Reasoning

To answer a question Q about the video, we feed the video captions C and speech transcriptions S along with Q into a reasoning LLM. We design several prompts to guide the LLM to reason jointly over visual and speech information (for complete prompts see Supplementary Material). Unlike prior video reasoning approaches, SiLVR performs reasoning entirely in the language space. However, the limited context window of LLMs poses a significant challenge when processing long videos with rich multimodal content. To address this issue, we introduce a simple adaptive token reduction scheme (see Algorithm 1). Our adaptive token reduction scheme dynamically adjusts the temporal granularity for sampling video tokens. Specifically, it starts with a small clip length and progressively increases it to reduce the total number of generated tokens.

Algorithm 1 Adaptive Token Reduction

Require: Video V, Question Q, LLM F, Captioner M, Speech Recognition Model W, Initial Clip Length L

```
S ← extractSubtitles(V, W)
limit ← getContextLength(F)
while True do
    v ← divideVideo(V, L)
    C ← generateCaptions(v, M)
    Z ← concat(S, C)
    if countTokens(Z) > limit then
        L ← L × 2
    else
        break
    end if
end while
return answer(Z, Q, F)
```

This allows us to effectively fit the input tokens within the LLM's context window for videos of varying durations while maintaining strong video reasoning performance.

3.3 Implementation Details

We use NVILA (Liu et al., 2024b) as the default visual captioning model. We use our adaptive token reduction scheme for all videos as described above. For speech transcription, we use Whisper-large-v3 (Radford et al., 2022). Due to its strong reasoning performance, we use DeepSeek-R1 (Guo et al., 2025) as the default LLM and set the temperature to 1.0 for all experiments. We use the official evaluation code provided by each benchmark, or use LMMs-Eval (Zhang et al., 2024a) when the official code is unavailable. Additional implementation details are provided in the supplementary materials.

4 Experiments

4.1 Benchmarks and Evaluation Metrics

We conduct experiments on eight complex video-language understanding benchmarks: Video-MMMU (Hu et al., 2025), Video-MMLU (Song et al., 2025), MMVU (Zhao et al., 2025), MMWorld (He et al., 2024), Video-MME (Fu et al., 2024), CGBench (Chen et al., 2024), EgoLife (Yang et al., 2025a) and CinePile (Rawal et al., 2024). Following Video-R1 (Feng et al., 2025), we group these benchmarks into two categories: Reasoning Benchmarks and General Benchmarks. The reasoning benchmarks include Video-MMMU, Video-MMLU, MMVU, and MMWorld, which primarily evaluate the reasoning capabilities of large video-language models. Specifically, Video-MMMU and Video-MMLU focus on lecture-based video understanding, where the model must extract knowledge from lecture videos to answer the questions. MMVU requires models to apply domain-specific knowledge and perform expert-level reasoning to analyze specialized-domain videos. MMWorld focuses on a diverse set of reasoning questions (e.g., counterfactual thinking, future prediction, etc.) across videos from seven broad disciplines. The other four benchmarks (i.e., Video-MME, CGBench, EgoLife, and CinePile) are general video-language benchmarks, which contain various types of questions and offer a comprehensive assessment of the video-language models. Specifically, Video-MME includes three splits (short, medium, and long) based on the duration of the video. CGBench and EgoLife are designed for long video understanding, with an average video duration of more than an hour. We focus on VideoQA for all benchmarks and report QA accuracy as our primary evaluation metric. Additionally, we conduct Grounded VideoQA experiments on CGBench to assess the model's temporal grounding ability and use the mean Intersection over Union (mIoU) to evaluate the results.

Table 1: **Main results.** We evaluate our method on a set of video reasoning benchmarks (Video-MMMU, Video-MMLU, MMVU, MMWorld) and general video benchmarks (Video-MME, CGBench, EgoLife, CinePile). We use the comprehension split of Video-MMMU and the long split of VideoMME (with subtitles). Based on these results, we observe that SiLVR achieves the best-reported results on Video-MMMU (comprehension), Video-MMLU, Video-MME (long split, with subtitles), CGBench, and EgoLife, outperforming strong proprietary models such as Gemini 2.0 and GPT-4o. We **bold** and underline the best and the second best models in each benchmark, respectively.

Model	Video Reasoning Benchmarks				General Video Benchmarks			
	Video-MMMU	Video-MMLU	MMVU	MMWorld	Video-MME	CGBench	EgoLife	CinePile
Proprietary Models								
Gemini 1.5 Flash	49.0	47.8	58.8	-	68.8	33.5	-	58.8
Gemini 1.5 Pro	53.5	-	65.4	51.0	77.4	37.8	36.9	**60.1**
Gemini 2.0 Flash	-	-	65.9	-	-	-	-	-
GPT-4o	62.0	44.9	67.4	**62.5**	72.1	44.9	36.2	56.1
Claude 3.5 Sonnet	75.7	71.3	65.2	54.5	-	40.3	-	-
Kimi-k1.6	76.7	-	-	-	-	-	-	-
OpenAI o1	-	-	**75.5**	-	-	-	-	-
Open-Source LMMs								
LLaVA-OV-7B	31.0	33.4	37.9	-	-	30.9	30.8	49.3
VideoLLaMA3-7B	46.0	-	45.0	-	61.0	-	-	-
Aria	53.0	-	49.3	-	66.3	-	-	-
Qwen-2-VL-72B	-	-	50.3	-	74.3	45.3	-	-
DeepSeek-VL2	-	-	52.1	-	-	-	-	-
SiLVR (ours)	**82.7**	**83.1**	68.2	59.9	**77.7**	**51.8**	**42.0**	59.4

4.2 Main Results

Video Reasoning Benchmarks. We present our results on video reasoning benchmarks on the left side of Table 1. Our results indicate that SiLVR achieves the best performance on two reasoning benchmarks: Video-MMMU (comprehension) and Video-MMLU. Specifically, on Video-MMMU, SiLVR outperforms the prior best method, Kimi-k1.6 (Team et al., 2025) by a large margin (**+6.0%**). It also significantly outperforms other strong proprietary models such as Gemini 1.5 Pro and GPT-4o by **29.2%** and **22.7%**, respectively. On Video-MMLU, SiLVR outperforms the previous state-of-the-art model Claude 3.5 Sonnet by a substantial **11.8%**. Additionally, on MMVU, we observe that our modular framework, with DeepSeek-R1 as the LLM, outperforms the unified multimodal model DeepSeek-VL2 by a significant margin of **15.9%**. These results suggest that despite the simplicity of our approach, it delivers strong performance across a wide range of video-language reasoning tasks.

General Video Benchmarks. We present our results on general video benchmarks on the right side of Table 1. Based on these results, we observe that SiLVR achieves state-of-the-art performance on three general benchmarks: Video-MME (long split, with subtitles), CGBench, and EgoLife. Specifically, on Video-MME and EgoLife, SiLVR outperforms the prior best performing method Gemini 1.5 Pro by **0.3%** and **5.1%**, respectively. On CGBench, SiLVR achieves 51.8% accuracy, outperforming the previous state-of-the-art method Qwen-2-VL-72B by a significant **6.9%** margin. SiLVR also surpasses strong proprietary models, outperforming GPT-4o by **6.9%** and Claude 3.5 Sonnet by **11.5%** on CGBench. Additionally, it is worth noting that Video-MME (long), EgoLife, and CGBench are designed for very long-form video understanding, with average video durations exceeding 60 minutes. Our strong results demonstrate that SiLVR is highly effective in comprehending long videos.

4.3 Reasoning Analysis

In this section, we conduct a more in-depth analysis of the video reasoning capabilities of our approach. To do this, we systematically compare the performance of our framework when using reasoning (e.g., DeepSeek-R1)

Table 2: **Performance comparison between a reasoning (DeepSeek-R1) and a non-reasoning (Llama 4) LLM.** Using DeepSeek-R1 reasoning LLM leads to significantly better results over a non-reasoning Llama 4 on all eight benchmarks. However, we also observe that the average gain on video reasoning benchmarks (VideoMMMU, VideoMMLU, MMVU, MMWorld) is significantly larger than on general video benchmarks (VideoMME, CGBench, EgoLife, CinePile). These results demonstrate that the strong reasoning ability of DeepSeek-R1 is crucial for solving complex video reasoning tasks.

Model	Video Reasoning Benchmarks				General Video Benchmarks			
	VideoMMMU	VideoMMLU	MMVU	MMWorld	VideoMME	CGBench	EgoLife	CinePile
SiLVR (Llama 4)	56.3	57.2	60.6	57.2	67.8	53.2	38.5	45.6
SiLVR (DeepSeek-R1)	**82.7**	**83.1**	**68.2**	**59.9**	**77.7**	**59.4**	**42.0**	**51.8**
	Average Gain: **+15.7**				Average Gain: **+6.5**			

vs. non-reasoning (e.g., Llama 4) LLMs across multiple benchmarks. Additionally, we break down the performance of our approach across different types of video reasoning questions (e.g., temporal, causal, long-context, knowledge acquisition, etc.).

Reasoning vs Non-Reasoning LLMs. To study the impact of a strong reasoning LLM within our framework, we compare the performance of our method when using a reasoning LLM (DeepSeek-R1) vs. a non-reasoning LLM (Llama 4). The results are presented in Table 2. Our results suggest several interesting trends. First, we observe that DeepSeek-R1 consistently outperforms Llama 4 across all benchmarks, indicating that it is a much stronger LLM than Llama 4. Second, we note that using DeepSeek-R1 leads to much larger performance gains on the reasoning benchmarks, where DeepSeek-R1 surpasses Llama 4 by a substantial **26.4%** on Video-MMMU and **25.9%** on Video-MMLU with an average improvement of **+15.7%** on all video reasoning benchmarks. In contrast, while DeepSeek-R1 also produces better results on general video benchmarks, the improvements over Llama 4 are much smaller (i.e., average improvement of **6.5%** on general video benchmarks vs. **15.7%** on the reasoning benchmarks). These results suggest that the strong reasoning ability of DeepSeek-R1 is critical for solving complex video reasoning tasks and that our framework's simple and modular design allows us to take full advantage of DeepSeek-R1's strong reasoning abilities on these complex video reasoning problems.

Performance Breakdown Across Different Tasks. In Figure 3, we report the performance gains of using a reasoning LLM (DeepSeek-R1) over a non-reasoning LLM (Llama 4) for different question categories on VideoMME, which contains 12 manually annotated categories. The four categories that we report on the left of Figure 3, belong to reasoning (e.g., temporal, spatial, object, and action reasoning). The other eight categories are classified as non-reasoning and require general perception capabilities (e.g., action recognition, OCR, etc.). Based on the results in Figure 3, we observe that compared to Llama 4, using DeepSeek-R1 achieves a significantly larger improvement on reasoning questions (a gain of **+11.1%**) compared to non-reasoning questions (a gain of **+4.9%**). This result is consistent with our observations in Table 2, which confirms that reasoning LLMs bring greater benefits for tasks that require complex reasoning.

4.4 Efficiency Analysis

In Table 3, we analyze the inference efficiency of our framework and compare it against prior video understanding models. To ensure fairness, we follow the evaluation setup of VideoTree (Wang et al., 2024e), which excludes ASR inputs, and measure the average processing time per video on the Video-MME (Long) benchmark using a single A6000 GPU. In addition to the best-performing variant of our method (SiLVR-best), we also report a more efficient variant (SiLVR-fast), which uses a larger clip length to reduce the number of sampled clips and generated captions. From Table 3, we can observe that both SiLVR-best and SiLVR-fast achieve higher accuracy than DrVideo (Ma et al., 2024) and VideoTree (Wang et al., 2024e) while invoking the LLM only once per video, compared to 5 and 34 calls required by the baselines. Additionally, SiLVR-fast is highly efficient, running **5.3x** faster than DrVideo and **2.0x** faster than VideoTree, while still achieving

Table 3: **Efficiency comparison with other methods on Video-MME (Long).** To ensure fair comparison, we follow the evaluation setup of VideoTree and use one single A6000 GPU for all methods. Both SiLVR-best and SiLVR-fast significantly outperforms VideoTree and DrVideo while calling the LLM only once. Notably, SiLVR-fast runs **5.3x** faster than DrVideo and **2.0x** faster than VideoTree. These results demonstrate that SiLVR is both efficient and effective in complex video reasoning tasks.

Method	Captioning (s) ↓	Reasoning (s) ↓	Total (s) ↓	#LLM Calls ↓	Accuracy ↑
DrVideo	370	76	446	5	51.7
VideoTree	147	15	162	34	54.2
SiLVR-fast (ours)	48	35	**83**	**1**	57.2
SiLVR-best (ours)	378	64	442	**1**	**62.7**

Table 4: **Results on knowledge acquisition and temporally grounded QA tasks.** SiLVR achieves the highest $\Delta_{\text{knowledge}}$ on VideoMMMU and the best mIoU on CGBench.

Model	VideoMMMU ($\Delta_{\text{knowledge}}$)	CGBench (mIoU)
Qwen-2.5-VL-72B	9.7	-
Gemini-1.5 Pro	8.7	3.85
Claude-3.5 Sonnet	11.4	4.17
GPT-4o	15.6	5.73
VideoMind-7B	-	7.10
SiLVR (ours)	**17.2**	**11.84**

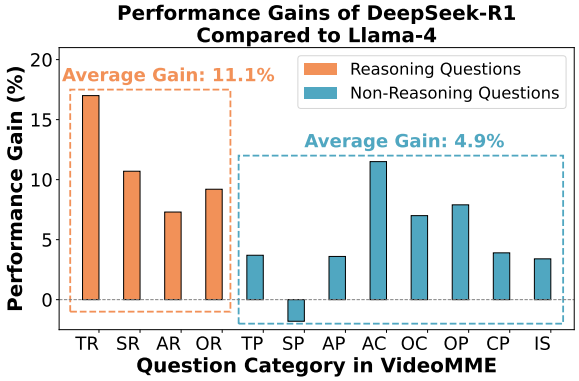

Figure 3: **Performance breakdown across different question categories.** Using DeepSeek-R1 as an LLM yields larger gains on reasoning questions (**+11.1%**) than general perception questions (**+4.9%**). Full category names are in the supplementary materials (Table 12).

+5.5% and **+3.0%** higher accuracy, respectively. These results demonstrate that our framework is able to achieve an effective balance between efficiency and accuracy, making it both effective and efficient in complex video reasoning tasks.

4.5 Results on Other Tasks

Knowledge Acquisition from Videos. We also evaluate our method on the novel knowledge acquisition task on Video-MMMU (Hu et al., 2025). The task requires models to answer questions both before and after watching a reference lecture video, with the goal of measuring how much knowledge the model gains from the video. The metric for the knowledge acquisition task is defined as:

$$\Delta_{\text{knowledge}} = \frac{\text{Acc}_{\text{post}} - \text{Acc}_{\text{pre}}}{100\% - \text{Acc}_{\text{pre}}} \times 100\% \tag{1}$$

where Acc_{pre} and Acc_{post} denote the accuracy before and after watching the video, respectively.

Our results in Table 4 show that SiLVR achieves **17.2%** in $\Delta_{\text{knowledge}}$, outperforming the prior best method GPT-4o by **1.6%**. SiLVR also outperforms strong proprietary models such as Gemini-1.5 Pro and Claude-3.5 Sonnet by **8.5%** and **5.8%**, respectively. These results demonstrate that SiLVR is not only effective in complex video reasoning, but also has strong knowledge acquisition capabilities.

Temporally Grounded QA. In Table 4, we also present our results on the temporally grounded QA task on CGBench (Chen et al., 2024). The task requires the model to temporally localize relevant video segments

Table 5: **Impact of ASR.** We remove the ASR module and evaluate our method on eight benchmarks. Incorporating ASR consistently improves performance, highlighting the importance of speech information in video reasoning.

Method	Video-MMMU	Video-MMLU	MMVU	MMWorld	Video-MME	CGBench	EgoLife	CinePile
SiLVR (w/ ASR)	**82.7**	**83.1**	**68.2**	**59.9**	**77.7**	**51.8**	**42.0**	**59.4**
SiLVR (w/o ASR)	70.7	78.2	61.7	57.4	62.7	40.9	35.1	47.4

Table 6: **Token reduction analysis.** Accuracy on VideoMME (overall) when selectively dropping speech vs. visual caption tokens. Based on these results, we observe that speech tokens are more informative than visual caption tokens. Furthermore, our adaptive token reduction strategy outperforms all static baselines.

Dropping Rate		Accuracy
Subtitles	Captions	
50%	-	65.3
75%	-	56.0
-	50%	68.9
-	75%	67.7
No Compression		70.3
Adaptive Token Reduction		**76.7**

needed to answer the question (usually less than 10 seconds) in long videos that span over 60 minutes. Our results in Table 4 show that SiLVR achieves the highest performance in mIoU, outperforming concurrent work VideoMind (Liu et al., 2025) by a notable **4.74%**. In addition, SiLVR also outperforms GPT-4o and Claude-3.5 Sonnet by **6.11%** and **7.67%**, respectively. These results suggest that SiLVR can correctly answer complex questions and temporally ground the answer to relevant segments in the video, which improves interpretability in video reasoning.

4.6 Ablation Studies

Unless otherwise specified, all ablation experiments are conducted on the VideoMME (Long) benchmark.

Impact of the ASR. To analyze the impact of ASR, we remove the ASR module of our method and evaluate our method on eight benchmarks. The results are shown in Table 5. From the table, we observe that ASR plays a crucial role across many benchmarks. On the lecture understanding benchmark Video-MMMU, incorporating ASR leads to a significant 12.0% accuracy gain, which aligns with the intuition that spoken content in lectures provides essential cues. Additionally, adding ASR leads to a 12.0% accuracy improvement on the movie understanding benchmark CinePile. Furthermore, in EgoLife and CGBench, which focus on analyzing short clues within long videos, ASR also provides notable gains. Finally, on general-purpose benchmarks like Video-MME and MMVU, adding ASR improves performance by a large margin, demonstrating that speech complements visual signals in multiple video understanding tasks.

Speech vs. Visual Caption Token Importance. To evaluate the relative contribution of visual and audio information, we vary the fraction of tokens from speech transcripts and video captions and report the QA performance on VideoMME. As shown in Table 6, the reduction of 50-75% speech tokens (while keeping all visual caption tokens) leads to a significant decrease in performance (**11.4%-20.7%**). In comparison, dropping the same fraction of visual caption tokens (while keeping all speech tokens) results in a much smaller performance drop (**7.8%-9.0%**). The No Compression baseline retains all tokens, achieving an overall accuracy of 70.3%. Our adaptive token reduction scheme, which prioritizes speech tokens and selectively reduces visual caption tokens, achieves an accuracy of 76.4%, outperforming other baselines.

Table 7: **Performance of adaptive token reduction (ATR) vs. fixed clip length baselines.** ATR achieves the highest overall accuracy, outperforming the best fixed-length baseline (8s) by **2.5%**. These results suggest that ATR effectively reduces redundant tokens while preserving strong performance.

Clip Length (s)	ATR	1	2	4	8	64
Accuracy	**76.7**	59.9	61.3	68.5	74.2	70.3

Table 8: **Performance with different visual captioners.** Qwen-2.5-VL 72B achieves the best overall accuracy. We use NVILA 7B for all experiments because it provides the best accuracy-cost trade-off.

Captioner	Overall	Short	Medium	Long
LLaVA-OV 7B	67.2	57.7	68.1	75.9
NVILA 7B	70.3	63.2	70.4	**77.3**
Qwen-2.5-VL 7B	70.9	63.8	**72.9**	76.1
Qwen-2.5-VL 72B	**71.2**	**65.0**	72.2	76.4

Table 9: **Performance of our framework with different LLMs.** Llama-4 Maverick achieves 66.2% overall accuracy, providing an effective trade-off between model sizes and performance. DeepSeek R1 achieves the highest overall accuracy, outperforming DeepSeek V3 and GPT-4.1 by **3.5%** and **0.8%**, respectively.

LLM	Overall	Short	Medium	Long
Llama-4-Scout 17B	63.0	56.7	64.4	67.8
Llama-4-Maverick 17B	66.2	57.2	68.3	73.0
DeepSeek V3	66.8	56.0	69.1	75.3
GPT-4o	67.3	57.0	69.2	75.8
GPT-4.1	69.5	59.6	**71.1**	**77.9**
DeepSeek R1	**70.3**	**63.2**	70.4	77.3

Analysis of Adaptive Token Reduction. In Table 7, we compare Adaptive Token Reduction (ATR) with several static baselines that use fixed video clip lengths. Among all baselines, the variant that uses an 8-second clip length achieves the highest accuracy of 74.2%. We note that a shorter clip variant (e.g., 1s) generates a large number of captions for long videos, which often exceeds the context window of the LLMs, thus leading to degraded performance. In contrast, a longer clip variant (e.g., 64s) reduces the number of captions at the cost of sacrificing the granularity of visual information, which also leads to lower accuracy. Compared to these static baselines, our proposed ATR consistently outperforms all fixed clip length baselines, surpassing the best-performing variant (8s) by a significant margin of **2.5%**. These results demonstrate that ATR effectively reduces redundant tokens by adaptively adjusting the clip length, offering flexibility and strong performance.

Visual Captioning Model. Next, we study the effect of different visual captioners. As shown in Table 8, Qwen-2.5-VL 72B achieves the highest overall accuracy, most likely due to the larger LLM (72B), which leads to higher-quality captions. We also observe that NVILA 7B and Qwen-2.5-VL 7B achieve similar performance, outperforming LLaVA-OV 7B by **2.9%** and **3.7%**, respectively. Since NVILA 7B is faster than Qwen-2.5-VL 7B and achieves similar performance, we use NVILA 7B for all experiments. We do not use Qwen-2.5-VL 72B due to the prohibitive computational cost.

Different LLMs. Lastly, in Table 9 we study the effect of different LLMs, including Llama-4-Scout 17B, Llama-4-Maverick 17B, GPT-4o, GPT-4.1, DeepSeek V3, and DeepSeek R1. Our results indicate that DeepSeek-R1, as the LLM backbone, achieves the best overall accuracy on Video-MME, outperforming the proprietary GPT-4.1 by **0.8%**. Additionally, we note that Llama-4-Maverick achieves 66.2% accuracy with only 17B parameters, offering an effective trade-off between model size and performance. Lastly, we observe

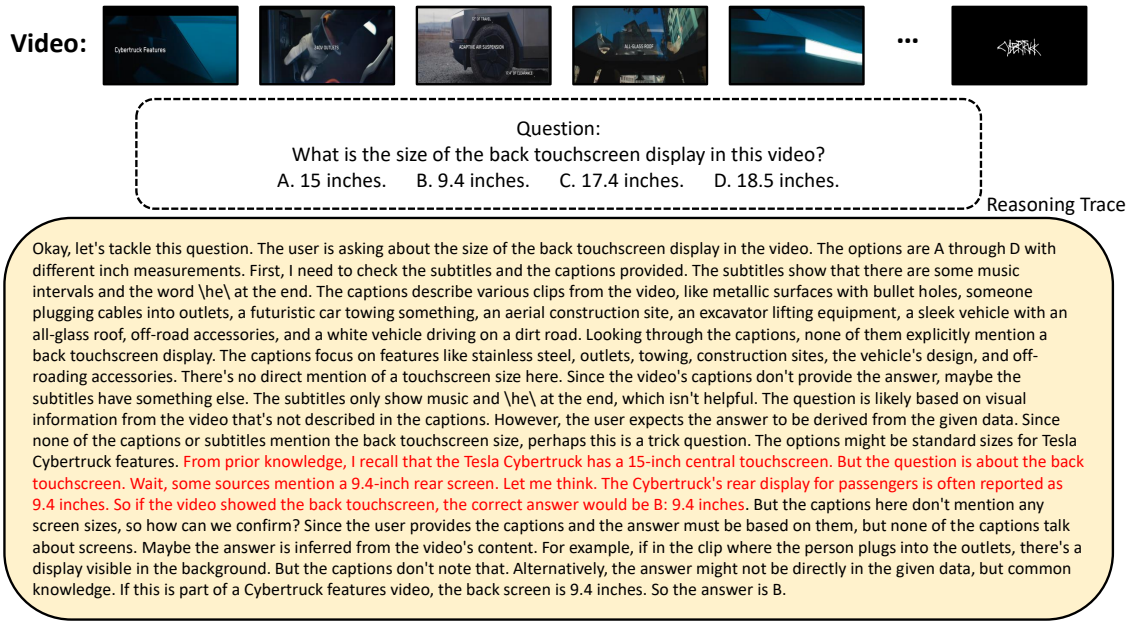

Figure 4: **Example 1 of SiLVR's reasoning trace.** The question asks about the size of the back touchscreen in the car shown in the video. The visual captioning module of SiLVR fails to capture the details about the touchscreen, which appears briefly for only about one second. However, by identifying the vehicle type and leveraging external knowledge from the LLM, SiLVR infers the correct answer.

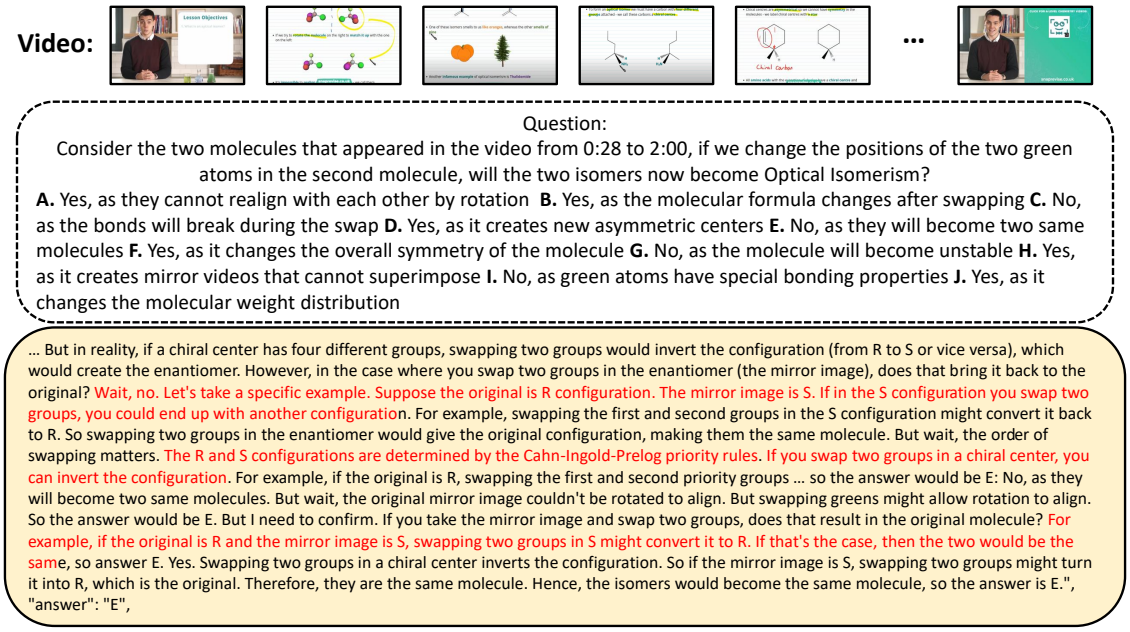

Figure 5: **Example 2 of SiLVR's reasoning trace.** Through step-by-step reasoning, SiLVR is capable of solving complex domain-specific questions. Notably, SiLVR does not immediately terminate the reasoning process upon reaching a plausible answer. Instead, it continues to verify the correctness of the generated answer before finalizing its response.

that DeepSeek R1 outperforms DeepSeek V3 by a significant **3.5%**, highlighting the effectiveness of using a reasoning LLM within our framework.

4.7 Qualitative Results

We present two reasoning traces of SiLVR in Figure 4 and Figure 5. In Figure 4, the video showcases the Tesla Cybertruck, and the question asks about the size of its rear touchscreen display. Since the display appears only briefly, the question is particularly challenging. As shown in the figure, SiLVR first recognizes the vehicle as a Tesla Cybertruck. Leveraging both visual cues and the LLM's prior knowledge, SiLVR then correctly infers that the size of the rear touchscreen is 15 inches. This example demonstrates SiLVR's capability to incorporate visual information with the LLM's prior knowledge for complex video reasoning. Figure 5 shows a chemistry tutorial video and a follow-up question. SiLVR initially makes a tentative prediction but does not terminate the reasoning process immediately. Instead, it continues to validate the predicted answer through step-by-step reasoning. These two examples highlight that SiLVR is effective across diverse video domains, highlighting the generalizability of our proposed method. Additional qualitative analyses are provided in the supplementary materials.

5 Conclusion

We present SiLVR, a simple, modular, and training-free language-based video reasoning framework. SiLVR achieves state-of-the-art performance on Video-MME (long), Video-MMMU (comprehension), Video-MMLU, CGBench, and EgoLife. SiLVR also achieves strong results in video-based knowledge acquisition and temporally grounded QA tasks, demonstrating strong generalization. Lastly, we systematically analyze the reasoning capabilities of SiLVR and perform ablations on several key design choices. We encourage the research community to build on our simple yet effective video reasoning framework and hope that it will inspire new ideas in video reasoning research.

Limitations

As with most modular frameworks, the performance of our method depends on its individual modules. On the visual perception side, our method relies on the visual captioning model, which may produce hallucinations or descriptions that lack fine-grained visual details. However, since our framework is agnostic to the specific use of visual captioning models, we believe that future advances in visual captioning models will mitigate this issue. On the reasoning side, our framework may underperform when the reasoning trace generated by the LLM is incorrect. However, we view this as a broader limitation of current LLMs, and anticipate that future advances in long-context modeling and reasoning for LLMs will further enhance the performance of our framework.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics, 2023. URL https://arxiv.org/abs/2302.12433.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

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.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. arXiv preprint arXiv:2412.12075, 2024.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. arXiv preprint arXiv:2503.24376, 2025.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In European Conference on Computer Vision, pp. 75–92. Springer, 2024.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong 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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. arXiv preprint arXiv:2406.08407, 2024.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with apps. NeurIPS, 2021.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14271–14280, 2024.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. arXiv preprint arXiv:2503.09590, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13700–13710, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213, 2022.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22195–22206, 2024b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958, 2025.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. arXiv preprint arXiv:2409.20365, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122, 2023.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. arXiv preprint arXiv:2503.13444, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? arXiv preprint arXiv:2403.00476, 2024a.

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. arXiv preprint arXiv:2412.04468, 2024b.

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

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13235–13245, 2024.

Kun Ouyang. Spatial-r1: Enhancing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805, 2025.

Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: pioneering multimodal reasoning with chain-of-thought. arXiv preprint arXiv:2504.05599, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024. URL https://arxiv.org/abs/2405.08813.

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, pp. 14313–14323, 2024.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. arXiv preprint arXiv:2503.11579, 2025.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434, 2024.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. arXiv preprint arXiv:2307.16449, 2023.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. arXiv preprint arXiv:2504.14693, 2025.

Zitian Tang, Shijie Wang, Junho Cho, Jaewook Yoo, and Chen Sun. How can objects help video-language understanding? arXiv preprint arXiv:2504.07454, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. Nature, 625(7995):476–482, 2024.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. In Forty-first International Conference on Machine Learning, 2024.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In European Conference on Computer Vision, pp. 142–160. Springer, 2024a.

Xiaodong Wang and Peixi Peng. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In European Conference on Computer Vision, pp. 58–76. Springer, 2024b.

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, Mohit Bansal, Gedas Bertasius, and David Crandall. Timerefine: Temporal grounding with time refining video llm. arXiv preprint arXiv:2412.09601, 2024c.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In ICLR, 2023.

Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. arXiv preprint arXiv:2407.05355, 2024d.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. arXiv preprint arXiv:2503.13377, 2025.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. arXiv preprint arXiv:2405.19209, 2024e.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. St-think: How multimodal large language models reason about 4d worlds from ego-centric videos. arXiv preprint arXiv:2503.12542, 2025.

Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, et al. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. arXiv preprint arXiv:2408.08152, 2024.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. URL https://arxiv.org/abs/2411.10440.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. arXiv preprint arXiv:2503.03803, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615, 2025b.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. arXiv preprint arXiv:2312.17235, 2023a.

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

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937, 2025.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. arXiv preprint arXiv:2410.16198, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713, 2024c.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023c.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. arXiv preprint arXiv:2501.12380, 2025.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6586–6597, 2023.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. arXiv preprint arXiv:2109.00110, 2021.

Our appendix consists of Additional Ablation Study (Section A), More Experimental Results (Section B), Additional Implementation Details (Section C), and Qualitative Results (Section D).

A Additional Ablation Study

Time-aware Caption Representation. We investigate the effectiveness of integrating time information into captions on the Video-MMMU benchmark. The results are presented in Table 10. The baseline concatenates all captions in temporal order without explicit timestamps, whereas our proposed time-aware method explicitly includes time information. Specifically, before each caption, we add a timestamp indicating the interval from which it was extracted (e.g. 00:00:24 –> 00:00:32: A clear glass bottle being filled with water, surrounded by seashells and a white cloth.) As shown in Table 10, incorporating timestamp information leads to a notable **2.25%** improvement in overall accuracy compared to the baseline. Furthermore, the time-aware method consistently outperforms the baseline across all question categories. These results demonstrate that explicitly providing time information effectively enhances the temporal perception and reasoning capabilities of the LLMs. Consequently, we adopt the time-aware caption representation for all experiments.

Method	Overall	Perception	Comprehension	Adaptation
w/o time	72.86	81.00	80.67	56.90
w/ time	**75.11**	**82.67**	**82.67**	**60.00**

Table 10: **Time-aware Caption Representation.** Incorporating time information into the captions by adding timestamps depicting time intervals from which the captions were extracted significantly boosts the performance on Video-MMMU.

B More Experimental Results

HourVideo. We evaluate our method on the development set of HourVideo (Chandrasegaran et al., 2024), a benchmark specifically designed for 3D reasoning over long videos. As shown in Table 11, our method performs surprisingly well, despite not incorporating any explicit 3D modeling. Specifically, it achieves an accuracy of 36.3%, outperforming the concurrent method VAMBA (Ren et al., 2025) by a notable margin of **2.7%**. These results demonstrate that our method is effective in reasoning about both the 3D physical world and hour-long videos.

Method	Overall Accuracy
Aria	**39.2**
Gemini 1.5 Pro	37.4
Qwen2-VL 7B	33.8
VAMBA	33.6
SiLVR (ours)	36.3

Table 11: **Performance of Our Method on HourVideo.** SiLVR outperforms the concurrent work VAMBA by a significant **2.7%**.

Detailed VideoMME Results. In Table 12, we present a detailed breakdown of our method's performance on VideoMME. From these results, we observe that our method achieves the lowest accuracy on the Counting Problem. This is likely due to the complexity of counting tasks, which require precise temporal localization of multiple events and subsequent reasoning. Any missed or incorrectly detected events could lead to incorrect answers, making the Counting Problem particularly challenging.

Question Category	Accuracy
Temporal Reasoning (TR)	74.6
Spatial Reasoning (SR)	94.6
Action Reasoning (AR)	76.1
Object Reasoning (OR)	79.5
Temporal Perception (TP)	85.5
Spatial Perception (SP)	74.1
Attribute Perception (AP)	80.2
Action Recognition (AC)	68.1
Object Recognition (OC)	82.5
OCR Problems (OP)	83.5
Counting Problem (CP)	50.7
Information Synopsis (IS)	88.5

Table 12: **Detailed Results on VideoMME.** Our method achieves the highest accuracy on Spatial Reasoning while achieving the lowest performance on the challenging Counting Problem.

C Additional Implementation Details

C.1 Captioner

For all LLaVA and Qwen models, we use the prompt "Briefly describe the video within 40 words" to generate captions for each clip. We set `max_new_tokens` to 200 and employ greedy decoding. We utilize the following model variants from Hugging Face: `lmms-lab/llava-onevision-qwen2-7b-ov` for LLaVA-OV 7B, `Qwen/Qwen2.5-VL-7B-Instruct` for Qwen2.5-VL 7B, and `Qwen/Qwen2.5-VL- 72B-Instruct` for Qwen2.5-VL 72B. For the NVILA model, we use the NVILA-8B-Video variant with the prompt “generate caption” to produce captions for each clip. We set `max_new_tokens` to 128 and employ greedy decoding similar to LLaVA and Qwen models.

We use 4 H100 GPUs for generating captions.

C.2 LLM

We use the default temperature of 1.0 for all LLM experiments. We use the DeepSeek API to run DeepSeek-R1 and DeepSeek-V3 efficiently. To accelerate inference, we implement a parallel processing pipeline with up to 64 concurrent processes, each handling raw captions and subtitles before sending requests to the API. During off-peak hours, this setup allows us to evaluate our method on the complete VideoMME benchmark in under 2 hours at a cost of less than $20. For GPT models, we use the OpenAI API. For Llama-4 models, we use 4×H100 GPUs for local inference. However, we can only process a subset of videos locally due to GPU memory constraints. For long videos, we use Lambda Cloud’s API service.

C.3 Prompt Design

Different VideoQA benchmarks include different types of questions (e.g., multiple-choice, open-ended, or mixed). Additionally, the Grounded VideoQA task in CGBench requires models to predict the temporal boundaries (start/end timestamps) of video segments relevant to each question. To accommodate these differences, we design task-specific prompts. Specifically, for multiple-choice questions, we use the following prompt template:

The video's subtitles are listed below.
`Subtitles`.

The video's captions are listed below. Each caption describes a `Clip Length` seconds clip.
`Captions`.

Select the best answer to the following multiple-choice question based on the video and the subtitles. Respond with only the letter (A, B, C, D, E, etc.) of the correct option.

Question: `Question`.
Options: `Options`.

The answer is:

For open-ended questions, we use the following prompt template:

The video's subtitles are listed below.
`Subtitles`.

The video's captions are listed below. Each caption describes a `Clip Length` seconds clip.
`Captions`.

Based on the video and the subtitles. Answer the following question with one sentence. Answer the following question based on the video and the subtitles. The answer is short. Please directly respond with the short answer.

Question: `Question`.

The answer is:

For the Grounded VideoQA task, we use the following prompt to generate the start and end seconds of the question-related clips:

The video's subtitles are listed below.
`Subtitles`.

The video's captions are listed below. Each caption describes a `Clip Length` seconds clip.
`Captions`.

Your task is to determine in which intervals the clue exists that contain visual information needed to answer the question.

Question: `Question`.

Only output the answer in the following format:
[[start1, end1], [start2, end2], ...]
In this output format, each 'start' and 'end' represents the beginning and end of an interval in seconds (integer) where relevant clues can be found.
You must provide at least one interval and at most five intervals.
Here are some example outputs.
Example 1: [[5, 7]]
Example 2: [[200, 207], [209, 213], [214, 220]]

C.4 Evaluation

We evaluate our method on the validation set of MMVU and HourVideo. For the VideoQA task on CGBench, we adopt the `long-acc` setting in which the model takes the entire long video as the input and answers the given questions.

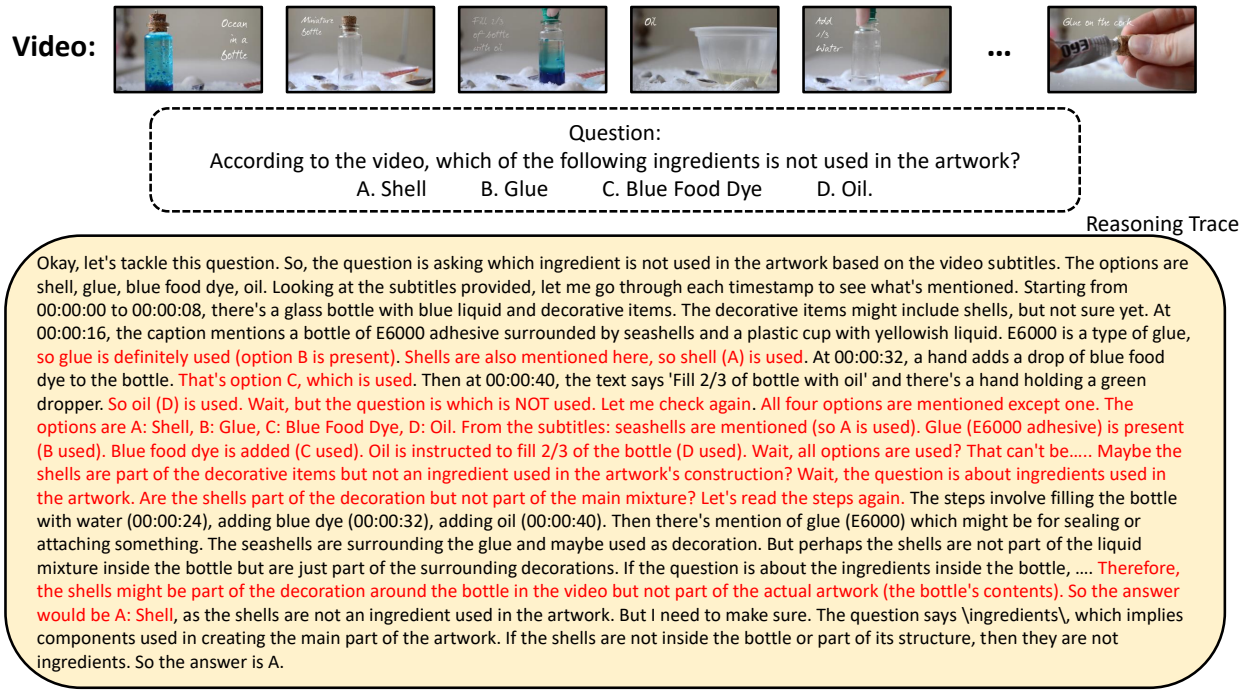

Figure 6: **Example 1 of SiLVR's Reasoning Trace.** The question asks which ingredients are not used in the video. Initially, SiLVR identified all listed items as potential ingredients. However, through self-correction, SiLVR correctly recognized that the shell is used as decoration rather than an ingredient.

We follow the official evaluation code of each benchmark to make a fair comparison with prior methods. If the official evaluation code is not provided, we use the code from LMMs Eval (Zhang et al., 2024a).

For the Grounded VideoQA task on CGBench, we need to decode the timestamps from the LLM outputs. In practice, we found that the Grounded QA prompt as shown in Section C.3 works surprisingly well, with each output sentence following the list format exactly. Therefore, we simply use the `eval` function in Python to decode the output text into a list object. We then use the official code provided by CGBench to compute the mIoU between the prediction and the ground truth.

D Additional Qualitative Results

We present and analyze several of SiLVR's reasoning traces in Figure 6, and Figure 7. From the figures, we observe that SiLVR effectively integrates information from both visual and speech modalities. Furthermore, SiLVR is capable of performing complex video reasoning steps, including step-by-step reasoning, self-correction, self-verification, and using relevant prior knowledge from the LLM to answer a given question.

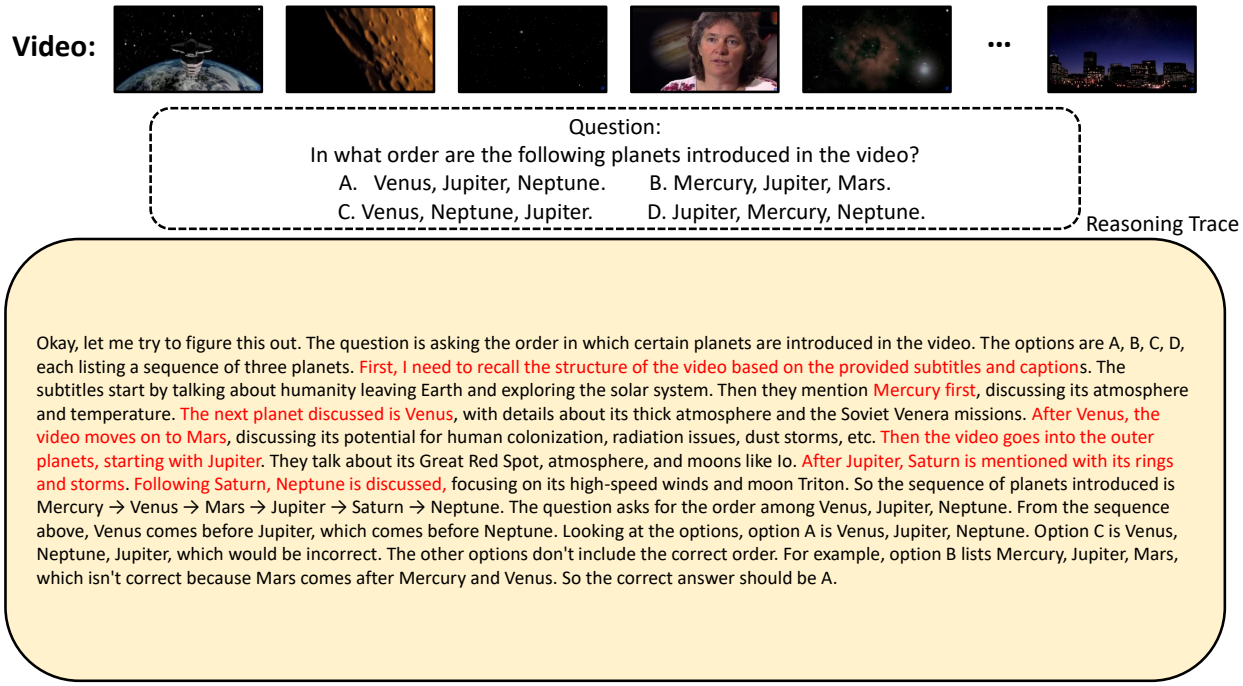

Figure 7: **Example 2 of SiLVR's Reasoning Trace.** The video sequentially introduces six planets in detail: Mercury, Venus, Mars, Jupiter, Saturn, and Neptune. SiLVR accurately identifies the correct order of the planets and systematically inspects all answer choices, eliminating the incorrect ones through logical reasoning.