VIDEOGAIA: A BENCHMARK FOR GENERAL AI ASSISTANTS ON AGENTIC VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video understanding is a fundamental task for evaluating the capabilities of multimodal large language models (MLLMs). However, existing leading models have already achieved approximately 90% accuracy on the Video-MME leaderboard, suggesting that conventional single-turn video understanding tasks are becoming increasingly saturated and insufficient for assessing the intelligence of advanced MLLMs. Towards this end, we introduce **VideoGAIA**, an agentic **Video** understanding benchmark for **G**eneral artificial intelligence (**AI**) **A**ssistants. Moving beyond one-shot video question answering, VideoGAIA formulates video understanding as a multi-turn, tool-augmented interaction process, where models must iteratively perceive videos, invoke external tools, gather complementary information, and integrate multimodal evidence across turns. VideoGAIA contains 271 model-human co-designed tasks covering diverse and complex real-world scenarios. Each video-question-answer instance is independently verified by three human experts to ensure both correctness and appropriate difficulty. All evaluated MLLMs, including frontier models such as GPT-5.5 and Kimi-K3, achieve less than 60% accuracy on VideoGAIA, highlighting its value as a high-quality and timely benchmark for evaluating next-generation MLLMs. We hope that VideoGAIA will facilitate the transition from conventional video understanding toward **agentic video understanding**. Code is available at https://anonymous.4open.science/r/VideoGAIA-anonymous.

1 INTRODUCTION

Video understanding is a fundamental task in multimodal understanding and has long served as a foundation for evaluating multimodal large language models (MLLMs). Over the past several years, numerous benchmarks (Wu et al., 2024; Zhao et al., 2025; Li et al., 2024) have been proposed to assess model capabilities in multimodal perception, temporal reasoning, and object recognition.

Despite their success, the rapid progress of foundation models has driven the performance of existing video understanding benchmarks close to saturation. On widely adopted benchmarks such as Video-MME (Fu et al., 2025), VideoMMMU (Hu et al., 2026), and MLVU (Zhou et al., 2024), the strongest MLLMs (Qwen Team, 2026c; OpenAI, 2026b; Google, 2026a) now achieve accuracies approaching 90%, suggesting that conventional single-turn video question answering has become a nearly solved problem. This trend calls for a new evaluation paradigm that better reflects the capabilities required by next-generation multimodal systems. Meanwhile, the emergence of multimodal agents is reshaping how artificial intelligence (AI) systems interact with the world, yet the development and evaluation of video agents remain largely underexplored. Towards this end, we introduce **VideoGAIA**, an agentic video understanding benchmark for general AI assistants. Going beyond traditional single-turn video understanding, VideoGAIA provides a comprehensive evaluation suite that measures the ability of video agents to iteratively perceive videos, engage in multi-turn interactions, invoke external tools, and integrate information from multiple sources to solve complex real-world tasks (shown in Figure 1). Even with access to additional tools, the frontier model Qwen3.7-Plus (Qwen Team, 2026c) still achieves substantially lower accuracy on VideoGAIA than on other video understanding benchmarks.

To design a video agent benchmark that addresses the most pressing needs of the community, we first ask a fundamental question: *what are contemporary MLLMs already good at, and where do they still fall short?* On the one hand, continued scaling in model size and training data has endowed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

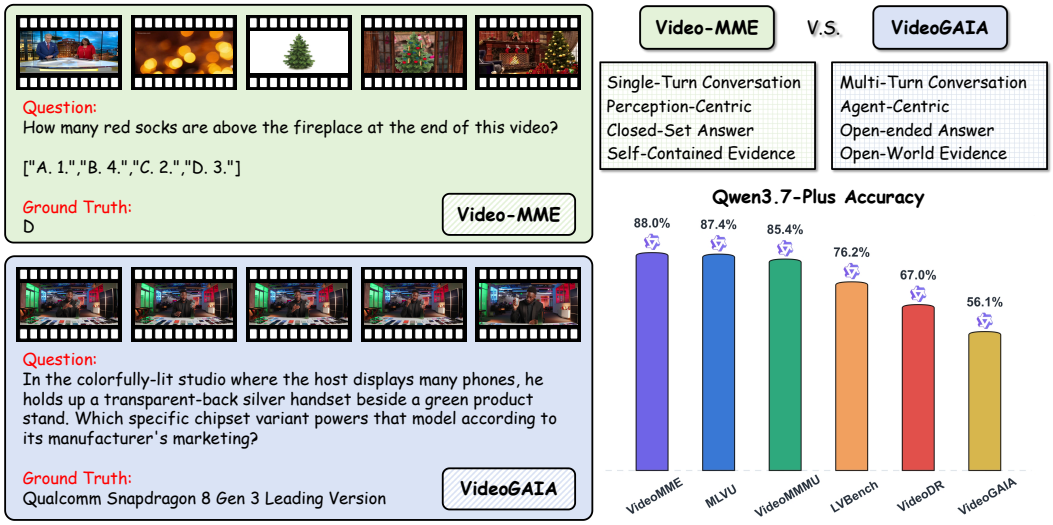
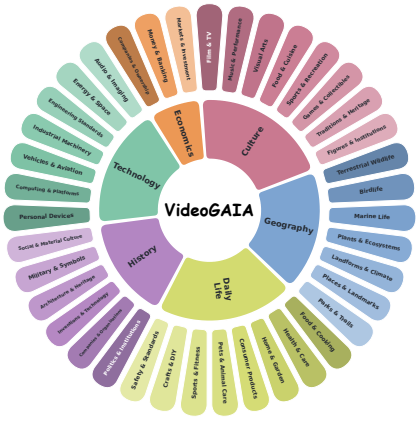


Figure 1: VideoGAIA focuses on agentic video understanding, which is more challenging compared with conventional video understanding benchmarks.

frontier models with strong general capabilities and extensive world knowledge (Qwen Team, 2025; Kimi Team, 2026b). They have become highly proficient in foundational tasks such as instruction following, visual recognition, and temporal understanding, which largely explains the saturation of single-turn video understanding benchmarks. On the other hand, foundation models are undergoing a broader transition from conversational chatbots to general-purpose productivity tools (OpenAI, 2026b; Z.ai, 2026; Luo et al., 2025; Zhang et al., 2026c). Capabilities essential to this transition, including tool use, long-horizon reasoning, information aggregation, and context management, remain under active development and are far from fully mature. Video understanding should therefore evolve accordingly for the agentic era. Its evaluation should no longer be limited to asking a model to watch a video and answer a single question passively. Instead, a capable video agent should jointly interpret textual instructions and visual-temporal evidence, identify information that cannot be obtained from the video alone, proactively use available tools to seek complementary information, and efficiently synthesize the collected evidence across multiple interaction steps. As shown in the left part of Figure 1, the ultimate goal of a video benchmark is not merely to test whether a model can answer questions about a video, but whether it can use videos as a source of evidence to accomplish concrete, real-world tasks. This perspective forms the central design principle of VideoGAIA and motivates our shift from conventional video question answering toward agentic video understanding.

The construction of VideoGAIA follows a multi-stage human-model co-design pipeline. We begin with a pool of 100K publicly available videos collected from the Internet and employ frontier models, including Gemini-3.1-Pro (Google, 2026a) and GPT-5.5 (OpenAI, 2026b), to support annotation and quality-control stages. These stages include frame-level video captioning, evidence-grounded task generation, rubric-based quality control, cross-modal filtering, tool-necessity assessment, and human crowdsourced annotation. Each model-filtered task is independently reviewed by at least three human experts. Only tasks that receive unanimous approval, or majority approval followed by an additional final review, are retained. The resulting task pool spans six task categories, including Culture, Geography, Daily Life, History, Technology, and Economics, and is further divided into two difficulty levels, L1 and L2. After more than 100 hours of annotation by 15 annotators, this process ultimately yields 271 high-quality video-agent tasks that constitute the VideoGAIA test set.

Following data curation, we conduct a comprehensive evaluation of 20 cutting-edge frontier MLLMs. To ensure a fair comparison, all models are evaluated under the same experimental setting using a ReAct-style (Yao et al., 2022) agent loop equipped with three tools: web search for retrieving external information, page visit for converting webpages into textual content, and thinking with videos for revisiting and inspecting relevant video segments. The results show that all evaluated models achieve accuracies between 30% and 60%. Even the strongest model Seed2.0-Pro (Seed, 2026) reaches only 58.30% accuracy, substantially lower than the performance on conventional video understanding



(a) Real-World Tasks

Rank	Model	Accuracy
1	Seed2.0-Pro	58.30%
2	Qwen3.7-Plus	56.09%
3	Kimi-K3	54.61%
4	GPT-5.5	53.14%
5	Kimi-K2.6	52.77%
6	Qwen3.5-Plus	52.03%
6	GPT-5.4	52.03%
8	Gemini-3.1-Pro	50.92%
8	GPT-5.2	50.92%
10	GLM-5V-Turbo	50.18%

Rank	Model	Accuracy
10	Qwen3.5-397B	50.18%
10	Qwen3.8-Max	50.18%
13	Kimi-K2.5	49.82%
14	Qwen3.6-Plus	49.08%
15	MiMo-V2.5	48.71%
16	Claude-Opus-4.6	47.97%
17	GLM-4.6V	46.86%
18	Gemini-3.5-Flash	43.91%
19	Claude-Opus-4.8	36.90%
20	Claude-Opus-4.7	34.69%

(b) Leaderboard

Figure 2: **(a) Real-World Tasks.** VideoGAIA spans 6 key domains and 39 subtasks. **(b) Leaderboard**. Performance comparison of 20 frontier MLLMs is shown under the agent loop setting.

benchmarks. These results in the leaderboard (as shown in Figure 2) indicate that agentic video understanding remains far from solved and presents substantial opportunities for future research. We hope it will encourage further research exploration of model architectures, training strategies, and agent harnesses, ultimately enabling video agents to genuinely understand videos, assist users in completing real-world tasks, and deliver meaningful productivity gains.

To summarize, the main contributions of this work are threefold:

- **Paradigm Shift.** We advocate a transition from single-turn video understanding to agentic video understanding. This new paradigm encourages video agents to move beyond answering questions about videos and instead understand video content as evidence for accomplishing concrete tasks.
- **High-quality Benchmark.** We introduce **VideoGAIA**, a carefully curated benchmark for agentic video understanding that contains 271 human-verified video-agent tasks spanning six real-world categories. Each task is constructed through a multi-stage human-model co-design pipeline and independently reviewed by at least three human experts, ensuring strong correctness, appropriate difficulty, and genuine requirements for multimodal reasoning and tool use.
- **Comprehensive Evaluation.** We systematically evaluate 20 cutting-edge frontier MLLMs under a unified agentic setting. All evaluated models achieve less than 60% accuracy, substantially below the peak performance reported on conventional video understanding benchmarks. These results reveal the limitations of current video agents and highlight substantial room for future development.

2 Related Work

2.1 Video Understanding Benchmarks

Existing video understanding benchmarks evaluate capabilities ranging from temporal perception and long-context comprehension to multimodal reasoning and domain knowledge (Zhang et al., 2026a; Wang et al., 2024a; Yang et al., 2024; Fang et al., 2024). MVBench (Li et al., 2024) organizes diagnostic tasks around temporal understanding, while Video-MME (Fu et al., 2025) covers videos of diverse durations, domains, and modalities. MLVU (Zhou et al., 2024) and LongVideoBench (Wu et al., 2024) focus on long-video comprehension, whereas MMVU (Zhao et al., 2025) and VideoMMMU (Hu et al., 2026) emphasize knowledge-intensive and discipline-oriented reasoning. Despite their diversity, most benchmarks adopt a closed-world, single-turn protocol in which models answer predefined questions based on a fixed video without actively determining what evidence to inspect or consulting external resources. VideoGAIA complements this line of work by evaluating adaptive evidence acquisition, multi-step tool use, and information synthesis, with video serving as evidence for completing a real-world task rather than only as the object of a question.

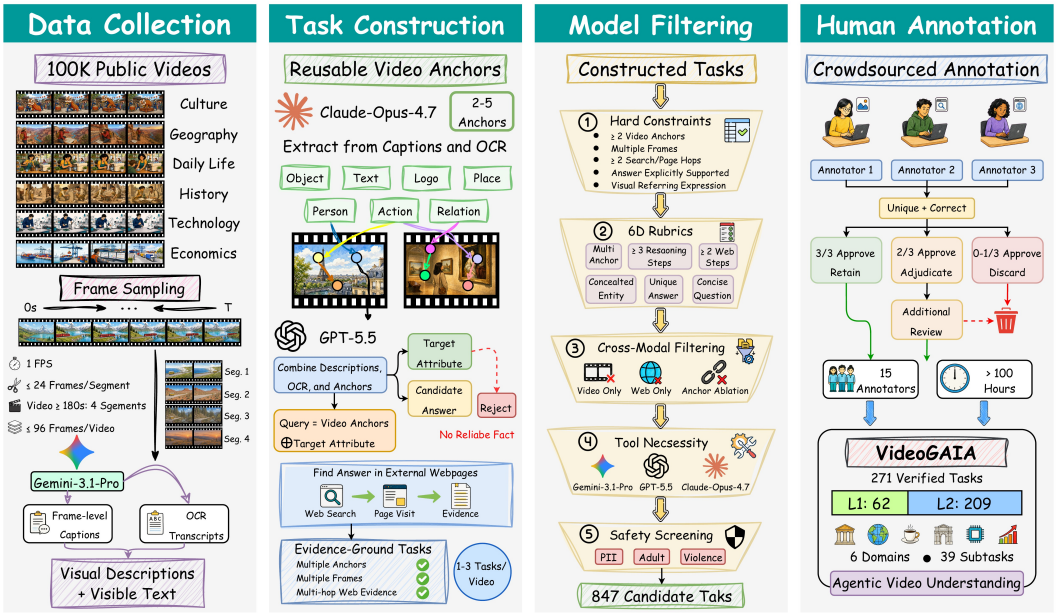


Figure 3: The benchmark construction pipeline of VideoGAIA.

2.2 Agentic Video Understanding

Recent work has begun to formulate video understanding as an interactive evidence-seeking process. VideoAgent (Wang et al., 2024b; Fan et al., 2024) iteratively selects informative visual evidence and employs structured memory or retrieval tools, while Deep Video Discovery (Zhang et al., 2026d), VideoChat-A1 (Wang et al., 2026), and VideoThinker (Li et al., 2026) explore multi-granular retrieval, coarse-to-fine inspection, and learned tool use. VideoWebArena (Jang et al., 2024) evaluates agents that execute skills demonstrated in video tutorials, whereas Video-BrowseComp (Liang et al., 2026), VideoDR (Liu et al., 2026), and LongVidSearch (Yu et al., 2026) combine video evidence with search or multi-hop reasoning. Although these studies establish important foundations for video agents, many focus on specific retrieval pipelines, controlled environments, or evidence navigation within videos. In contrast, VideoGAIA evaluates general-purpose assistants across real-world domains, requiring them to iteratively inspect videos, search the web, read webpages, and synthesize multimodal evidence, while providing a unified comparison of 20 frontier MLLMs under the same agent harness.

3 VideoGAIA

In this section, we describe the data collection process (Sec. 3.1), task construction procedure (Sec. 3.1), and multi-stage quality-control protocol (Sec. 3.2). We then present a statistical overview of VideoGAIA (Sec. 3.3). Figure 3 illustrates an overview of the benchmark construction pipeline.

3.1 Data Collection and Task Construction

We first collect 100K public videos from online platforms. These seed videos serve as the starting point for constructing video-agent tasks, with each video processed through the following pipeline.

Frame-Level Video Captioning. We sample each video at 1 frame per second (FPS) and divide it into segments, retaining at most 24 frames per segment. For videos longer than 180 seconds, we process up to four segments, resulting in an upper bound of approximately 96 frames per video. We then use Gemini-3.1-Pro (Google, 2026a) to generate detailed frame-level visual descriptions and extract visible textual information through optical character recognition (OCR).

Evidence-Grounded Task Generation. Next, we use Claude-Opus-4.7 (Anthropic, 2026b) to extract two to five reusable video anchors from the visual descriptions and OCR transcripts. Each anchor corresponds to an object, text string, logo, place, person, action, or spatial relation. Given the video captions, OCR information, and extracted anchors, GPT-5.5 (OpenAI, 2026b) identifies target attributes suitable for task construction and proposes candidate answers. A seed video is rejected if no reliable candidate fact can be identified. We further require each task to depend on multiple video anchors rather than a single isolated frame or information inferable from the question alone. To retrieve the external evidence required to construct each task, we combine the video-anchor descriptions with the target attribute to form search queries. Search results are obtained through the Serper API (Serper, 2025), and the corresponding webpages are converted into textual content using the Jina API (Jina, 2025). This design requires models to first recognize relevant entities or clues from the video and then retrieve the target information from external webpages, rather than exposing the answer directly. Up to three candidate tasks are independently generated from each seed video.

3.2 Quality Control and Analysis

After task generation, all candidate tasks undergo a strict multi-stage quality-control pipeline to eliminate incorrect, ambiguous, or insufficiently agentic samples.

Rubric-Based Quality Control. We first enforce a set of hard constraints. Each task must contain at least two reasoning hops that require web search or webpage content extraction, and its reference answer must be explicitly supported by the title, search snippet, or a verbatim passage in the retrieved evidence. The task must also involve at least two video anchors, require observations from multiple frames, conceal the answer entity through visual referring expressions, and establish a multi-hop reasoning chain from video evidence to external webpages. Candidates satisfying these constraints are subsequently scored by Claude-Opus-4.8 (Anthropic, 2026c) according to six-dimensional rubrics: (1) the task strictly requires at least two video anchors and cannot be solved from a single frame; (2) the solution involves at least three distinct reasoning steps connecting video evidence with web information; (3) at least two steps require web search or full-page content extraction; (4) the question refers to relevant entities through visual descriptions without directly revealing the answer; (5) the answer is unique and corresponds to a single canonical fact; and (6) the question is concise, grammatically correct, and of appropriate length. Each criterion is assigned a score of either 1 or 2, resulting in a maximum score of 12. Only tasks receiving at least 8 points are retained.

Cross-Modal Filtering. We further examine whether each remaining task genuinely requires multimodal evidence. Specifically, Gemini-3.1-Pro (Google, 2026a) is asked to solve the task using only the video, while GPT-5.5 (OpenAI, 2026b) receives only the web content. We additionally perform anchor ablation by removing each video anchor and testing whether the task remains answerable. A candidate is filtered out if it can be solved in either unimodal setting or if all video anchors are dispensable, as such cases do not genuinely require multimodal and agentic reasoning.

Tool-Necessity Assessment. For candidates that pass previous stages, we disable all external tools and ask Gemini-3.1-Pro (Google, 2026a), GPT-5.5 (OpenAI, 2026b), and Claude-Opus-4.7 (Anthropic, 2026b) to answer the question directly using only the video and their world knowledge. If any model produces the correct answer, the task is removed. Such tasks reduce to single-turn video understanding and therefore do not adequately evaluate tool-augmented video agents.

Safety Screening. We screen all remaining samples for personally identifiable or sensitive information, including government-issued identification numbers and credit-card details. We also remove videos containing adult content, graphic violence, or other material flagged by model safety filters. Any task involving such content is excluded from the benchmark.

Human Crowdsourced Annotation. After the preceding quality-control stages, only 847 candidate tasks remain from the initial pool of 100K seed videos. Because model-based filtering alone cannot guarantee task correctness, each candidate is independently verified by three human annotators. The annotators solve the task from scratch by inspecting the video, conducting web searches, and determining whether the proposed answer is both unique and correct. Tasks unanimously approved by all three annotators are retained directly. Tasks approved by two of the three annotators undergo

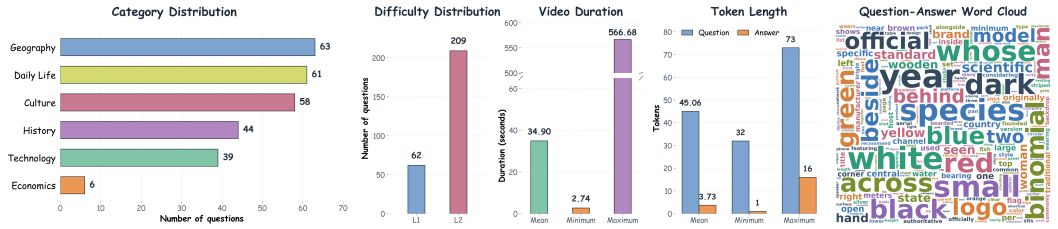


Figure 4: Overview of VideoGAIA dataset statistics, including category and difficulty distributions, video duration, question and answer token lengths, and a question-answer word cloud.

an additional adjudication process involving further human review, while tasks approved by one or none of the annotators are discarded. After more than 100 hours of annotation by 15 annotators, we obtain 271 verified tasks that constitute the VideoGAIA test set. Following difficulty calibration, 62 tasks are assigned to level 1 (L1) and the remaining 209 tasks to the more challenging level 2 (L2).

3.3 Data Statistics

As shown in Figure 4, we summarize the final VideoGAIA benchmark. It contains 271 tasks across six real-world categories: Geography (63), Daily Life (61), Culture (58), History (44), Technology (39), and Economics (6). This broad coverage requires agents to connect heterogeneous visual evidence with external knowledge. VideoGAIA includes 62 L1 and 209 L2 tasks, with L2 accounting for 77.12% of the benchmark, reflecting its emphasis on multi-step evidence acquisition and cross-modal reasoning. Video durations average 34.90 seconds and range from 2.74 to 566.68 seconds, covering both short clips with transient visual clues and longer videos that require temporal navigation. English questions average 45.06 tokens, whereas reference answers average 3.73 tokens. Questions are relatively detailed because they encode multiple visual anchors, referring expressions, and the external attribute to retrieve, while answers remain concise and canonical. This design preserves open-ended evaluation without multiple-choice cues. The joint word cloud further highlights the lexical diversity of entities, attributes, dates, locations, species, products, and official names.

4 Experiments

4.1 Evaluation Setup

We evaluate 20 frontier MLLMs using a unified agent loop protocol based on the ReAct (Yao et al., 2022) framework. Each evaluated model receives the question together with 20 coarse frames sampled uniformly across the complete video. It may then interleave reasoning with three tools: *web search* for information seeking, *page visit* for extracting web page contents, and *thinking with videos*, which samples up to 20 additional frames from a model-selected temporal segment. The video tool is visual-only and does not expose audio. We allow at most 40 agent steps per task and use the same prompts, tools, frame resolution, and step budget for every model. Tool outputs are not inserted into the trajectory verbatim. Instead, the evaluated model itself summarizes the context returned by each tool before it is supplied to the next agent turn. We report pass@1 accuracy over all tasks and each result is averaged across three independent runs. Predictions are judged against the human-verified reference answer by GPT-5.5 (OpenAI, 2026b), with DeepSeek-V4-Pro (DeepSeek-AI, 2026) used only when the primary judge remains unavailable or empty after several retries.

4.2 Main Results

Table 1 reports category-level and overall agent loop accuracy, while Figure 5 separates performance on L1 and L2 across frontier models. Seed2.0-Pro (Seed, 2026) ranks first overall accuracy at 58.30%, followed by Qwen3.7-Plus (Qwen Team, 2026c) at 56.09% and Kimi-K3 (Kimi Team, 2026b) at 54.61%. Notably, no evaluated model reaches 60% accuracy, despite access to both video reinspection and open-web tools. A family-level comparison reveals a broadly positive, but not universal, generational trend. Newer models improve over their predecessors in the Kimi family

Table 1: The pass@1 (%) results of various models under the agent loop setting on VideoGAIA.

Model	Culture	Geography	Daily Life	History	Technology	Economics	Overall
Agent Loop							
Qwen3.5-397B (Qwen Team, 2026a)	43.10	52.38	55.74	45.45	56.41	33.33	50.18
Qwen3.5-Plus (Qwen Team, 2026a)	46.55	47.62	54.10	63.64	56.41	16.67	52.03
Qwen3.6-Plus (Qwen Team, 2026b)	46.55	46.03	50.82	54.55	48.72	50.00	49.08
Qwen3.7-Plus (Qwen Team, 2026c)	50.00	58.73	57.38	56.82	58.97	50.00	56.09
Qwen3.8-Max (Qwen Team, 2026d)	44.83	42.86	62.30	50.00	51.28	50.00	50.18
GLM-4.6V (Z.ai, 2025)	48.28	42.86	50.82	50.00	41.03	50.00	46.86
GLM-5V-Turbo (Hong et al., 2026)	51.72	44.44	54.10	52.27	48.72	50.00	50.18
Seed2.0-Pro (Seed, 2026)	51.72	60.32	60.66	54.55	66.67	50.00	58.30
MiMo-V2.5 (MiMo Team, 2026)	53.45	47.62	44.26	52.27	48.72	33.33	48.71
Kimi-K2.5 (Team, Kimi et al., 2026)	50.00	46.03	50.82	50.00	53.85	50.00	49.82
Kimi-K2.6 (Kimi Team, 2026a)	50.00	46.03	59.02	59.09	51.28	50.00	52.77
Kimi-K3 (Kimi Team, 2026b)	50.00	49.21	65.57	52.27	56.41	50.00	54.61
Gemini-3.1-Pro (Google, 2026a)	41.38	49.21	57.38	54.55	53.85	50.00	50.92
Gemini-3.5-Flash (Google, 2026b)	37.93	30.16	47.54	45.45	64.10	66.67	43.91
Claude-Opus-4.6 (Anthropic, 2026a)	37.93	39.68	52.46	59.09	51.28	83.33	47.97
Claude-Opus-4.7 (Anthropic, 2026b)	46.55	30.16	31.15	34.09	28.21	50.00	34.69
Claude-Opus-4.8 (Anthropic, 2026c)	39.66	30.16	49.18	36.36	25.64	33.33	36.90
GPT-5.2 (OpenAI, 2025)	50.00	39.68	57.38	54.55	53.85	66.67	50.92
GPT-5.4 (OpenAI, 2026a)	56.90	42.86	52.46	65.91	43.59	50.00	52.03
GPT-5.5 (OpenAI, 2026b)	50.00	46.03	62.30	54.55	56.41	33.33	53.14

(49.82% for Kimi-K2.5 (Team, Kimi et al., 2026), 52.77% for Kimi-K2.6 (Kimi Team, 2026a), and 54.61% for Kimi-K3 (Kimi Team, 2026b)), the GLM family (46.86% for GLM-4.6V (Z.ai, 2025) versus 50.18% for GLM-5V-Turbo (Hong et al., 2026)), and the Gemini family (43.91% versus 50.92% for the two evaluated generations). However, this pattern does not hold across all families: Claude-Opus-4.7 (Anthropic, 2026b) and Claude-Opus-4.8 (Anthropic, 2026c) score 34.69% and 36.90%, respectively, both below Claude-Opus-4.6 (Anthropic, 2026a) at 47.97%. This non-monotonic scaling suggests that improvements in general model capability do not automatically transfer to reliable multimodal tool use. Performance also varies substantially by category: Seed2.0-Pro (Seed, 2026) leads on Geography and Technology, Kimi-K3 (Kimi Team, 2026b) is strongest on Daily Life, and GPT-5.4 (OpenAI, 2026b) performs best on Culture and History. Economics contains only six tasks, so its per-category percentages should be interpreted with caution. These results indicate that VideoGAIA measures a combination of visual grounding, information seeking, and evidence integration for which no current model provides uniformly strong behavior.

4.3 Analysis and Findings

Difficulty Analysis. The results shown in Figure 5 demonstrate a consistent decline from L1 to L2 for every evaluated model. Among the three strongest models overall, Seed2.0-Pro (Seed, 2026) decreases from 75.81% to 53.11%, Qwen3.7-Plus (Qwen Team, 2026c) from 79.03% to 49.28%, and Kimi-K3 (Kimi Team, 2026b) from 77.42% to 47.85%. The 22.70%–29.75% point gaps confirm that L2 captures substantially harder trajectories rather than merely redistributing category frequencies. In particular, L2 tasks more often require agents to resolve indirect visual references, issue several targeted searches, and reconcile evidence across video observations and webpages.

Tool Analysis. We count every explicit tool call made by five representative models across all 271 trajectories. As shown in Figure 6, web search dominates all tool-use profiles, accounting for 60.5%–90.6% of calls, whereas page visits never exceed 14.0%. Kimi-K3 (Kimi Team, 2026b) allocates the largest share to thinking with videos (25.5%), suggesting a stronger tendency to revisit

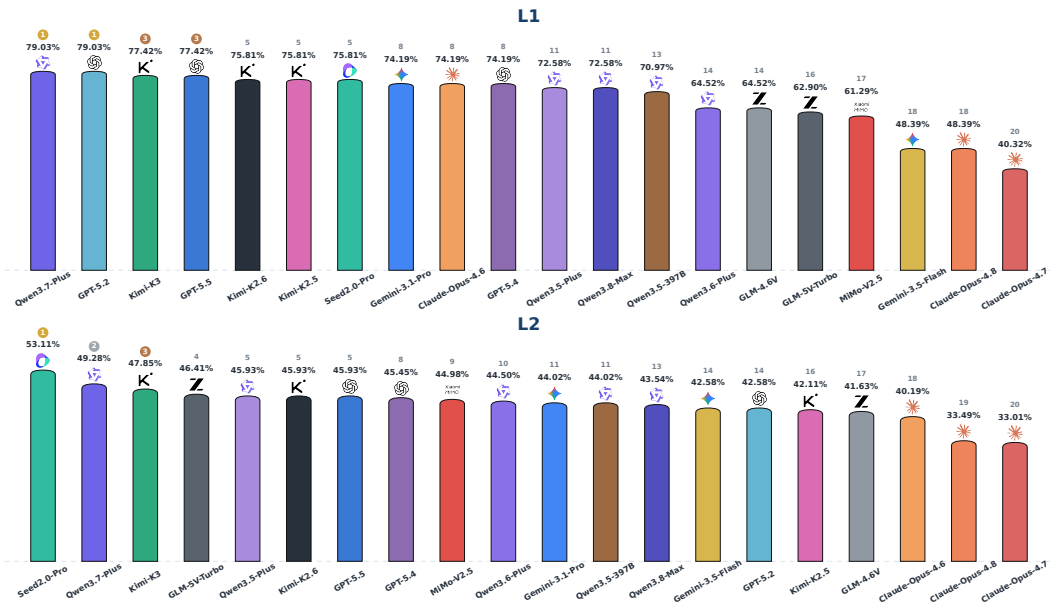


Figure 5: Performance comparison among various frontier models under two difficulty levels.

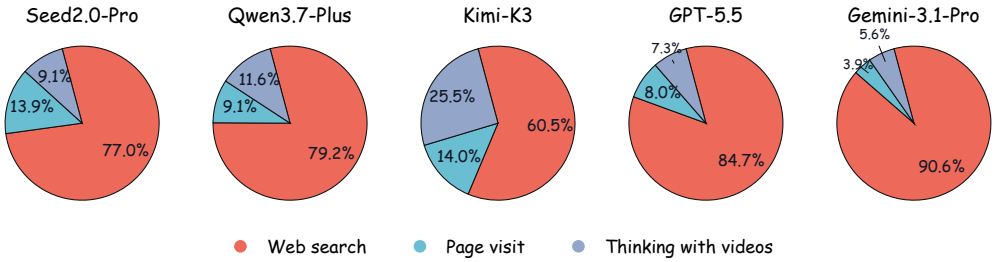


Figure 6: Tool-use distributions of five representative models under the agent loop setting.

temporal evidence. In contrast, Gemini-3.1-Pro (Google, 2026a) makes 2,557 calls (90.6% of them searches), while GPT-5.5 (OpenAI, 2026b) makes only 262 calls, yet GPT-5.5 (OpenAI, 2026b) attains higher accuracy. This phenomenon reflects that a larger tool budget alone does not ensure success. Effective video agents should learn to select informative video segments, open authoritative sources, and stop once sufficient evidence has been acquired.

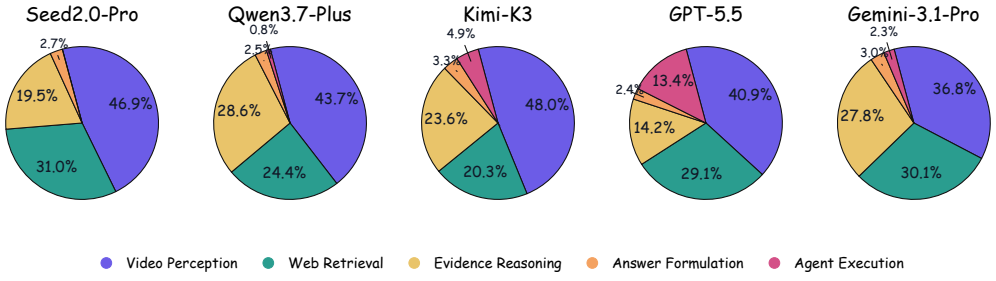


Figure 7: Distribution of dominant error causes among incorrect predictions from five models.

Error Analysis. We conduct a trace-level audit of all 615 incorrect predictions produced by the same five representative models and assign each trajectory one mutually exclusive dominant cause: video perception, web retrieval, evidence reasoning, answer formulation, or agent execution. Figure 7 shows

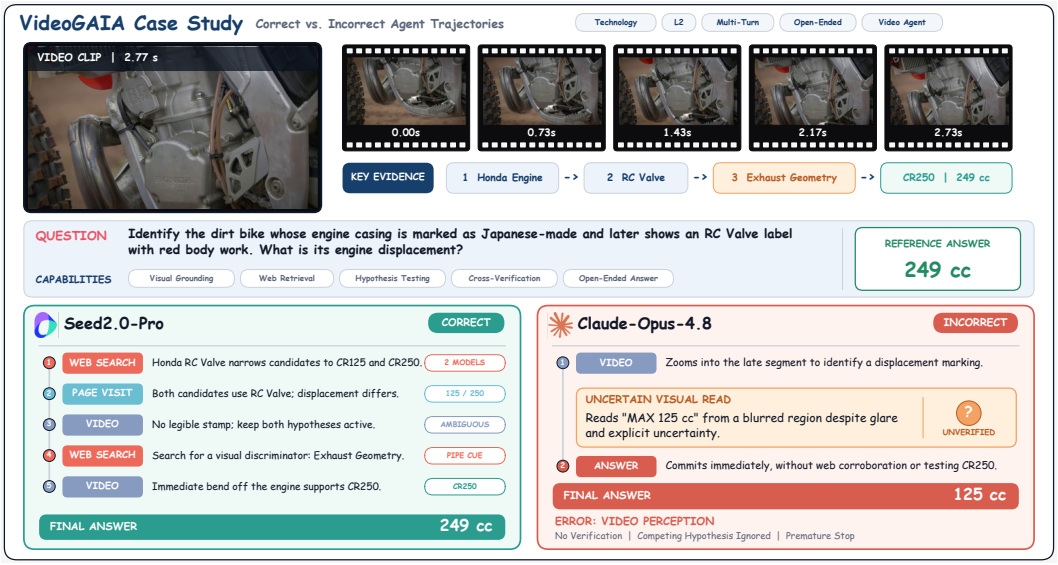


Figure 8: The visualization of correct and incorrect video agent trajectories on VideoGAIA.

that video perception is the largest source of error for every model (36.8%–48.0%), demonstrating that a wrong visual anchor often contaminates all downstream searches. Web retrieval contributes 20.3%–31.0% and evidence reasoning contributes 14.2%–28.6%, revealing separate bottlenecks in finding evidence and using evidence once found. Answer formulation accounts for only 2.4%–3.3%, so most failures cannot be explained by surface-form mismatch. GPT-5.5 (OpenAI, 2026b) exhibits a comparatively high 13.4% agent-execution error rate, primarily due to malformed action outputs and consequent parse failures, whereas this category remains below 5% for the other models. These results suggest that progress on VideoGAIA requires joint improvements in fine-grained video grounding, deliberate retrieval, cross-source reasoning, and reliable tool protocols.

Case Analysis. In Figure 8, we provide a concrete example of agentic video understanding on VideoGAIA. The initial video reveals Honda branding, a "Made in Japan" casing, an RC Valve, and red bodywork, but these anchors do not uniquely determine the displacement because both CR125 and CR250 remain plausible. Seed2.0-Pro (Seed, 2026) therefore performs adaptive evidence acquisition. The first search shows that both CR125 and CR250 use the RC Valve system and therefore remain plausible candidates. The first video revisit confirms that no displacement marking is legible, while a second search identifies exhaust-pipe geometry as a model-specific discriminative cue. Seed2.0-Pro (Seed, 2026) then reinspects the video clip, observes the immediate bend associated with CR250, and returns the correct 249 cc answer. Claude-Opus-4.8 (Anthropic, 2026c), by contrast, asks the video tool to read a putative stamp and interprets a blurred region as "MAX 125 cc," even though its own observation notes glare and uncertainty. It neither gathers external evidence nor evaluates the CR250 hypothesis before terminating. The contrast shows that VideoGAIA tests more than visual recognition or retrieval in isolation. A successful video agent must maintain uncertainty, decide what evidence to acquire next, and integrate retrieved knowledge with targeted video inspection.

5 Conclusion

In this work, we introduce VideoGAIA, a benchmark that advances agentic video understanding. VideoGAIA contains 271 human-verified video tasks across six real-world categories and requires models to iteratively ground visual evidence, use external tools, and integrate information across turns. Under a unified evaluation of 20 frontier MLLMs, all models remain below 60% accuracy, revealing substantial limitations in video grounding, cross-modal reasoning, and reliable tool use. Our further analyses show that successful task execution depends not on making more tool calls alone, but on adaptive evidence acquisition and verification. We hope VideoGAIA serves as a rigorous testbed for developing video agents that can use videos as evidence to solve practical real-world tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and language polishing. We did not use generative AI tools for generating references or designing the experimental methodology, and the other required disclosure items are not applicable to this work. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REFERENCES

Anthropic. Claude opus 4.6 system card, 2026a. URL https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

Anthropic. Claude opus 4.7 system card, 2026b. URL https://www-cdn.anthropic.com/037f06850df7fbe871e206dad004c3db5fd50340/Claude%20Opus%204.7%20System%20Card.pdf.

Anthropic. Claude opus 4.8 system card, 2026c. URL https://www-cdn.anthropic.com/0b4915911bb0d19eca5b5ee635c80fef830a37ea.pdf.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. Opensearch-vl: An open recipe for frontier multimodal search agents. arXiv preprint arXiv:2605.05185, 2026.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

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.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. arXiv preprint arXiv:2406.14515, 2024.

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. In CVPR, 2025.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. arXiv preprint arXiv:2508.05748, 2025.

Google. Gemini 3.1 pro model card, 2026a. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf.

Google. Gemini 3.5 flash model card, 2026b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf.

Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, et al. Glm-5v-turbo: Toward a native foundation model for multimodal agents. arXiv preprint arXiv:2604.26752, 2026.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 27798–27828, 2026.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. arXiv preprint arXiv:2601.22060, 2026.

Lawrence Jang, Yinheng Li, Dan Zhao, Charles Ding, Justin Lin, Paul Pu Liang, Rogerio Bonatti, and Kazuhito Koishida. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. arXiv preprint arXiv:2410.19100, 2024.

Jina. Jina reader, 2025. URL https://jina.ai/reader/.

Kimi Team. Kimi k2.6: Advancing open-source coding, 2026a. URL https://www.kimi.com/blog/kimi-k2-6.

Kimi Team. Kimi k3: Open frontier intelligence, 2026b. URL https://www.kimi.com/blog/kimi-k3.

Chenglin Li, Qianglong Chen, Feng Han, Yikun Wang, Xingxi Yin, Yan Gong, Ruilin Li, Yin Zhang, and Jiaqi Wang. Videothinker: Building agentic videollms with llm-guided tool reasoning. arXiv preprint arXiv:2601.15724, 2026.

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, 2024.

Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Nicu Sebe, Zheng Liu, and Lizi Liao. Video-browser: Towards agentic open-web video browsing. arXiv preprint cs.CV/2512.23044, 2026.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Jisheng Dang, Rui Xu, Sen Hu, Jianheng Hou, et al. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning. arXiv preprint arXiv:2601.06943, 2026.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. arXiv preprint arXiv:2503.21460, 2025.

MiMo Team. Xiaomi mimo-v2.5: A leap in agency and multimodality., 2026. URL https://mimo.xiaomi.com/mimo-v2-5.

OpenAI. Introducing gpt-5.2, 2025. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. Gpt-5.4 thinking system card, 2026a. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

OpenAI. Gpt-5.5 system card, 2026b. URL https://openai.com/index/gpt-5-5-system-card/.

Qwen Team. Qwen3-max: Just scale it, September 2025.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026b. URL https://qwen.ai/blog?id=qwen3.6.

Qwen Team. Qwen3.7-Plus: Multimodal agent intelligence, May 2026c. URL https://qwen.ai/blog?id=qwen3.7-plus.

Qwen Team. Qwen3.8-Max-Preview, 2026d.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. arXiv preprint arXiv:2607.00248, 2026.

Serper. Serper api, 2025. URL https://serper.dev/.

Team, Kimi, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276, 2026.

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

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.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 10467–10475, 2026.

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

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. arXiv preprint arXiv:2412.14171, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629, 2022.

Rongyi Yu, Chenyuan Duan, and Wentao Zhang. Longvidsearch: An agentic benchmark for multi-hop evidence retrieval planning in long videos. arXiv preprint arXiv:2603.14468, 2026.

Z.ai. Glm-4.6v: Open source multimodal models with native tool use, 2025. URL https://z.ai/blog/glm-4.6v.

Z.ai. Glm-5.2: Built for long-horizon tasks, 2026. URL https://z.ai/blog/glm-5.2.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In International Conference on Learning Representations, volume 2026, pp. 48767–48807, 2026a.

Fan Zhang, Vireo Zhang, Shengju Qian, Haoxuan Li, Zheng Lian, Hao Wu, Yuan Gao, Xinyu Geng, Xin Wang, and Pheng-Ann Heng. Struct-searcher: Agentic structural thinking advances multimodal deep information seeking, 2026b. URL https://arxiv.org/abs/2606.07689.

Fan Zhang, Vireo Zhang, Shengju Qian, Haoxuan Li, Hao Wu, Jinyang Wu, Donghao Zhou, Zhihong Zhu, Zheng Lian, Xin Wang, et al. Orchestra-o1: Omnimodal agent orchestration. arXiv preprint arXiv:2606.13707, 2026c.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. Advances in Neural Information Processing Systems, 38:89863–89895, 2026d.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. arXiv preprint arXiv:2508.11630, 2025.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8475–8489, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning, 2026. URL https://arxiv.org/abs/2505.14362.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **3**

2.1 Video Understanding Benchmarks 3

2.2 Agentic Video Understanding 4

3 VideoGAIA **4**

3.1 Data Collection and Task Construction 4

3.2 Quality Control and Analysis 5

3.3 Data Statistics 6

4 Experiments **6**

4.1 Evaluation Setup 6

4.2 Main Results 6

4.3 Analysis and Findings 7

5 Conclusion **9**

A More Related Works **15**

A.1 Multimodal AI Agents 15

B Implementation Details **15**

B.1 System Prompt for Video Agents 15

B.2 System Prompt for Judge Models 17

C More Experimental Results **17**

C.1 Efficiency Analysis 17

C.2 Additional Visualization 18

C.3 Additional Case Study 19

D Limitations and Social Impact **19**

A More Related Works

A.1 Multimodal AI Agents

Large language model agents combine planning, tool use, memory, and feedback to act beyond passive generation (Yao et al., 2022; Luo et al., 2025). Multimodal agents extend this paradigm along two particularly relevant directions. First, thinking with images systems actively manipulate visual evidence during reasoning. DeepEyes (Zheng et al., 2026) learns when to ground and revisit image regions through reinforcement learning, while Thyme (Zhang et al., 2025) generates executable code for operations such as cropping, rotation, enhancement, and numerical computation. Second, multimodal search agents integrate visual and textual retrieval for open-world information seeking. WebWatcher (Geng et al., 2025), Vision-DeepResearch (Huang et al., 2026), and OpenSearch-VL (Chen et al., 2026) develop multi-turn search workflows and training recipes for acquiring evidence across modalities, while Struct-Searcher (Zhang et al., 2026b) explicitly organizes heterogeneous and potentially conflicting evidence through structural reasoning. VideoGAIA connects these directions in the video domain, where evidence may be brief or temporally dispersed, by jointly evaluating targeted video inspection, web retrieval, webpage reading, and cross-modal verification.

B Implementation Details

B.1 System Prompt for Video Agents

System Prompt for Video Agents

```
You are a video agent. You solve video-based questions by interleaving reasoning with tools. You have exactly these tool actions: VideoAnalysisAction, WebSearchAction, and ExtractUrlContentAction.

==== Core Rules ====
1. The initial user message includes coarse frames sampled uniformly from the whole video. First reason from those attached frames.
2. Use VideoAnalysisAction only when the coarse frames are insufficient, a specific time segment needs closer inspection, or you need more frames from a segment.
3. VideoAnalysisAction is visual-only. Do not ask for audio, transcription, speech, music, or sound. Never include analyze_audio.
4. Use WebSearchAction for open-web search. Use ExtractUrlContentAction only after search finds a promising URL.
5. After each tool call you will see only a summarized observation wrapped in <observation></observation>.
6. Final answer should be concise and match the expected answer style.
7. Do not invent tools. The "action" value inside <action> must be exactly one available tool name.
8. Use exactly one tool action per assistant turn. After an observation, decide the next single action or final answer.

==== Output Format ====
Every assistant turn MUST use tags. No markdown and no code fences.

For a tool call, output exactly:
<think>Brief reasoning for the next step.</think>
<action>{"action":"<EXACT_TOOL_NAME>","params":{"<tool-specific parameter>":"<value>"},"memory":"brief notes"}</action>

For the final answer, output exactly:
<think>Brief reasoning that the evidence is sufficient.</think>
<answer>concise final answer</answer>
```

```

The <action> content MUST be valid JSON. Do not put multiple actions in one <action> tag.
Do not output <answer> in the same turn as <action>.
Do not use "command", "tool", "tool_name", or "function_call" fields.

==== Tool Parameter Formats ====
VideoAnalysisAction params:
{
  "query": "<specific visual question to answer from the video>",
  "video_path": "<local mp4 path>",
  "max_frames": 20,
  "start_time": 0,
  "end_time": 12.5
}

WebSearchAction params:
{
  "query": "<web search query>",
  "k": 5,
  "gl": "us",
  "hl": "en"
}

ExtractUrlContentAction params:
{
  "url": "<URL from WebSearchAction>",
  "browse_query": "<what information you need from this page>"
}

==== Tool Call Examples ====
Example 1 - inspect a segment more closely when the attached coarse frames are not enough:
<think>The coarse frames show the relevant event around the middle of the video, but I need a closer look at that segment.</think>
<action>{"action":"VideoAnalysisAction","params":{"query":"Inspect this segment carefully and identify the visible event, objects, people, text, and temporal clues needed to answer the question. Ignore audio.","video_path":"/path/to/video.mp4","start_time":24,"end_time":38,"max_frames":20},"memory":"Zoom in on the uncertain middle segment using visual frames only."}</action>

Example 2 - search the web when the video contains a named entity or external fact:
<think>The visual evidence gives a named entity, but I need external verification.</think>
<action>{"action":"WebSearchAction","params":{"query":"official name of the building with sign Example Hall","k":5,"gl":"us","hl":"en"},"memory":"Need external verification for the visible sign/entity."}</action>

Example 3 - read a promising search result:
<think>A search result looks relevant, so I should extract the needed page content.</think>
<action>{"action":"ExtractUrlContentAction","params":{"url":"https://example.com/article","browse_query":"Find the specific fact needed to answer the VideoDR question."},"memory":"Search found a relevant page; extract the detail."}</action>

Example 4 - provide final answer after enough evidence:
<think>The observations are sufficient to answer the question concisely.</think>
```

```
<answer>the Golden Gate Bridge</answer>

Example 5 - provide the best available answer when steps are nearly exhausted:
<think>The step budget is low and the evidence does not identify the requested detail confidently.</think>
<answer>unknown</answer>
```

B.2 System Prompt for Judge Models

System Prompt for Judge Models

```
You are a strict but fair Video Agent answer judge. Evaluate whether the predicted answer correctly answers the question.

Question:
{question}

Ground-truth answer and accepted aliases:
{gold}

Predicted answer:
{pred}

Rules:
- Use semantic equivalence, not exact string matching.
- Mark correct when the prediction identifies the same underlying entity, place, event, direction, code, quantity, date, or fact as the ground truth.
- Accept aliases, abbreviations, translations, official names vs common names, and formatting differences.
- Extra explanation is acceptable if it contains the correct answer and does not contradict it.
- For numeric answers, require the same value except normal rounding, unit conversion, or formatting differences.
- If the prediction is empty, only hedged, irrelevant, or contradicts the ground truth, mark incorrect.

Return only one character:
1 if the predicted answer is correct.
0 if the predicted answer is incorrect.

Do not output JSON, explanations, markdown, or any other text.
```

C More Experimental Results

C.1 Efficiency Analysis

As shown in Figure 9, we examine whether longer or more expensive trajectories translate into better task performance. We measure response length as the mean number of visible output tokens produced per task across all turns of the ReAct (Yao et al., 2022) loop. The corresponding cost is an output-only computation by summing these output tokens over all 271 tasks and applying each model's public standard output-token price. Input tokens, external tool services, judge calls, hidden reasoning, and other provider-specific charges are excluded. From the results, we observe that response length has a weak, statistically non-significant negative correlation with accuracy, while log total output cost shows a somewhat stronger but still non-significant negative correlation. Seed2.0-Pro (Seed, 2026) reaches the highest accuracy with 391 output tokens per task and an estimated total output cost of $0.251, and Qwen3.7-Plus (Qwen Team, 2026c) ranks second with 426 tokens and $0.136. In

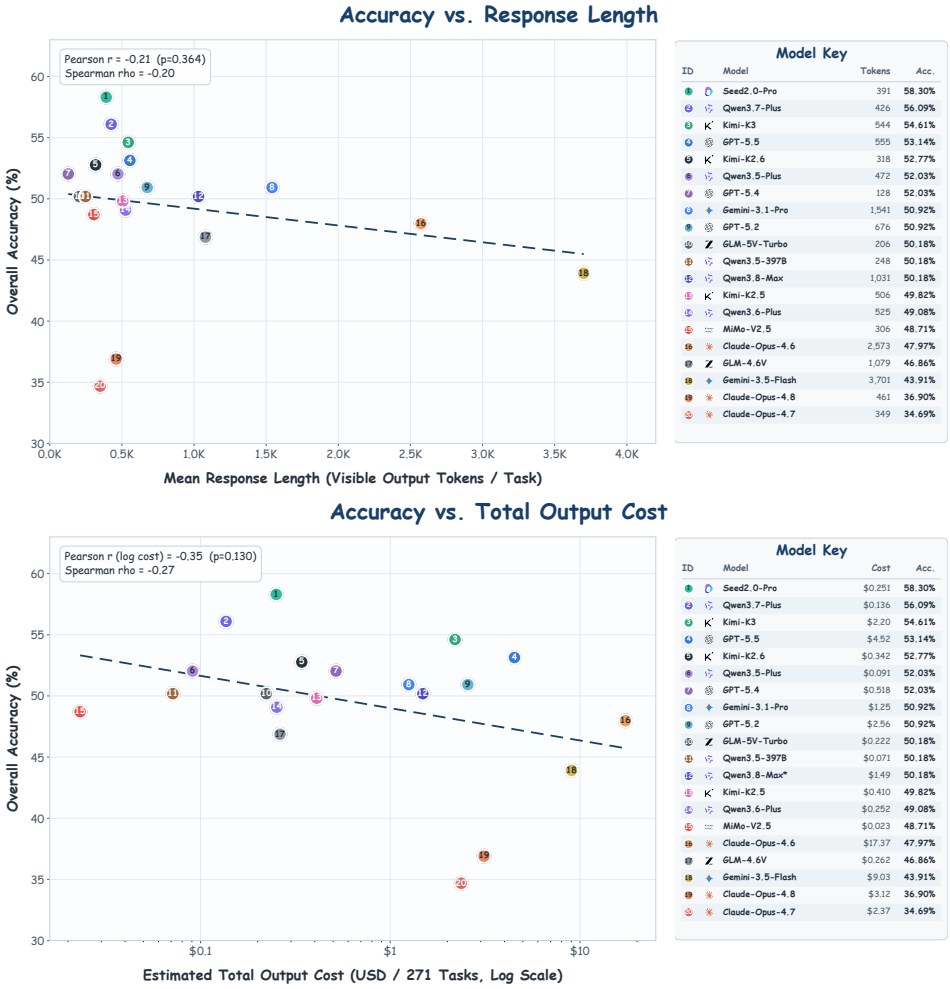


Figure 9: Overall accuracy versus response length and output cost across 20 models on VideoGAIA.

contrast, Gemini-3.5-Flash (Google, 2026b) and Claude-Opus-4.6 (Anthropic, 2026a) produce much longer or more expensive trajectories without corresponding accuracy gains. These results do not imply that reasoning should always be shorter. Rather, they show that trajectory volume alone is a poor proxy for useful computation and that efficient video agents must acquire targeted evidence, avoid redundant searches, and terminate when the answer is sufficiently verified.

C.2 Additional Visualization

As shown in Figure 10, we provide category-level performance comparison for all 20 evaluated frontier video agents. The markedly different polygon shapes show that no model dominates uniformly across the six domains. Seed2.0-Pro (Seed, 2026) is particularly strong in Geography and Technology, Kimi-K3 (Kimi Team, 2026b) performs best in Daily Life, and GPT-5.4 (OpenAI, 2026a) leads in Culture and History. The radar plots also reveal substantial within-family variation, reinforcing that aggregate accuracy can conceal distinct strengths and weaknesses.

Figures 11 and 12 extend the main-text analyses to the remaining 15 models and therefore complete the coverage of all evaluated systems. Web search remains the largest tool category for every additional model, accounting for 53.1%–88.1% of calls, although the relative use of page visits and targeted video inspection varies considerably. The error distributions are similarly consistent with the previous analysis for five representative models. Video perception is the largest dominant cause for all 15 models, comprising 34.9%–57.3% of their incorrect trajectories, while web retrieval and

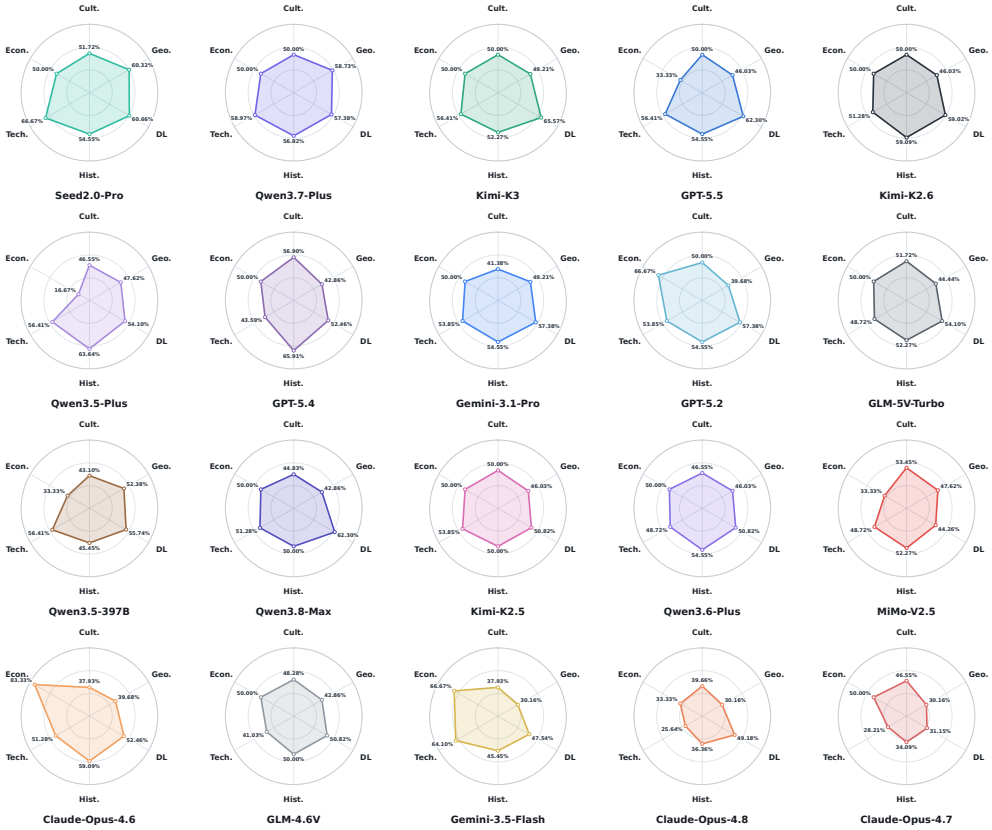


Figure 10: Radar plot comparison of 20 video agents across six categories on VideoGAIA.

evidence reasoning explain much of the remainder. Answer-formulation errors are generally limited, whereas agent-execution failures become substantial for a small number of models. Together, these visualizations indicate that robust performance depends on grounding the correct visual anchor early, retrieving discriminative evidence, and maintaining a reliable action protocol across turns.

C.3 Additional Case Study

Figures 13, 14, 15, 16, 17, and 18 present one complete trajectory from each VideoGAIA domain. In the successful cases, Seed2.0-Pro (Seed, 2026) and Qwen3.7-Plus (Qwen Team, 2026c) ground distinctive visual clues before issuing focused web searches, while Kimi-K3 (Kimi Team, 2026b) revisits the relevant video segment to resolve the color of the racing buoys. The failed trajectories remain superficially plausible but confuse closely related facts. Qwen3.6-Plus (Qwen Team, 2026b) returns August 4 rather than the annotated August 3 inauguration date, GLM-4.6V (Z.ai, 2025) conflates the 2001 release of the SM7B with the original SM7 introduction in 1973, and Gemini-3.1-Pro (Google, 2026a) follows a recent ownership snippet instead of the reference NBCUniversal (Comcast). Together, these examples show that tool access alone is insufficient, and reliable video agents must formulate discriminative queries, preserve uncertainty among nearby hypotheses, and cross-check retrieved facts against the exact visual and temporal context of the question.

D Limitations and Social Impact

VideoGAIA has several limitations. First, the current task collection does not comprehensively cover multilingual scenarios or depend heavily on audio information. Second, to ensure consistent and controlled evaluation across models, we employ a simple ReAct-style agent loop as the video-agent harness. We do not evaluate the models within more sophisticated agent systems, such as Codex or

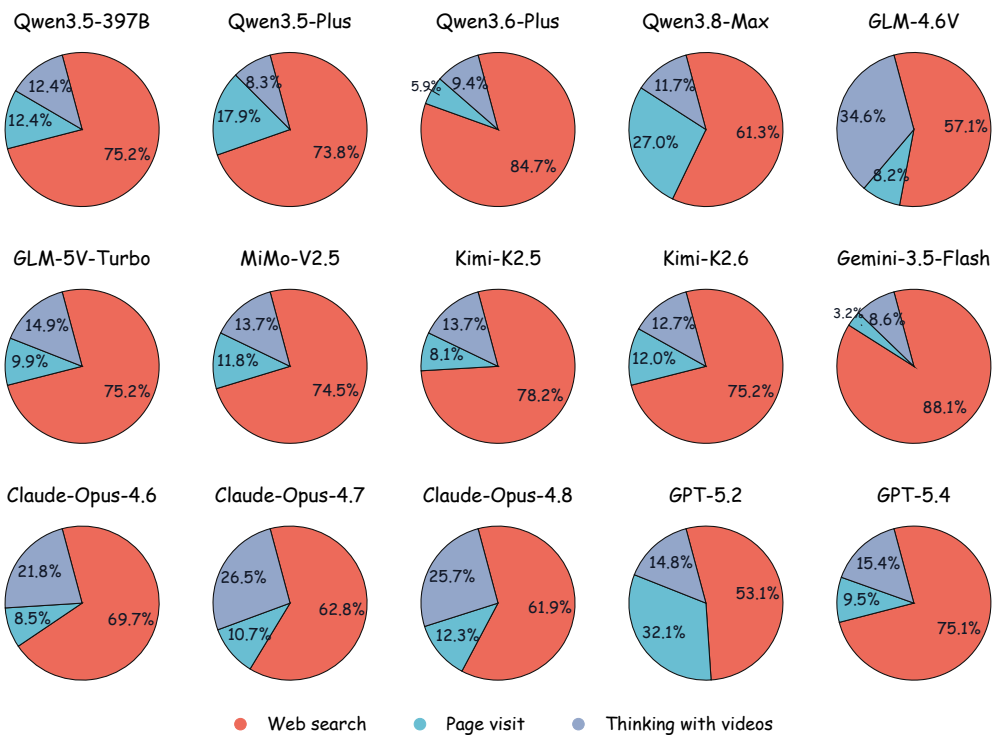


Figure 11: Tool-use distributions of 15 frontier models under the agent loop setting. Percentages are computed over all explicit tool calls on both correct and incorrect trajectories.

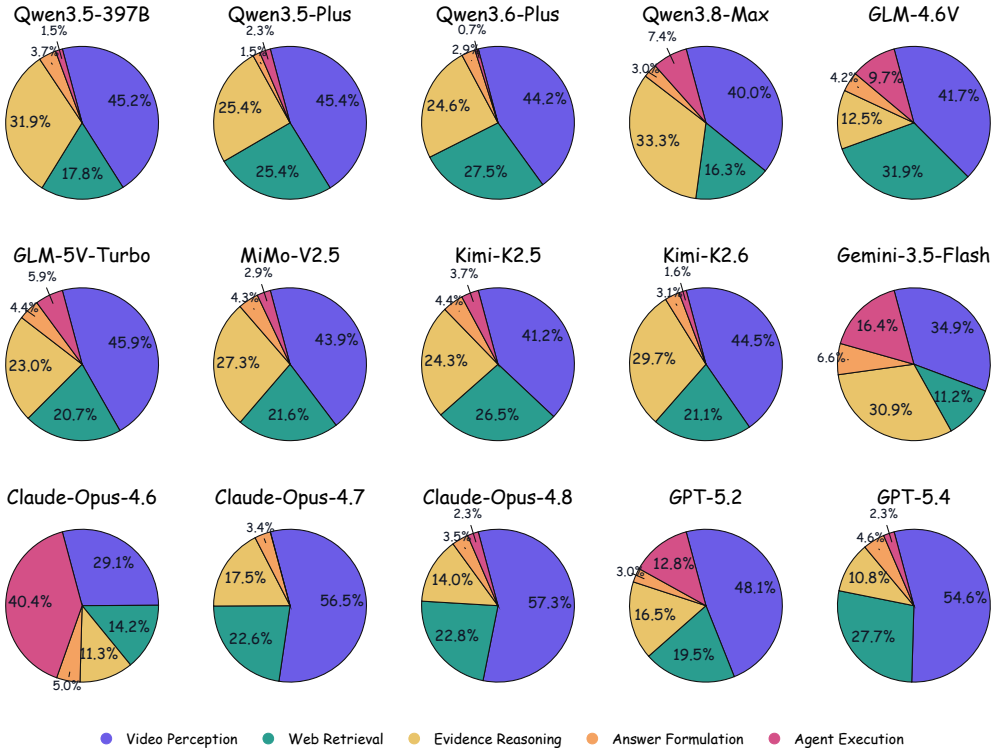


Figure 12: Distribution of dominant error causes among incorrect predictions from 15 frontier models. Each failed trajectory is assigned exactly one category.

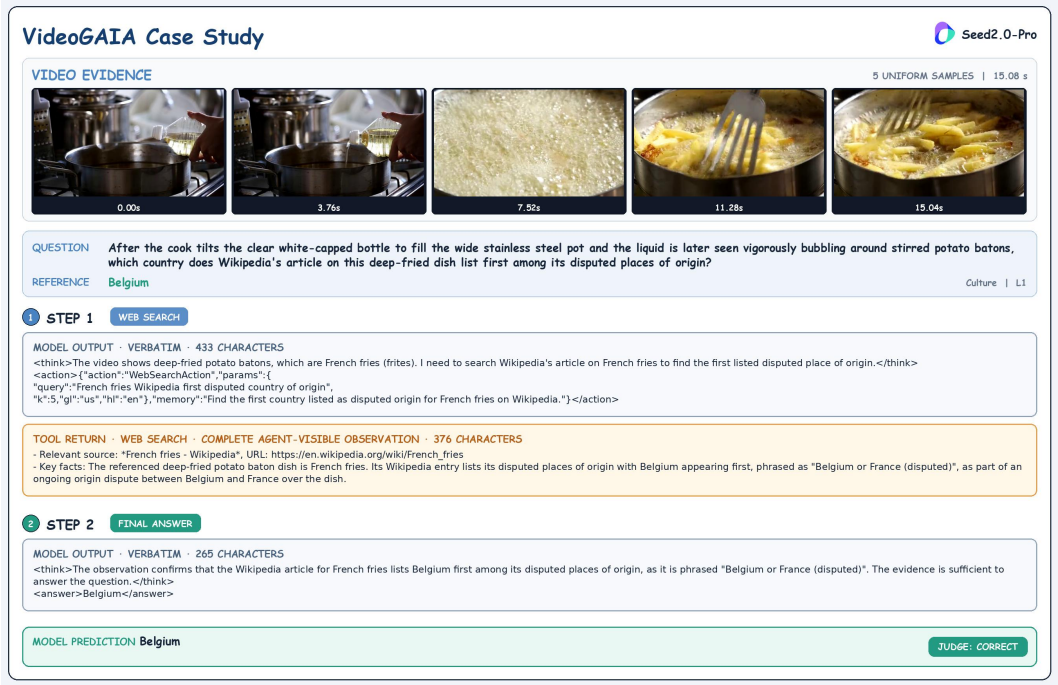


Figure 13: An example from the Culture domain of VideoGAIA.

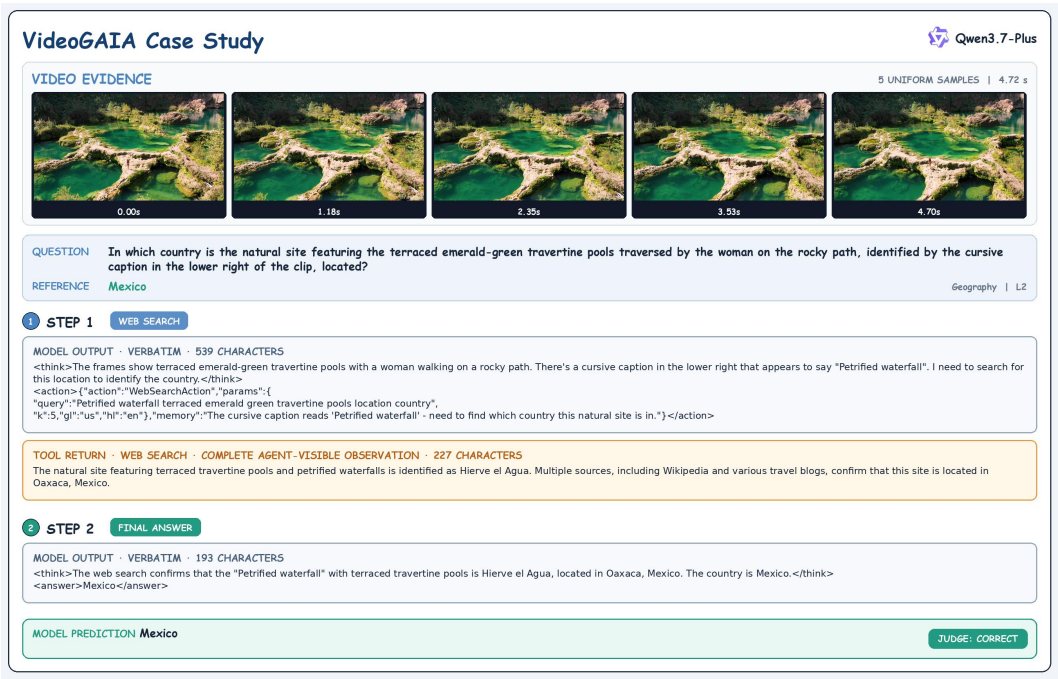


Figure 14: An example from the Geography domain of VideoGAIA.

OpenClaw, which may provide stronger context management, tool orchestration, and long-horizon planning capabilities. We leave the extension to broader modalities, languages, and advanced agent harnesses for future work. Nevertheless, these limitations do not undermine the central contribution of VideoGAIA in establishing and advancing the paradigm of agentic video understanding.

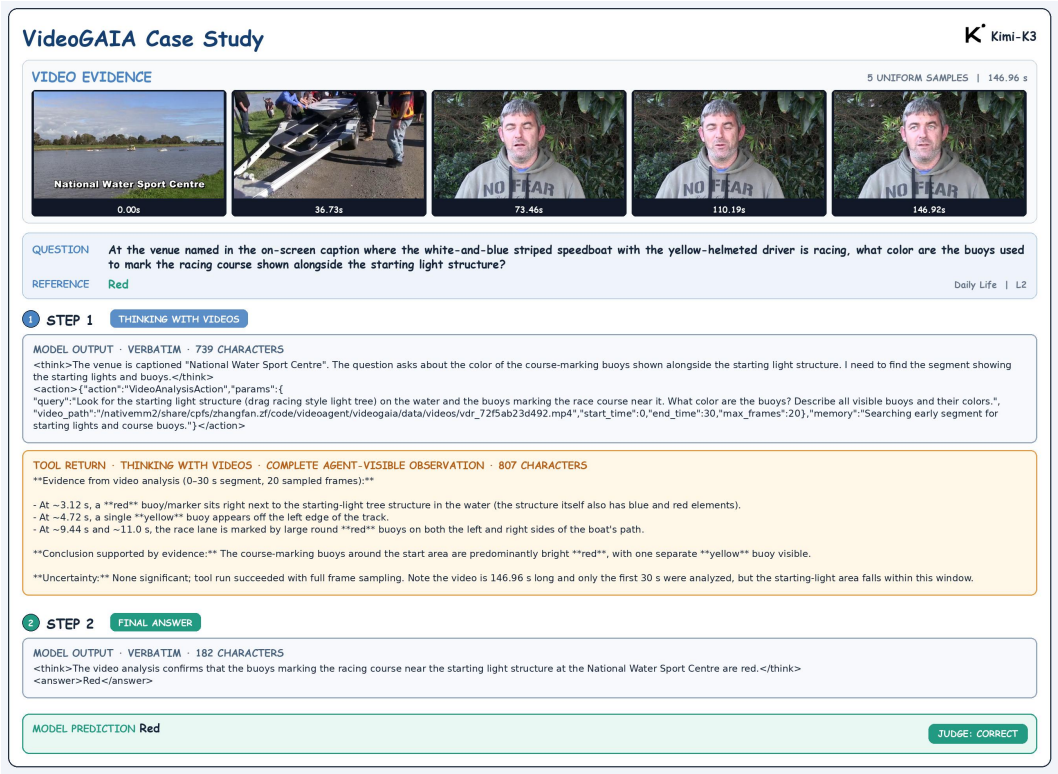


Figure 15: An example from the Daily Life domain of VideoGAIA.

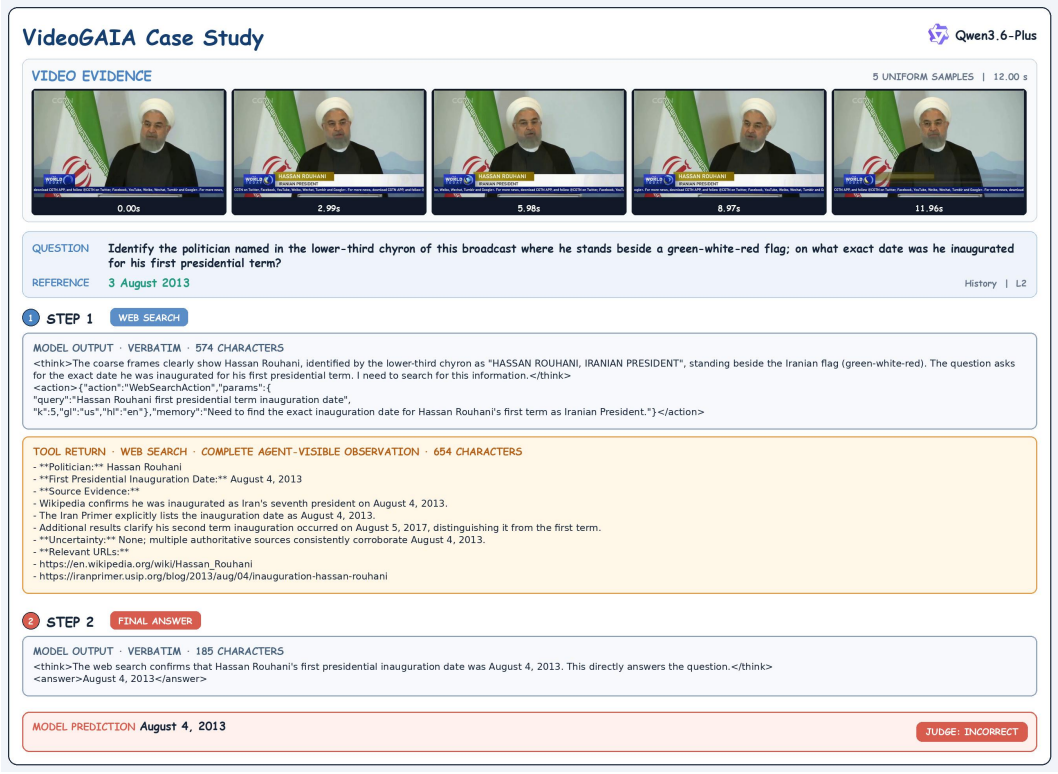


Figure 16: An example from the History domain of VideoGAIA.

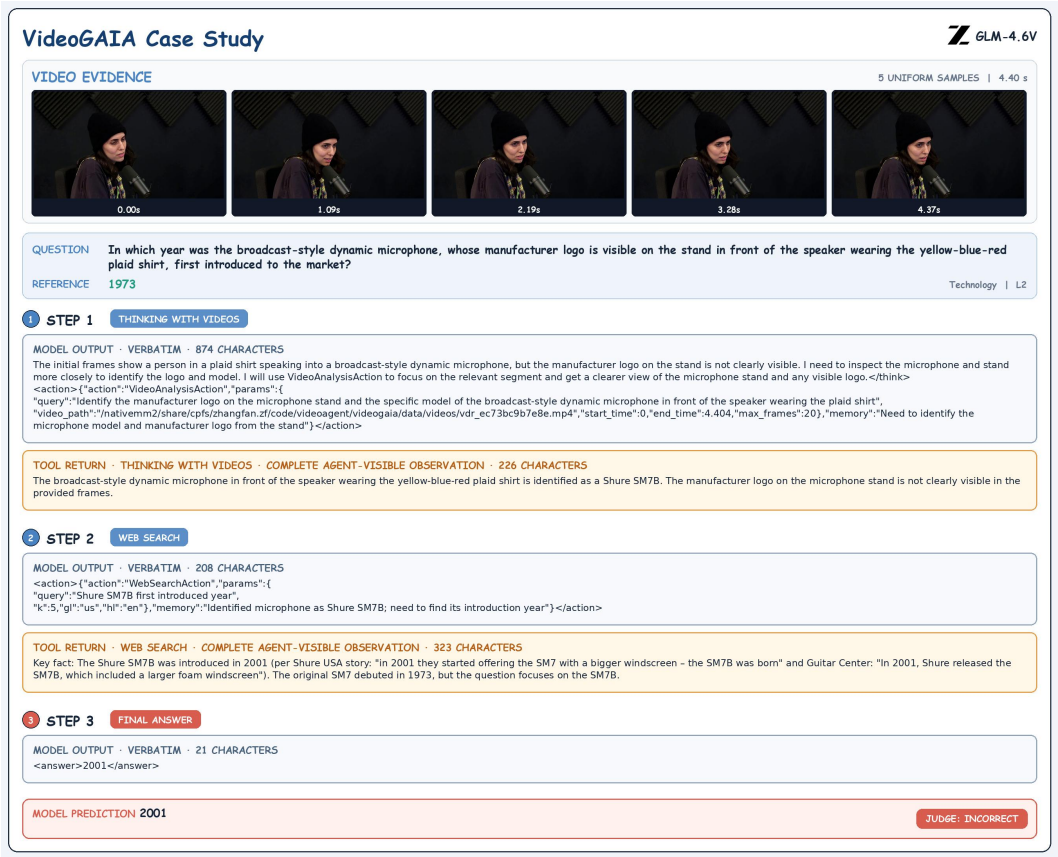


Figure 17: An example from the Technology domain of VideoGAIA.

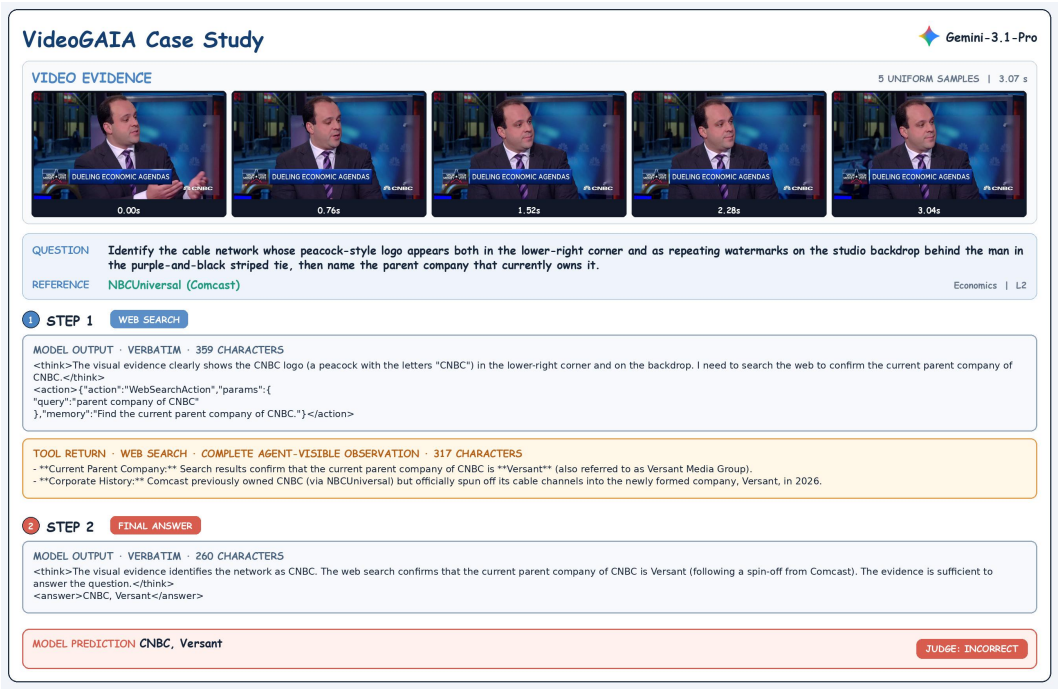


Figure 18: An example from the Economics domain of VideoGAIA.

VideoGAIA has the potential to benefit research in both video understanding and multimodal agents. We hope that it will draw greater attention to video agents and encourage the community to move beyond multimodal agents that primarily operate over static images and text. In the real world, visual information is inherently continuous, dynamic, and temporally evolving. Developing agents capable of perceiving such information, seeking complementary evidence, and acting upon it is therefore essential for transforming multimodal agents from conversational systems into practical productivity tools. To date, we have not identified any potential negative societal impacts.