Video-Holmes: Can MLLM Think like Holmes for Complex Video Reasoning?

Junhao Cheng[1,2], Yuying Ge[1], Teng Wang[1], Yixiao Ge[1], Jing Liao[2], and Ying Shan[1]

[1] ARC Lab, Tencent PCG, China
[2] City University of Hong Kong, Hong Kong SAR, China
https://video-holmes.github.io/Page.github.io/

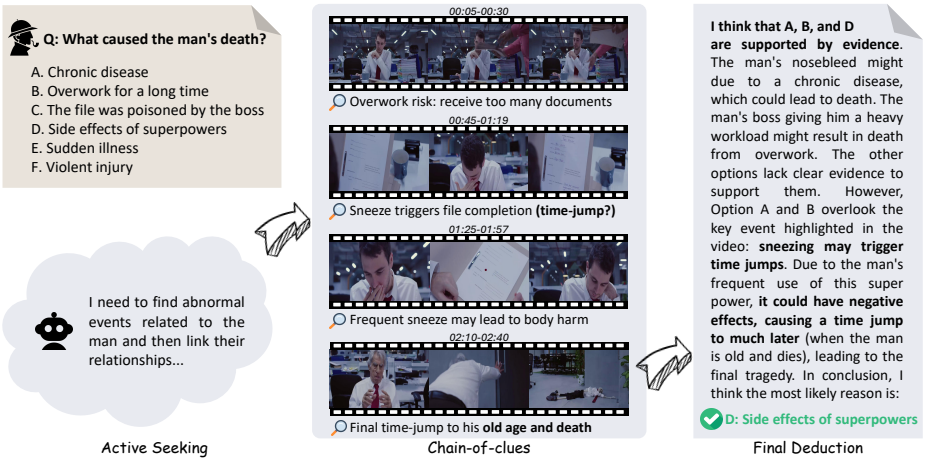

Fig. 1: An example of Video-Holmes. Models are required to actively locate and connect multiple relevant visual clues scattered across video clips to render the final answer.

Abstract. Recent advances in Chain-of-Thought reasoning have been reported to enhance video reasoning capabilities of multimodal large language models (MLLMs). This progress naturally raises a question: *Can these models perform complex video reasoning in a manner comparable to human experts?* However, existing video benchmarks primarily evaluate visual perception and grounding, with questions that can be answered based on explicit prompts or isolated visual cues (e.g., "What is the woman wearing?"). Such benchmarks do not fully capture the intricacies of real-world reasoning, where humans must actively search for, integrate, and analyze multiple clues before reaching a conclusion. To address this, we introduce **Video-Holmes**, a benchmark inspired by the reasoning process of Sherlock Holmes, designed to evaluate the complex video reasoning capabilities of MLLMs. Video-Holmes consists of 1,837 questions derived from 270 manually annotated **suspense short films**, spanning seven carefully designed tasks. Each task is constructed by first identifying key events and causal relationships within films, and then designing questions that require models to actively locate and connect multiple relevant visual clues scattered across video clips. Evaluation of state-of-

the-art MLLMs reveals that while these models generally excel at visual perception, they encounter difficulties with integrating information and often miss critical clues. For example, the best-performing model, Gemini-3.0-Pro, achieves an accuracy of only 49.6%, with most models scoring below 40%. We aim for Video-Holmes to serve as a *"Holmes-test"* for multimodal reasoning, motivating models to reason more like humans and emphasizing the ongoing challenges in this field.

Keywords: Video Reasoning Benchmark · Complex Reasoning

1 Introduction

The development of chain-of-thought (CoT) reasoning [44] and reinforcement learning (RL) post-training strategies [37] has contributed to significant improvements in the reasoning abilities of LLMs [18, 31, 33]. By generating human-like reasoning steps, these models have shown strong performance in addressing complex reasoning tasks. Furthermore, these advancements have been successfully adapted to MLLMs for video reasoning [5,11,14,24,34,43,51,55,56]. This progress naturally raises a question: *Can these models perform complex video reasoning in a manner comparable to human experts?*

However, most existing evaluation benchmarks for video reasoning [8, 9, 19, 21, 23, 25, 27, 28, 36, 52, 59] are limited by their predominant focus on assessing the visual perception and grounding capabilities of models, where questions can be answered based on explicit prompts or isolated visual cues (e.g., "What is the woman wearing?"). Such benchmarks do not fully capture the intricacies of real-world reasoning, where humans must actively search for, integrate, and analyze multiple clues before reaching a conclusion. Empirical results show that even models with advanced thinking abilities [14] achieve only marginal gains (e.g., from 68.3% to 69.4% on Video-MME [12]), which raises doubts about the extent to which these tasks require genuine reasoning.

To address this issue, we introduce Video-Holmes, a benchmark inspired by the reasoning process of Sherlock Holmes. It is designed to assess the complex video reasoning abilities of MLLMs and analyze the underlying factors that drive both their success and failure cases. As demonstrated in Table 1, Video-Holmes differs from existing benchmarks in several key aspects: (1) We utilize suspense short films as the video sources with detailed manual annotations. These videos are characterized by rich elements of suspense, reasoning, and supernatural themes, making them particularly challenging for models to comprehend. (2) The questions in Video-Holmes require models to actively locate and connect multiple relevant visual clues scattered across different video segments to infer the final answer. As illustrated in Figure 1, models first need to identify the abnormal scene involving the man and then progressively integrate the extracted visual clues to deduce the cause of the man's death, just like the reasoning process of Sherlock Holmes. (3) We provide detailed analysis of models' reasoning processes, examining the factors that lead to both correct and incorrect answers.

Table 1: Comparison between Video-Holmes and existing video reasoning benchmarks across several key aspects: the video source domain (**Domain**), annotation methodology (**Anno.**), the number of reasoning QA pairs (**RQA Pairs**), necessity for models to actively seek out clues (**Active Seeking**), necessity for models to link multiple clues (**Chain-of-Clues**), whether provide reasoning process analysis (**RPA**), and whether provide audio information (**Aud.**).

Benchmarks	Domain	Anno.	RQA Pairs	Active Seeking	Chain-of-Clues	RPA	Aud.
VSI-Bench [52]	Indoor scenes	A&M	5,130	✗	✗	✗	✗
MVBench [23]	Open	A	4,000	✗	✗	✗	✗
TempCompass [27]	Open	A&M	7,540	✗	✗	✗	✗
Video-MMMU [21]	Academic	M	900	✗	✗	✗	✓
MMVU [59]	Science	M	3,000	✗	✗	✗	✓
Video-MME [12]	Open	M	1,944	✗	✗	✗	✓
VCR-Bench [36]	Open	A&M	1,034	✗	✗	✓	✗
Video-Holmes (Ours)	Suspense short films	A&M	1,837	✓	✓	✓	✓

Our comprehensive evaluation of state-of-the-art (SOTA) MLLMs reveals that, while these models generally excel at visual perception, they encounter difficulties with integrating information and often miss critical clues. For example, the best-performing model, Gemini-3.0-Pro, achieves an accuracy of only 49.6%, with most models scoring below 40%.

To demonstrate the intrinsic utility and transferability of our high-reasoning-demand data, we provide an additional training set of 233 short films (1,551 questions). While sourced from suspense movies, our data offers a high "reasoning density" that goes beyond its specific domain. Results show that RL post-training on this training set yields cross-domain gains on general video understanding benchmarks (e.g., +2.3% on Video-MME), proving its value as a potent engine for enhancing generalized video understanding in broader scenarios.

We make the following contributions in this work:

- We present Video-Holmes, a benchmark for complex video reasoning. Video-Holmes comprises 270 manually annotated suspense short films, along with 1,837 challenging questions, which require models to actively locate and link multiple relevant visual clues, offering the community a high-quality and challenging video reasoning benchmark.
- We conduct extensive experiments on Video-Holmes to evaluate existing SOTA MLLMs. We diagnose the reasoning processes of these models and observe that while they perform well in visual perception, they face significant challenges in integrating clues and frequently overlook critical clues. These observations provide valuable insights for future research.
- We craft an auxiliary training set to validate the potential of models trained on high-reasoning-demand data to generalize effectively to broader scenarios.

2 Related Work

MLLMs for Video Understanding and Reasoning. The advancements in MLLMs for image understanding and reasoning [4, 13, 60, 61] have enabled their extension to video-based tasks. Methods such as VideoChat [22], Video-ChatGPT [29], CogVLM2Video [20], InternVL [7, 62], LLaVA-Video [58], and Qwen-VL [41, 50] treat videos as sequences of images, allowing large language models (LLMs) to perform video understanding and reasoning. CoT reasoning and RL post-training strategies have been shown to enhance the reasoning capabilities of LLMs [18, 31], and recent works extend these techniques to MLLMs for multimodal reasoning tasks involving images [10, 26, 38–40, 49] or videos [5, 11, 24, 42, 57]. For instance, Video-R1 [11] and VideoChat-R1 [24] employ supervised fine-tuning (SFT) followed by reinforcement learning post-training using Group Relative Policy Optimization (GRPO) [37] based on Qwen-VL-2.5 [2]. These methods achieve higher performance compared to vanilla Qwen-VL-2.5 across diverse reasoning tasks.

Video Reasoning Benchmarks. Early video understanding benchmarks primarily assess model capabilities within specific scenarios. For instance, MSRVTT-QA [48], ActivityNet-QA [53], and NExT-QA [46] focus on fundamental tasks such as action recognition and video question answering. Recently, benchmarks like MMBench [47], TempCompass [27], and MVBench [23] evaluate reasoning over short video clips, while LongVideoBench [45] and Video-MME [12] extend evaluations to longer video sequences. However, these tasks are generally straightforward and do not require complex reasoning. With the success of chain-of-thought (CoT) reasoning, there is increasing interest in advancing video reasoning in more challenging scenarios. Benchmarks such as MMVU [59] and Video-MMMU [19] evaluate reasoning in academic and scientific domains, while VSI-Bench [52] focuses on indoor environments. The recent VCR-Bench [36] introduces a benchmark specifically designed to assess CoT reasoning in video tasks. Despite these developments, such benchmarks [9,25,28] primarily evaluate visual perception and grounding abilities, with questions that can be answered based on explicit prompts or isolated visual cues and do not fully capture the intricacies of real-world reasoning. In contrast, Video-Holmes requires models to actively locate and link multiple relevant visual clues, engaging them in a more complex and demanding reasoning scenario.

3 Video-Holmes

As shown in Figure 2, the construction pipeline of Video-Holmes involves three main steps: video collection and annotation, task definition, and question-answer-explanation generation.

Video Collection and Annotation. Suspense short films serve as an ideal source for evaluating the complex video reasoning capabilities of MLLMs, as they are characterized by compact narratives enriched with hints, plot twists, and supernatural elements. We utilize the keyword "suspense short films" to search

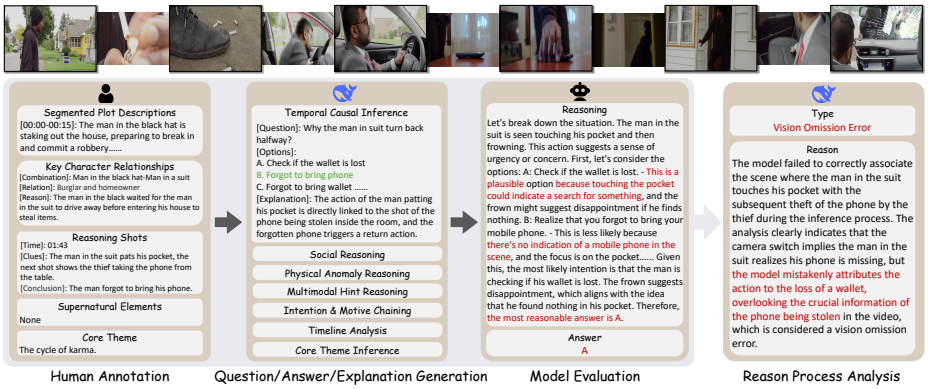

Fig. 2: Construction and evaluation pipeline of Video-Holmes. We first collect and manually annotate suspense short films from YouTube. Next, we design 7 challenging tasks and employ DeepSeek [18] to generate question-answer-explanation triplets. Finally, we benchmark SOTA MLLMs and leverage DeepSeek to analyze their responses.

videos from YouTube with durations between 1 and 5 minutes. We incorporate eight sub-keywords[3] (e.g., "Supernatural") in this process to ensure diversity. From the initial pool of over 2,500 retrieved videos, we performed full manual annotation on the candidates, during which videos were discarded if they: (1) featured overly simplistic plots with low reasoning demand, (2) contained too few visual clues to support deduction, or (3) were of poor production quality. This rigorous pipeline resulted in a final collection of 270 videos for the Video-Holmes test set, and 233 videos reserved as the auxiliary training set. Each video is annotated following a structured template that considers the following aspects:

- **Segmented Plot Descriptions:** Annotators are asked to divide the video into segments based on the storyline, and provide detailed descriptions for each segment.
- **Key Character Relationships:** Annotators are asked to present the relationships between key characters in the video, along with evidence that supports the identification.
- **Reasoning Shots:** Annotators are asked to identify reasoning shots in the video, providing timestamps, visual clues, and the inferred conclusions.
- **Supernatural Elements:** Annotators are asked to specify any supernatural elements present in the videos and the implications they introduce, whether positive or negative.
- **Core Theme:** Annotators are asked to summarize the core themes of the videos based on their overall content.

During the annotation process, we implemented a multi-stage workflow to ensure high-quality video annotations. The procedure consists of the following steps:

[3] Details are provided in Appendix D

1. **Primary Annotation**: Carried out by trained annotators, all of whom hold at least an undergraduate degree.
2. **Quality Verification**: A random sample of 20% of the annotations was reviewed by senior staff. A full rejection policy was enforced: if any annotation was found to be too brief or contained logical flaws, the corresponding annotator was required to resubmit all their assignments until every annotation met our strict standards.

This combination of diverse short films featuring complex reasoning chains, along with high-quality and consistently formatted manual annotations, ensures the reliability and quality of Video-Holmes.

Task Definition. To comprehensively evaluate the differences in MLLMs' capabilities for complex video reasoning from multiple perspectives, we define seven distinct reasoning tasks for Video-Holmes. As illustrated in Figure 3, different from existing benchmarks that are primarily designed around clue-given questions, Video-Holmes focuses on tasks that require models to actively locate and link multiple relevant visual clues scattered across different video segments:

- **Social Reasoning (SR):** Identifying relationships between characters. This includes identifying identity associations across time (e.g., the same man in youth and old age).
- **Physical Anomaly Reasoning (PAR):** Identifying scenes in the video that deviate from real-world norms and reasoning about their underlying rules or implicit meanings.
- **Multimodal Hint Reasoning (MHR):** Decoding cues from multimodal hints, such as semantic implications of camera movements or gradual changes in object positions.
- **Intention & Motive Chaining (IMC):** Observing characters' actions or environmental cues to disentangle surface behaviors from underlying behavioral intentions.
- **Temporal Causal Inference (TCI):** Inferring causal mechanisms between events across time and space using cinematic language and multimodal clues.
- **Timeline Analysis (TA):** Integrating and reconstructing the narrative storyline of the film.
- **Core Theme Inference (CTI):** Extracting the core theme or deeper meaning of the video by analyzing its plot, dialogues, and symbolic elements.

Question-Answer Generation. We leverage DeepSeek-R1 [18] to automatically generate questions based on structured manual annotations and predefined task types. To ensure factuality, each question is derived strictly from the ground-truth annotations. Additionally, the model generates detailed explanations for each answer, which serve as references for analyzing MLLM reasoning processes. Following the automated generation, we conduct a manual audit of the entire dataset. During the initial review, approximately 1% (46 pairs) of the QA pairs are found to be erroneous or contained ambiguous answers. These cases are subsequently regenerated and rectified to ensure that all final questions are accurate and yield unique, unambiguous answers. After this rigorous verification, we construct the test set of 270 videos (1,837 QA pairs) and a training set of 233 videos

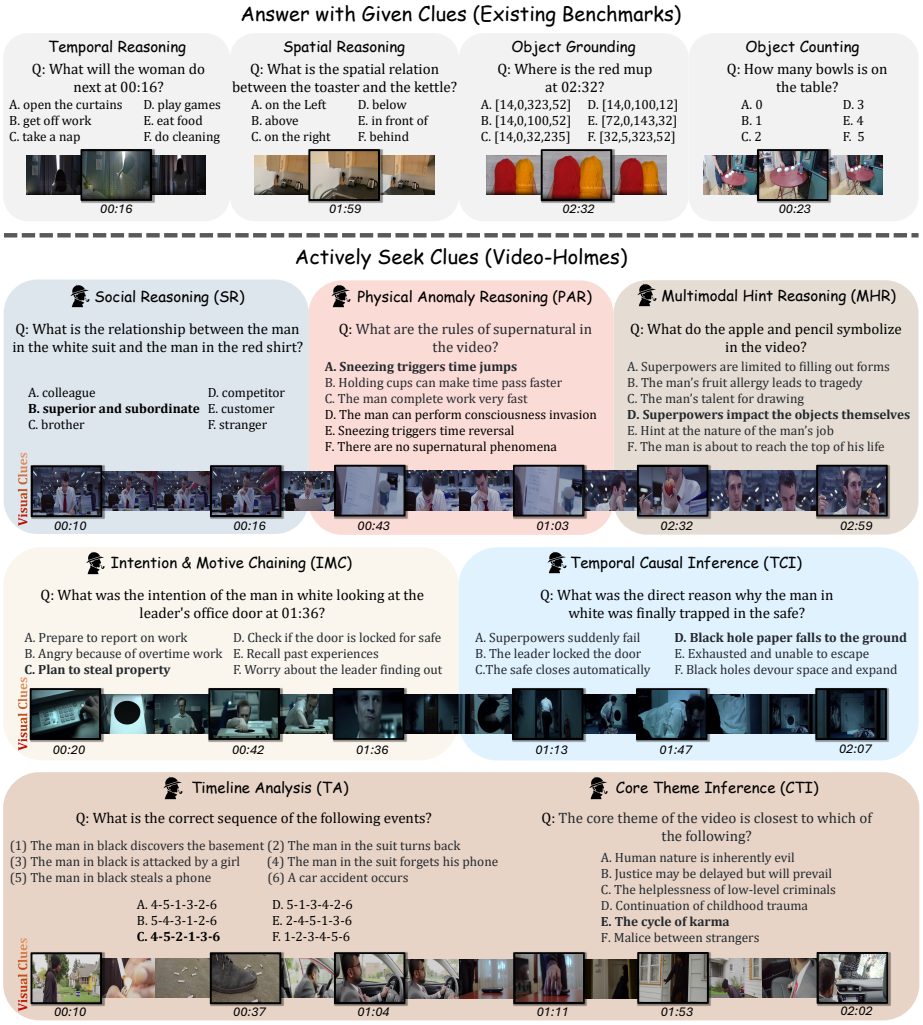

Fig. 3: Comparison of question types between Video-Holmes and existing benchmarks. Existing benchmarks primarily involve clue-given questions, where models depend on explicitly provided clues to derive answers. In contrast, Video-Holmes adopts an active seeking paradigm, requiring models to actively locate and connect multiple relevant visual clues scattered across different video segments.

(1,551 QA pairs) for Video-Holmes. Prompts for pipeline construction and key statistics of Video-Holmes are provided in Appendix B and D, respectively.

4 Experiments

4.1 Setup

Evaluation Models. We conduct an evaluation of several mainstream MLLMs, including the open-source models: InternVL2.5 (8B) [6], InternVL3 (8B) [62], Qwen2.5-VL (7B, 32B) [2], Qwen2.5-Omni (7B) [50] and Qwen3-Omni (30B) [51]. Additionally, we assess open-source models that incorporate RL post-training based on Qwen2.5-VL (7B): GRPO-CARE [5], Video-R1 [11], SpaceR [34], Video-RTS [43] and VideoChat-R1 [24]. We also include several advanced closed-source models in our evaluation: Gemini-2.0-Flash [35], Gemini-2.0-Flash-Thinking [14], Gemini-1.5-Pro [15], Gemini-2.5-Pro [16], Gemini-3.0-Pro [17], GPT-4o [30], OpenAI o4-mini [32], Claude 3.5 & 3.7 Sonnet [1].

Implementation Details. For models with native video input support such as Gemini, videos were processed directly without additional pre-processing. For models lacking native video input capabilities (e.g., GPT-4o), frames were uniformly extracted from the video along with corresponding timestamp, and multi-image input was utilized for evaluation. To ensure a fair comparison, all models were deployed following their official guidelines and using the officially released checkpoints. During inference, models were required to first generate a reasoning process before providing the final answer. Specifically, the models were instructed to produce a step-by-step solution to the given question. For further details regarding model implementation and evaluation prompts, please refer to Appendix A and B.

4.2 Main Results

Table 2 presents the performance of each model on Video-Holmes. Most models achieve an accuracy below 40%, with the best-performing model, Gemini-3.0-Pro, reaching an overall accuracy of 49.6%. The widely-used open-source model, Qwen2.5-VL (7B) achieves an overall accuracy of 27.8%, far worse than its performance on other video reasoning benchmarks. This performance gap suggests that Video-Holmes introduces unique challenges that are particularly demanding for current MLLMs in video reasoning.

Models trained with thinking strategies exhibit notable improvements over their vanilla version. For instance, Gemini-2.0-Flash-Thinking demonstrates a 12.5% performance gain compared to Gemini-2.0-Flash. This observation indicates that the Video-Holmes benchmark imposes substantial reasoning challenges and effectively distinguishes models' reasoning abilities. In contrast, other benchmarks do not reflect this pattern, as illustrated in Table 3.

Model performance across seven reasoning tasks in Video-Holmes remains relatively even, with most models achieving accuracy below 40% for each task. This highlights that each task in Video-Holmes poses substantial challenges, requiring advanced reasoning capabilities from existing methods.

We also establish a human baseline by recruiting 5 evaluators and reporting their average performance across the entire benchmark. The evaluation follows

Table 2: Model performance on Video-Holmes, where **SR** stands for Social Reasoning; **IMC** stands for Intention & Motive Chaining; **TCI** stands for Temporal Causal Inference; **TA** Timeline Analysis; **MHR** stands for Multimodal Hint Reasoning; **PAR** stands for Physical Anomaly Reasoning; **CTI** stands for Core Theme Inference. Blue represents the vanilla model, while Green represents its corresponding thinking version with RL post-training.

Model	Frames	SR	IMC	TCI	TA	MHR	PAR	CTI	Overall
Open Source Models									
InternVL2.5-8B	32	28.0	32.2	21.5	07.7	25.7	23.8	22.6	23.8
InternVL3-8B	32	29.5	40.7	37.9	35.1	24.6	38.9	24.1	32.3
Qwen2.5-Omni-7B	32	27.1	19.9	13.9	7.5	14.8	14.9	13.7	16.4
Qwen3-Omni-30B-A3B	32	52.5	48.1	38.4	55.2	41.6	42.8	44.7	46.2
Qwen2.5-VL-32B	32	43.2	44.2	31.5	51.0	36.4	31.4	32.2	38.4
Qwen2.5-VL-7B	32	38.4	34.8	17.6	30.0	27.1	18.6	25.2	27.8
GRPO-CARE	32	42.8	35.1	25.6	40.5	29.2	29.9	32.6	33.5
VideoChat-R1	32	42.1	38.8	24.5	39.5	29.5	27.8	29.3	33.0
Video-R1	32	48.6	41.7	28.9	34.5	31.0	33.5	35.9	36.5
SpaceR	32	48.2	39.4	26.0	33.0	28.9	35.1	35.6	35.2
Video-RTS	32	51.0	44.6	33.3	43.0	37.1	35.6	38.2	40.0
Closed Source Models									
GPT-4o	32	50.0	49.6	38.8	30.0	44.0	39.2	37.0	42.0
OpenAI o4-mini	32	36.3	31.2	20.5	34.0	30.1	30.9	27.4	29.9
Claude 3.5 Sonnet	20	48.6	43.5	30.8	41.0	39.8	36.6	33.7	39.3
Claude 3.7 Sonnet	20	45.9	48.2	33.7	39.5	40.7	39.7	38.1	41.0
Gemini-1.5-Pro	-	52.1	48.2	34.4	26.0	39.2	46.4	38.9	41.2
Gemini-2.5-Pro	-	46.6	49.3	46.9	53.0	40.1	44.3	37.4	45.0
Gemini-3.0-Pro	-	53.2	51.8	48.5	54.0	43.5	49.8	46.4	49.6
Gemini-2.0-Flash	-	41.8	33.7	23.1	20.5	30.1	26.8	33.7	30.6
Gemini-2.0-Flash-Thinking	-	43.4	46.9	43.1	51.0	37.9	43.6	39.3	43.1
Human	single-view	93.5	96.3	99.6	97.9	94.9	92.6	96.7	96.3

a strictly controlled single-view protocol: participants are instructed to read the questions beforehand, watch the short film exactly once, and then provide their answers. As shown in Table 2, humans achieve an impressive accuracy of 96.3%, significantly outperforming all SOTA MLLMs. Notably, the fact that humans do not achieve a perfect score under this single-view constraint—unlike the annotation phase, which mandates repeated reviewing to capture all nuances—underscores the profound complexity of the tasks. The result validates both the reliability of our ground-truth annotations and the inherent difficulty of Video-Holmes.

Table 3: Performance gains of reasoning models across benchmarks.

Model	MMVU	Video-MME	Video-MMMU	MVBench	VCR-Bench	MORSE-500	**Video-Holmes**
Gemini-2.0-Flash	65.2	68.3	54.3	55.3	51.7	16.0	30.6
Gemini-2.0-Flash-Thinking	64.5 (-0.7)	69.4 (+0.9)	56.4 (+1.9)	57.4 (+2.1)	54.1 (+2.4)	19.2 (+3.2)	43.1 **(+12.5)**

Table 4: Reasoning process analysis results. Where **VPE** represents visual perception error, **VOE** represents visual omission error, **RE** represents reasoning error, **TRAW** represents think right answer wrong, **TWAR** represents think wrong answer right, and **TRAR** represents think right answer right.

Model	Answer Wrong								Answer Right			
	VPE		VOE		RE		TRAW		TRAR		TWAR	
	Count	Ratio	Count	Ratio	Count	Ratio	Count	Ratio	Count	Ratio	Count	Ratio
Open Source Models												
InternVL2.5-8B	66	0.06	331	0.28	740	0.63	32	0.03	332	0.79	89	0.21
InternVL3-8B	32	0.03	248	0.22	810	0.73	22	0.02	419	0.76	135	0.24
Qwen2.5-Omni-7B	26	0.02	398	0.36	643	0.59	29	0.03	236	0.82	53	0.18
Qwen2.5-VL-32B	44	0.04	430	0.39	610	0.56	6	0.01	363	0.78	103	0.22
Qwen2.5-VL-7B	15	0.02	209	0.27	544	0.70	5	0.01	457	0.87	67	0.13
GRPO-CARE	32	0.03	325	0.30	688	0.63	54	0.05	543	0.80	133	0.20
VideoChat-R1	33	0.03	306	0.26	844	0.71	12	0.01	494	0.85	88	0.15
Video-R1	46	0.04	415	0.36	681	0.59	15	0.01	504	0.83	101	0.17
Closed Source Models												
GPT-4o	17	0.02	250	0.24	745	0.73	13	0.01	720	0.94	43	0.06
o4-mini	36	0.03	354	0.29	840	0.68	11	0.01	513	0.94	34	0.06
Claude 3.5 Sonnet	18	0.02	318	0.31	695	0.67	11	0.01	687	0.92	59	0.08
Claude 3.7 Sonnet	9	0.01	470	0.44	584	0.54	10	0.01	612	0.86	103	0.14
Gemini-1.5-Pro	75	0.07	250	0.24	692	0.67	18	0.02	637	0.85	114	0.15
Gemini-2.5-Pro	31	0.03	210	0.22	718	0.75	4	0.01	725	0.88	98	0.12
Gemini-3.0-Pro	22	0.02	185	0.20	710	0.77	9	0.01	865	0.95	46	0.05
Gemini-2.0-Flash	33	0.03	242	0.20	893	0.76	13	0.01	542	0.89	65	0.11
Gemini-2.0-Flash-Thinking	31	0.03	210	0.22	718	0.75	4	0.01	774	0.94	49	0.06

4.3 Analytical Study

Reasoning Process Analysis. We diagnose the factors contributing to the model's answers by comparing its reasoning process with human-annotated descriptions and answer explanations. Specifically, we categorize the main causes of incorrect answers into the following four types:

- **Visual Perception Error (VPE):** The model extracts incorrect visual information, leading to wrong answers.
- **Visual Omission Error (VOE):** The model omits critical visual information (i.e., key objects or events), resulting in an incorrect answer.
- **Reasoning Error (RE):** The model makes errors during the reasoning process, such as incorrectly associating multiple visual clues.
- **Think Right Answer Wrong (TRAW):** The model's reasoning is largely aligned with the ground-truth explanation, but it answers wrong.

For correctly answered questions, we define the following two categories:

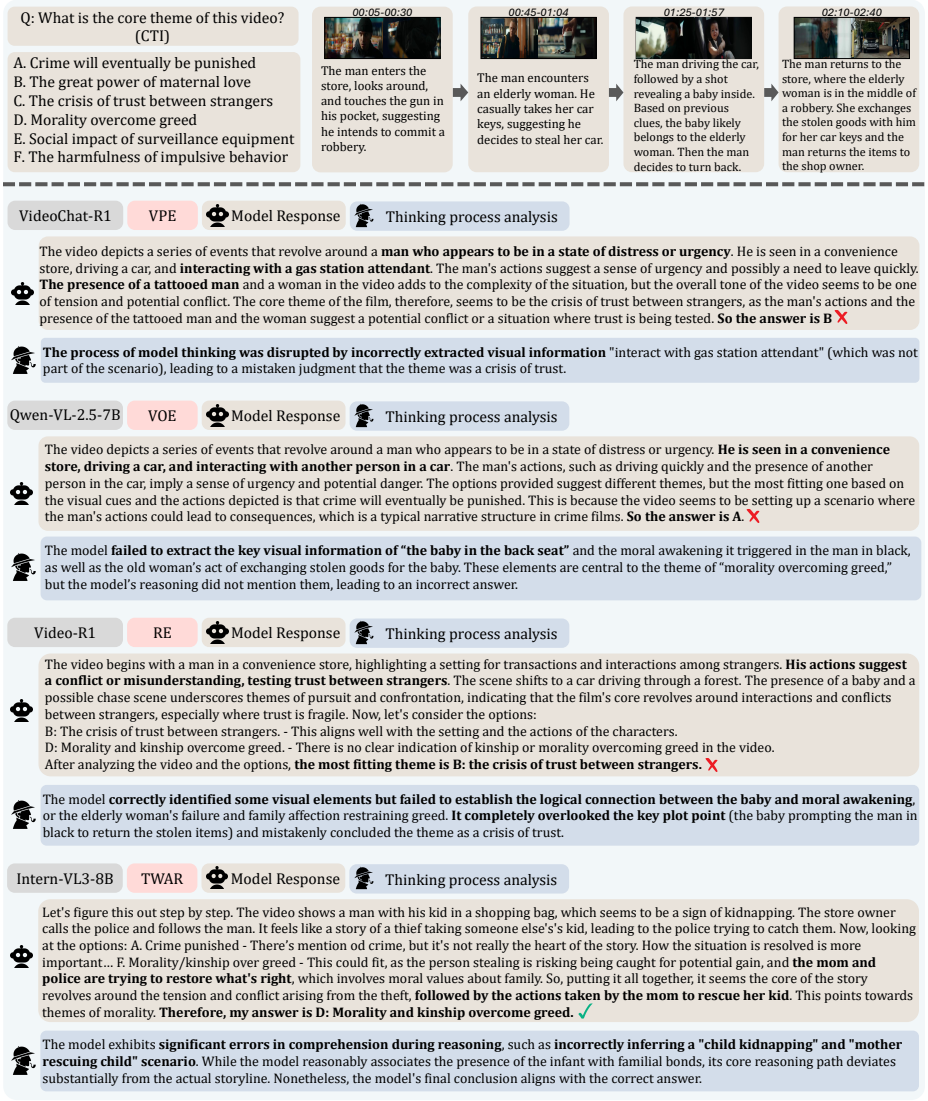

Fig. 4: Example of model reasoning process analysis. VideoChat-R1 misinterprets visual information, incorrectly perceiving a tattooed man, Qwen2.5-VL overlooks critical visual clues (the baby), Video-R1 fails to establish logical connections between the visual clues, and Intern-VL3 guesses the right option with a wrong reasoning process.

– **Think Wrong Answer Right (TWAR):** The model's reasoning process deviates significantly from the ground-truth explanation, yet it arrives at the correct answer.

Table 5: Analyze experiment results on Video-Holmes. **HA** stands for human annotation; **FLC** stands for frame-level caption; **VLC** stands for video-level caption.

(a) Number of input frames

Model	Frames	Acc
Qwen2.5-VL-7B	64	30.2 (+2.4)
Qwen2.5-VL-7B	80	33.0 (+5.2)
Qwen2.5-VL-7B	fps=1	42.5 (+14.9)
Qwen2.5-Omni-7B	fps=1	43.9 (+21.6)
Video-R1	64	37.4 (+0.9)
Video-R1	80	38.5 (+2.0)
Video-R1	fps=1	40.1 (+3.6)

(b) Audio input

Model	Audio	SR	Overall
Qwen2.5-Omni-7B	✗	27.1	16.4
Qwen2.5-Omni-7B	✓	38.4	24.4
Gemini-2.5-Pro	✗	46.6	45.0
Gemini-2.5-Pro	✓	54.8	51.3
Gemini-1.5-Pro	✗	52.1	41.2
Gemini-1.5-Pro	✓	59.6	45.7

(c) Reasoning or not

Model	Frames	Reasoning	Acc
Qwen2.5-VL-7B	32	✓	27.8
Qwen2.5-VL-7B	32	✗	29.4
Video-R1	32	✓	36.5
Video-R1	32	✗	28.2
Gemini-2.0-Flash	-	✓	30.6
Gemini-2.0-Flash	-	✗	28.5

(d) Text-only input

Model	Input	Acc
DeepSeek-R1	FLC	31.2
DeepSeek-R1	VLC	64.6
DeepSeek-R1	HA	92.0
OpenAI o3	FLC	25.4
OpenAI o3	VLC	61.3
OpenAI o3	HA	89.7

– **Think Right Answer Right (TRAR):** The model's reasoning process is largely aligned with the ground-truth explanation and answers right.

We provide the human annotations, questions, models' reasoning outputs (if validated), and the type definitions as inputs to DeepSeek, prompting[4] it to perform the analysis.

The results in Table 4 and Figure 4 show that both open-source and closed-source models generally demonstrate the ability to accurately extract visual information and provide answers consistent with their reasoning processes. Approximately 35% of errors are attributed to the omission of critical visual information, while a larger proportion (around 60%) stems from challenges in logical comprehension of multiple visual clues (Reasoning Errors).

For correctly answered questions, the proportion of responses based on valid reasoning (TRAR) exceeds 80% across most models. This highlights the difficulty of the Video-Holmes benchmark, where models struggle to infer correct answers through inconsistent reasoning.

Number of Input Frames. In main experiments, we adopt a 32-frame input to align with standard practices [12, 21] and ensure a fair comparison, as models have varying pre-training constraints. For example, Video-R1 is trained on 16-32 frames, which constrains its performance with larger inputs. As shown in Table 5(a), increasing the input frames for Video-R1 from 32 to 1 fps only yields a marginal performance improvement of 3.6%. In contrast, Qwen2.5-VL, which is trained with more frames, achieves a much higher gain of 14.9% under the same conditions. Given that Video-Holmes features dense clues, it stands to benefit from longer sequences. To verify this, we conduct experiments using

[4] Detailed in Appendix B

inputs with more frames at 1 fps on Qwen2.5-VL and Qwen2.5-Omni. The results in Table 5(a) confirm that this setup leads to substantial performance gains.

Audio Input. We evaluate the performance of several models with audio input as an additional modality. Table 5(b) demonstrates that integrating audio input enhances model performance, especially in social reasoning tasks where conversational cues offer essential insights into interpersonal dynamics. These results underscore the importance of audio information in multimodal reasoning.

Reasoning or Not. We conduct experiments where models directly generate answers without using CoT prompts. The results in Table 5(c) demonstrate that for stronger closed-source models, CoT prompting leads to higher accuracy compared to directly answering. In contrast, weaker open-source models exhibit the opposite trend. This suggests that the effectiveness of reasoning is contingent on the model's overall capability—only models with sufficiently strong reasoning abilities can fully benefit from CoT prompts. Conversely, weaker models may amplify errors during CoT.

Text-only Input. We conduct experiments using three types of text-only inputs for advanced reasoning models [18, 33]: (1) human-annotated movie plots, (2) frame-level captions generated by Qwen2.5-VL , and (3) video-level captions generated by Gemini-2.5-Pro. As shown in Table 5(d), models achieve around 90% accuracy with human annotations, whereas performance drops significantly with auto-generated captions. This gap arises because human annotations explicitly provide logical connections between critical visual clues. In contrast, frame-level captions often provide redundant and fragmented descriptions of static visuals, missing essential temporal dynamics. Although video-level captions outperform frame-level ones by better capturing event continuity, they fall short of human annotations due to the captioning model's own reasoning limitations. The results highlight the challenge of Video-Holmes in requiring models to jointly perceive and reason about logical relationships among visual clues.

Training & Generalization. We further investigate whether the complex reasoning required by Video-Holmes can enhance general video understanding. Through RL post-training of Qwen2.5-VL-7B on our auxiliary training set (233 videos, 1,551 questions) following the standard protocol from [5], we observe notable performance gains across different domains. As shown in Table 6, the model not only achieves a significant 12.9% improvement on Video-Holmes, but also demonstrates consistent boosts across established general video benchmarks, ranging from 1.8% on TempCompass to 4.8% on MVBench. This evidence confirms that our high-reasoning-demand data fosters transferable skills rather than merely overfitting to domain-specific artifacts, making it both highly challenging and valid for evaluating MLLMs' core reasoning capacities.

4.4 Human Studies

Justification of Reasoning Analysis. To verify the reliability of the automated reasoning analysis, we conduct a human study on 100 randomly sampled model responses. 10 human evaluators manually categorized these responses into the defined error and success types. As shown in Table 7, DeepSeek-R1 achieves a

Table 6: Performance on general video benchmarks. * denotes model post-trained on the training set of Video-Holmes via [5].

Model	Video-Holmes	Video-MME	TempCompass	Video-MMMU	MVBench
Qwen2.5-VL-7B	27.8	56.6	72.6	48.1	59.0
Qwen2.5-VL-7B*	40.7 (+12.9)	58.9 (+2.3)	74.2 (+1.8)	50.2 (+2.1)	63.8 (+4.8)

Table 7: Human-DeepSeek agreement across different reasoning analysis types.

Category	Answer Wrong				Avg. (W)	Answer Right		Avg. (R)	Total Avg.
Type	VPE	VOE	RE	TRAW		TWAR	TRAR		
Agreement (%)	94.1	91.5	90.2	93.3	94.0	96.0	96.0	96.0	95.0
Cohen's Kappa (κ)	0.89	0.86	0.82	0.88	0.88	0.91	0.94	0.92	0.90

Table 8: Human evaluation of different benchmarks. **Active Seeking** denotes the percentage of questions requiring models to actively search for visual evidence rather than relying on predefined text clues. **Chain-of-Clues** denotes the percentage of questions requiring multi-step logical connections. **Difficulty** is rated on a scale of 1 to 5.

Benchmark	Active Seeking (%)	Chain-of-Clues (%)	Difficulty (1-5)
Video-MME	6.0	18.0	1.2
VCR-Bench	24.0	46.0	3.2
Video-Holmes	**98.0**	**96.0**	**4.6**

high agreement rate of 94.0% for incorrect answers and 96.0% for correct answers, with an overall Cohen's Kappa of 0.90. This high consistency stems from the rich context provided in our prompts (e.g., complete human-annotated video plot, the question, and the detailed answer rationale), which effectively transforms this reasoning evaluation into a straightforward information-matching task.

Evaluation of Benchmark Difficulty. We conduct a human study to further validate the challenge of our benchmark. We randomly select 100 questions each from Video-Holmes, VCR-Bench, and Video-MME. 10 human evaluators are asked to assess these questions based on three criteria: (1) whether solving the question requires *Active Seeking* of visual evidence or merely relies on *Predefined Clues*; (2) whether it requires a *Chain-of-Clues* reasoning process; and (3) the overall difficulty on a scale from 1 to 5.

As shown in Table 8, Video-Holmes surpasses other benchmarks across all complexity metrics. Specifically, 98.0% of the questions in Video-Holmes require active seeking, compared to 24.0% in VCR-Bench and only 6.0% in Video-MME. Furthermore, Video-Holmes achieves the highest overall difficulty score of 4.6. This analysis confirms that Video-Holmes successfully forces models to actively search for and logically connect multiple pieces of evidence, making it a demanding and more realistic testbed.

5 Conclusion and Copyright

Conclusion. We present Video-Holmes, a benchmark designed to evaluate the complex video reasoning capabilities of MLLMs. Video-Holmes consists of 1,837 questions derived from 270 manually annotated suspense short films, spanning seven carefully designed tasks that require models to actively locate and link multiple relevant visual clues scattered across different video segments. We conduct a detailed analysis of model reasoning processes, examining the factors that lead to both correct and incorrect answers. Our evaluation of SOTA MLLMs reveals that, while these models excel at visual perception, they encounter substantial difficulties with reasoning and often miss critical clues. We aim for Video-Holmes to serve as a *"Holmes-test"* for multimodal reasoning, motivating models to reason more like humans and emphasizing the ongoing challenges in this field.

References

1. Anthropic: The claude 3 model family: Opus, sonnet, haiku (2024)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
4. Cao, Q., Cheng, J., Liang, X., Lin, L.: Visdiahalbench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12161–12176 (2024)
5. Chen, Y., Ge, Y., Wang, R., Ge, Y., Cheng, J., Shan, Y., Liu, X.: Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. arXiv preprint arXiv:2506.16141 (2025)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
8. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495 (2025)
9. Deng, A., Yang, T., Yu, S., Spencer, L., Bansal, M., Chen, C., Yeung-Levy, S., Wang, X.: Scivideobench: Benchmarking scientific video reasoning in large multimodal models. arXiv preprint arXiv:2510.08559 (2025)
10. Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv preprint arXiv:2503.17352 (2025)
11. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
12. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
13. Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. arXiv preprint arXiv:2310.01218 (2023)
14. Google: Gemini-2.0-flash-thinking (2024)
15. Google: Gemini-2.0-pro (2025)
16. Google: Gemini-2.5-pro (2025)
17. Google: Gemini-3.0-pro (2025)
18. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
19. He, X., Feng, W., Zheng, K., Lu, Y., Zhu, W., Li, J., Fan, Y., Wang, J., Li, L., Yang, Z., et al.: Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. arXiv preprint arXiv:2406.08407 (2024)

20. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500 (2024)
21. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826 (2025)
22. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
23. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., 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 (2024)
24. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
25. Li, X., Li, X., Hu, S., Huang, K., Zhang, W.: Causalstep: A benchmark for explicit stepwise causal reasoning in videos. arXiv preprint arXiv:2507.16878 (2025)
26. Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., Deng, Z.: Improved visual-spatial reasoning via r1-zero-like training. arXiv preprint arXiv:2504.00883 (2025)
27. Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., Hou, L.: Tempcompass: Do video llms really understand videos? arXiv preprint arXiv:2403.00476 (2024)
28. Liu, Y., Ouyang, K., Wu, H., Liu, Y., Sui, L., Li, X., Zhong, Y., Charles, Y., Zhou, X., Sun, X.: Videoreasonbench: Can mllms perform vision-centric complex video reasoning? arXiv preprint arXiv:2505.23359 (2025)
29. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
30. OpenAI: Hello gpt-4o (2024), https://openai.com/index/hello-gpt-4o/
31. OpenAI: Introducing openai o1 (2024)
32. OpenAI: o4-mini (2025)
33. OpenAI: Openai o3 (2025)
34. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
35. Pichai, S., Hassabis, D., Kavukcuoglu, K.: Introducing gemini 2.0: our new ai model for the agentic era (2024)
36. Qi, Y., Zhao, Y., Zeng, Y., Bao, X., Huang, W., Chen, L., Chen, Z., Zhao, J., Qi, Z., Zhao, F.: Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. arXiv preprint arXiv:2504.07956 (2025)
37. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
38. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
39. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)

40. Thawakar, O., Dissanayake, D., More, K., Thawkar, R., Heakl, A., Ahsan, N., Li, Y., Zumri, M., Lahoud, J., Anwer, R.M., et al.: Llamav-o1: Rethinking step-by-step visual reasoning in llms. arXiv preprint arXiv:2501.06186 (2025)
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
42. Wang, Y., Xu, B., Yue, Z., Xiao, Z., Wang, Z., Zhang, L., Yang, D., Wang, W., Jin, Q.: Timezero: Temporal video grounding with reasoning-guided lvlm. arXiv preprint arXiv:2503.13377 (2025)
43. Wang, Z., Yoon, J., Yu, S., Islam, M.M., Bertasius, G., Bansal, M.: Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 28114–28128 (2025)
44. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
45. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
46. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
47. Xu, C., Hou, X., Liu, J., Li, C., Huang, T., Zhu, X., Niu, M., Sun, L., Tang, P., Xu, T., et al.: Mmbench: Benchmarking end-to-end multi-modal dnns and understanding their hardware-software implications. In: 2023 IEEE International Symposium on Workload Characterization (IISWC). pp. 154–166. IEEE (2023)
48. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM Multimedia (2017)
49. Xu, G., Jin, P., Hao, L., Song, Y., Sun, L., Yuan, L.: Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440 (2024)
50. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
51. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
52. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. arXiv preprint arXiv:2412.14171 (2024)
53. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)
54. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. Advances in neural information processing systems (2021)
55. Zhang, C., Wang, Z., Ma, Y., Peng, J., Wang, Y., Zhou, Q., Song, J., Zheng, B.: Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. arXiv preprint arXiv:2509.23652 (2025)
56. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025)

57. Zhang, X., Wen, S., Wu, W., Huang, L.: Tinyllava-video-r1: Towards smaller lmms for video reasoning. arXiv preprint arXiv:2504.09641 (2025)
58. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
59. Zhao, Y., Xie, L., Zhang, H., Gan, G., Long, Y., Hu, Z., Hu, T., Chen, W., Li, C., Song, J., et al.: Mmvu: Measuring expert-level multi-discipline video understanding. arXiv preprint arXiv:2501.12380 (2025)
60. Zheng, K., He, X., Wang, X.E.: Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239 (2023)
61. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
62. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)