Watching, Reasoning and Searching: A Video Deep Research Benchmark on Open Web for Agentic Video Reasoning

Anonymous ACL submission

Abstract

In real-world video question answering scenarios, videos often provide only localized visual cues, while verifiable answers are distributed across the open web; models therefore need to jointly perform cross-frame clue extraction, iterative retrieval, and multi-hop reasoning-based verification. To bridge this gap, we construct the first video deep research benchmark, VideoDR. VideoDR centers on video-conditioned open-domain video QA, requiring cross-frame visual anchor extraction, interactive web retrieval, and multi-hop reasoning over joint video–web evidence; through rigorous human annotation and quality control, we obtain high-quality video deep research samples spanning six semantic domains. We evaluate multiple closed-source and open-source MLLMs under both the Workflow and Agentic paradigms, and the results show that Agentic is not consistently superior to Workflow: its gains depend on a model's ability to maintain the initial video anchors over long retrieval chains. Further analysis indicates that goal drift and long-horizon consistency are the core bottlenecks. In sum, VideoDR provides a systematic benchmark for studying video agents in open-web settings and reveals the key challenges for next-generation video deep research agents.

1 Introduction

In existing multimodal evaluations, video remains a significant weakness (Li et al., 2024a; Fang et al., 2024; Wu et al., 2024; Jang et al., 2024). On the one hand, video reasoning inherently requires cross-temporal cue tracking and spatiotemporal modeling (Wu et al., 2024; Yang et al., 2025b; Chen et al., 2025a); on the other hand, most existing evaluations adopt a closed-evidence setting, where models typically only need to answer questions within the given video, without interacting with evidence on the open web (Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025a; Yang et al., 2025a). As a result, the capability of using videos as clues and completing fact verification and reasoning synthesis on open webpages has not been systematically characterized.

Figure 1: An example of the VideoDR task: identifying a museum via video visual cues, then using multi-hop search to find the closest "don't miss" exhibit to the entrance and outputting its accession number WB.67.

Meanwhile, deep research agents are pushing question answering from static contexts toward active evidence exploration on the open web: instead of answering directly from a given context, systems must conduct multiple rounds of search, filtering, and cross-checking in real web environments, ultimately producing conclusions grounded in evidence (Chen et al., 2025b; Zheng et al., 2025). A large number of deep research benchmarks have emerged around this capability (Li et al., 2025a; Jin et al., 2025). However, overall, these benchmarks still mostly start from textual queries (Wei et al., 2025; Wu et al., 2025; Li et al., 2025b); even when multimodal information is introduced, visual content is often treated as static auxiliary information rather than key evidence that must be precisely tracked and propagated (Jiang et al., 2024).

However, in real use, videos often carry decisive clues. User questions about videos are typically

open-domain factoid questions: the knowledge to be verified does not directly appear in the video or its title, but is distributed across large and dynamically changing web corpora; meanwhile, the key clues relevant to the question lie along the video timeline and must be extracted through cross-frame association. In this research pattern, videos provide localized visual cues, and webpages provide verifiable answers—yet it is not directly covered by existing deep research benchmarks that take text as input (Wei et al., 2025), nor by video benchmarks that assume evidence is closed within the video (Li et al., 2024a; Fang et al., 2024; Zhou et al., 2025).

Based on these gaps, we propose VideoDR, the first open-domain benchmark that systematically evaluates video deep research. As shown in Figure 1, we extend deep research to an open-domain factoid question answering setting conditioned on video: models must extract and compose visual anchors from multiple frames (Yang et al., 2025b; Chen et al., 2025a; Zhang et al., 2025), use browser-based search to locate candidate evidence on the open web (Wei et al., 2025; Wu et al., 2025; Li et al., 2025b; Chen et al., 2025b; Zheng et al., 2025), and perform multi-hop reasoning in the joint evidence space of videos and webpages to output a unique and verifiable answer (Liang et al., 2025). To support this setting, we explicitly incorporate an interactive web search process into the task definition and adopt annotation and strict quality control over diverse real-world scenarios to systematically remove samples that can be answered by the video alone or by webpages alone, making the combined capability of video understanding, web search, and evidence-based reasoning the core evaluation target (Yao et al., 2022; Gao et al., 2023). Based on VideoDR, we evaluate mainstream models under both the Workflow and Agentic paradigms (Liu et al., 2025), including closed-source models GPT-4o (Hurst et al., 2024) and Gemini-3-pro-preview (Team et al., 2023), as well as open-source models MiniCPM-V 4.5 (Yu et al., 2025b), Qwen3-Omni-30B-A3B (Xu et al., 2025), and InternVL3.5-14B (Wang et al., 2025b), and conduct comprehensive performance and error analyses along three dimensions: difficulty, video duration, and semantic domain.

Our main contributions are as follows:

❶ **Video Deep Research Task**: We first define the Video Deep Research task, shifting video understanding from closed-context perception to active, multi-hop search and reasoning on the open web anchored by video cues.

❷ **VideoDR Benchmark**: We construct a high-quality VideoDR benchmark through rigorous human annotation and quality control. VideoDR benchmark ensures that the multi-step evidence gathering process maintains a strong dependency on multi-frame visual cues within the video.

❸ **Agent Capability Boundaries**: By benchmarking leading MLLMs across Workflow and Agentic paradigms, we delimit the capability boundaries of these two agentic approaches. Leveraging the diverse distribution of VideoDR across semantic domains, question lengths, and video durations, we systematically analyze the performance and error patterns under different paradigms. Our findings reveal that Goal Drift and Long-horizon Consistency are the core bottlenecks constraining the development of next-generation video deep research agents.

2 Related Work

Deep Research Benchmarks. Existing deep research evaluations typically assess search, reasoning, and tool use as a unified process (Wei et al., 2025; Wu et al., 2025; Li et al., 2025b; Zheng et al., 2025): one line of work tests multi-step query planning and information localization in live web environments (Wei et al., 2025; Wu et al., 2025), while another studies more controllable settings to improve reliability and stability (Xue et al., 2025; Chen et al., 2025b; Jin et al., 2025; Li et al., 2025a). Overall, however, most benchmarks still start from textual queries (Wei et al., 2025; Wu et al., 2025), and visual content is often downplayed as static auxiliary information rather than first-class evidence in the retrieval and verification loop (Jiang et al., 2024; Liang et al., 2025; Liu et al., 2025).

Video Reasoning Benchmarks. Existing video reasoning evaluations are conducted under a closed-evidence setting, where questions are designed to be answerable using only the video itself, and they primarily stress long-video temporal understanding and long-context reasoning (Fu et al., 2025; Li et al., 2024a; Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025a; Fang et al., 2024; Yang et al., 2025a; Li et al., 2024b; Nagrani et al., 2024; Yu et al., 2025a). Recent agentic video efforts also

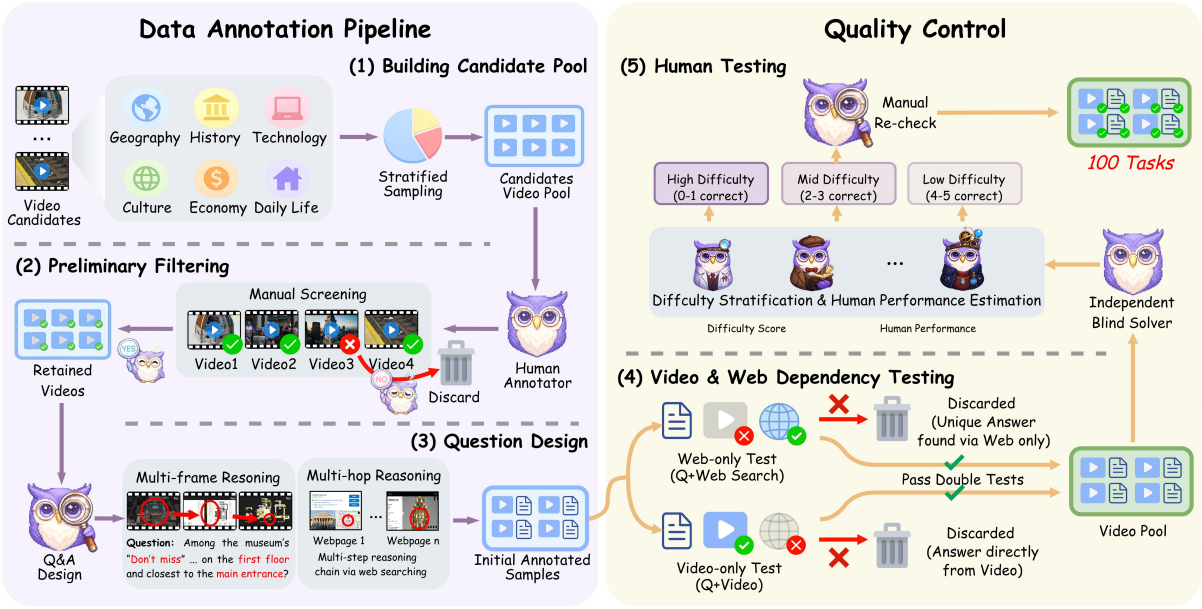

Figure 2: Overview of the VideoDR construction pipeline.

explore multi-round interaction and tool calling within the video-understanding loop (Zhang et al., 2025; Yang et al., 2025b; Chen et al., 2025a). In summary, systematic evaluations of multi-step evidence gathering and reasoning integration on the open web anchored to video cues remain relatively scarce.

3 Video Deep Research

3.1 Task Definition

We propose **Video Deep Research (VideoDR)**: an open-domain factoid question answering benchmark conditioned on a given video, designed to evaluate a model's ability to perform complex reasoning anchored in the video while leveraging the open web. Given a video V and a natural language question Q, the model can interactively call a browser search tool S, iteratively searching between video cues and web-page evidence, and finally output a factual answer A. Each VideoDR sample is constructed such that the model must exploit multi-frame cues from the video to locate candidate evidence on the open web, and then perform multi-hop reasoning over the joint evidence space of the video and the open web in order to obtain a unique answer. Formally, the VideoDR task can be expressed as:

$$f : (V, Q; S) \rightarrow A.$$

3.2 Data Annotation Process

In the data construction stage, we recruited three annotators with experience in video understanding and web search to create questions and annotate answers. Each annotator must actively locate several multi-frame visual cues in the video and design corresponding multi-hop questions and answers around these cues. To ensure annotation consistency, we provided unified annotation guidelines and examples before the annotation phase. As shown in Figure 2, the overall annotation pipeline consists of three steps.

Candidate Video Pool. Annotators first select videos from different platforms and then perform stratified sampling along three dimensions: source, domain, and duration, to cover diverse real-world scenarios. To ensure data quality, we apply a strict negative filtering strategy to remove the following three types of videos: ❶ single-scene clips with highly redundant visual semantics; ❷ popular topics whose information is overly explicit and can be obtained via text search without watching the video; ❸ isolated content on the open web for which no verifiable chain of evidence can be found.

Initial Filtering. At the initial screening stage, we manually remove clips that lack prominent *visual anchors*. We only retain videos that exhibit coherent visual cues at multiple time points and are suitable for cross-frame association, as candidates for subsequent annotation. For longer videos,

annotators are allowed to extract multiple semantically focused segments from the same video and construct separate questions for each segment.

Question Design. As shown in Figure 1, in this stage, annotators are required to design high-complexity questions for the retained video segments, under two strict constraints: ❶ **Multi-frame reasoning:** Each question must be grounded in video cues spanning multiple frames; it should be impossible to answer the question from a single-frame screenshot. ❷ **Multi-hop reasoning:** Each question must implicitly admit a decomposable multi-step reasoning path, forcing the model to perform at least one round of information exchange between video perception and external search. To ensure verifiability, we archived the web pages containing the key evidence supporting each answer.

3.3 Quality Control

To ensure that VideoDR is rigorous and unambiguous, we apply a two-stage quality verification procedure to the annotated samples.

3.3.1 Video & Web Dependency Testing

To verify that the annotated samples simultaneously depend on both the video and search, we design two ablation settings and conduct tests on all samples. Only samples that fail under both of the following conditions are retained as valid:

Web-only Test. The annotator is only given the textual question Q and can conduct the web search. If a unique and unambiguous answer can be obtained purely through web search, the question is deemed to lack dependence on visual anchors and is discarded.

Video-only Test. The annotator answers the question by watching only the video V. If the answer can be directly obtained from the video information alone, the question degenerates into a pure video understanding task and is discarded.

3.3.2 Human Testing

To verify the correctness of annotated questions and characterize their difficulty, we adopt a human evaluation protocol with five independent participants solving the tasks in a blind manner. For each sample i, five subjects independently produce answers under the condition that they can only access the annotated video and the textual question, without being given any reference answers, forming the answer set $\{a_{i1}, \ldots, a_{i5}\}$. Subjects are required to complete the task in a browser environment and must, for each question, submit a final answer obtained via video browsing and autonomous web search. This process serves two purposes: ❶ to validate the solvability of the questions and the correctness of the annotations under conditions close to real-world usage; and ❷ to provide empirical evidence for subsequent difficulty stratification and estimation of the human upper bound.

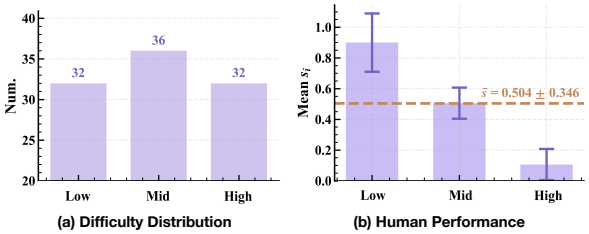

Figure 3: Human solvability across benchmark difficulty levels.

Difficulty. To quantify question difficulty at the sample level, we define the difficulty score of sample i based on the human success rate from the five-subject blind evaluation:

$$s_i = \frac{1}{5} \sum_{j=1}^{5} \mathbf{1}[a_{ij} \equiv A_i],$$

where a_{ij} denotes the answer provided by the j-th subject for sample i, A_i is the gold answer for that sample, and $\mathbf{1}[\cdot]$ is the indicator function (1 if the prediction is semantically equivalent to the gold answer, and 0 otherwise). We treat the distribution of $\{s_i\}$ as a proxy for question difficulty and stratify samples by the number of correct human answers: if only 0–1 out of 5 subjects answer correctly, the sample is labeled *High*; if 2–3 subjects answer correctly, it is labeled *Mid*; and if 4–5 subjects answer correctly, it is labeled *Low*. This difficulty partition is used in subsequent experiments to analyze how different models perform across human difficulty levels. Figure 3(a) shows the distribution of samples across three difficulty levels.

Human performance. Under the above setting, we estimate the human upper bound by the average accuracy of the 5 subjects over all samples:

$$\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s_i.$$

Figure 3(b) shows the human test results. Across all 100 samples, the mean sample-level success rate

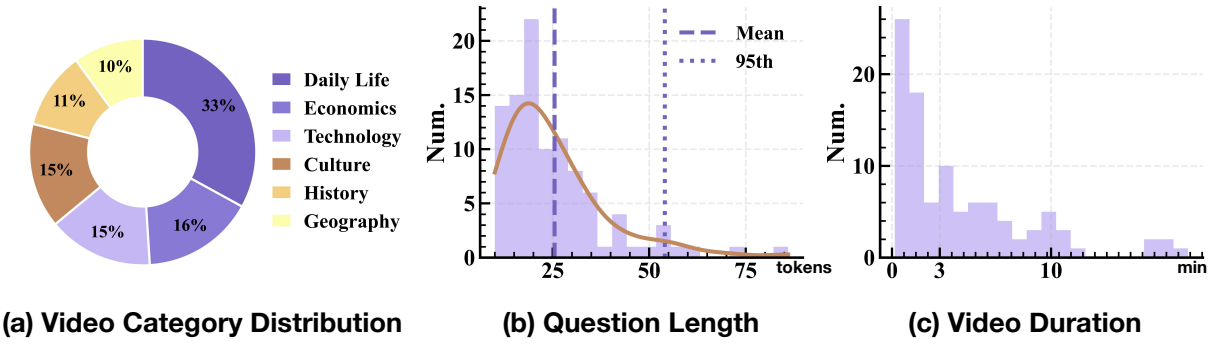

Figure 4: Data statistics of VideoDR, including (a) video category, (b) question length, and (c)video duration.

is $\bar{s} = 0.504$ with a standard deviation of 0.346. When results are stratified by difficulty, the mean s_i is 0.900 ± 0.190 for Low, 0.506 ± 0.101 for Mid, and 0.106 ± 0.101 for High. Overall, human participants can reliably solve the majority of samples, and we do not observe a systematic pattern of universal failure, which indirectly supports the consistency of the annotated answers. In addition, during dataset construction, we manually re-examine samples with clear disagreement or samples where multiple mutually inconsistent yet seemingly reasonable answers are provided; if ambiguity remains after multiple rounds of verification, the sample is discarded. The final retained dataset ensures that each question corresponds to a unique and verifiable gold answer, given the available evidence, thereby reducing evaluation noise.

3.4 Data Statistics

Finally, we construct 100 VideoDR samples. To characterize the benchmark's structural properties, and to provide a basis for stratified analyses in subsequent experiments, we summarize the dataset from three perspectives: ❶ video category distribution, ❷ question length distribution, and ❸ video duration distribution.

Video category Distribution. As shown in Figure 4(a), VideoDR spans six semantic domains: Daily Life, Economics, Technology, Culture, History, and Geography. The distribution is relatively balanced: Daily Life accounts for 33%, Economics 16%, Technology and Culture 15% each, History 11%, and Geography 10%. This distribution ensures broad coverage of diverse content types in open-domain settings.

Question Length. As shown in Figure 4(b), we compute the token-length distribution of the natural-language questions. The average question length is 25.54 tokens, and the lengths are concentrated: the 95th percentile is 54 tokens. This indicates that most questions are compact in phrasing, maintaining concise language while preserving necessary constraints. As a result, evaluation can focus more on the core process of starting from video cues and combining web evidence to perform multi-hop reasoning, rather than on extra comprehension burden induced by overly long inputs.

Video duration. As shown in Figure 4(c), on the video side, durations exhibit a clear long-tailed pattern: most videos cluster in a short-duration range, while a small number of videos longer than 10 minutes form the tail. This duration structure covers both rapid cue capture in short-video scenarios and cross-segment association in long-video scenarios, enabling a more comprehensive test of models' cross-frame understanding and evidence localization across different time scales.

4 Experiments

4.1 Setting

Baselines. We compare two common paradigms: Workflow and Agentic. Workflow adopts a two-stage design: a multimodal model first extracts question-relevant cross-frame visual cues from the video and produces a structured intermediate text, which is then provided together with the question as input for subsequent reasoning. Without further access to the original video, the system uses the search tool to retrieve candidate evidence from the open web, and uses the think tool to reflect on and filter evidence and to generate the next-round query, finally aggregating multi-round search results to produce an answer. In contrast, Agentic adopts a stronger end-to-end setting: it feeds the raw video and the question directly into a single multimodal agent, which performs video understanding, query

Table 1: Performance comparison across difficulty levels under the **Workflow** and the **Agentic** settings.

Model	Setting	Difficulty (%)			
		Low	Mid	High	Ave.
#Samples		32	36	32	100
Qwen3-Omni-30B-A3B	Workflow	59.38	30.56	21.88	37.00
	Agentic	65.62	27.78	18.75	37.00
InternVL3.5-14B	Workflow	37.50	22.22	21.88	27.00
	Agentic	46.88	22.22	21.88	30.00
MiniCPM-V 4.5	Workflow	50.00	13.89	12.50	25.00
	Agentic	18.75	19.44	9.38	16.00
Gemini-3-pro-preview	Workflow	90.62	61.11	56.25	69.00
	Agentic	93.75	69.44	65.62	76.00
GPT-4o	Workflow	50.00	30.56	46.88	42.00
	Agentic	62.50	38.89	28.12	43.00
Human		90.00	50.56	10.63	50.40

Table 2: Performance comparison across different video durations under the **Workflow** and the **Agentic** settings.

Model	Setting	Video Duration (%)			
		Short	Medium	Long	Ave.
#Samples		52	38	10	100
Qwen3-Omni-30B-A3B	Workflow	34.62	36.84	50.00	37.00
	Agentic	38.46	39.47	20.00	37.00
InternVL3.5-14B	Workflow	36.54	15.79	20.00	27.00
	Agentic	32.69	28.95	20.00	30.00
MiniCPM-V 4.5	Workflow	30.77	15.79	30.00	25.00
	Agentic	17.31	15.79	10.00	16.00
Gemini-3-pro-preview	Workflow	75.00	65.79	50.00	69.00
	Agentic	71.15	84.21	70.00	76.00
GPT-4o	Workflow	46.15	36.84	40.00	42.00
	Agentic	51.92	31.58	40.00	43.00
Human		51.92	50.53	42.00	50.40

generation, web retrieval, and evidence integration within the same execution loop. The agent likewise uses only the think tool and the search tool, and autonomously decides when to initiate search based on video cues, when to reflect, and how to organize multi-round evidence into the final conclusion.

MLLMs. We select mainstream multimodal models spanning both closed-source and open-source families as the core research agents. Closed-source models include Gemini-3-pro-preview (Team et al., 2023) and GPT-4o (Hurst et al., 2024); open-source models include MiniCPM-V 4.5 (Yu et al., 2025b), InternVL3.5-14B (Wang et al., 2025b), and Qwen3-Omni-30B-A3B (Xu et al., 2025) . All models are tested under both Workflow and Agentic paradigms to analyze their capability boundaries under different system organizations.

LLM as Judge. VideoDR is an open-domain factual question answering task. To avoid false mismatches, we adopt an LLM-as-judge protocol (Zheng et al., 2023) using DeepSeek-V3-0324 (Liu et al., 2024) to assess semantic equivalence between the model prediction and the reference answer. The judge outputs a binary correctness label, which we use to compute overall Accuracy.

4.2 Main Results

From the overall averages, our VideoDR setting yields a clear capability stratification among existing models: Gemini-3-pro-preview leads by a large margin under both paradigms (Workflow 69% / Agentic 76%), GPT-4o consistently forms the second tier (42% / 43%), and the three open-source models are overall weaker (Qwen3-Omni-30B-A3B 37% / 37%, InternVL3.5-14B 27% / 30%, MiniCPM-V 4.5 25% / 16%).

Difficulty stratification reveals the essential differences between the two paradigms. As shown in Table 1, all evaluated models exhibit a highly consistent decline in performance as the difficulty level increases from Low to Mid and then to High. Human performance decreases from 90.00% to 50.56% and then to 10.63%, and models also generally degrade as difficulty increases, suggesting that our difficulty definition based on human success rate indeed corresponds to longer and more fragile evidence chains. Gains from Agentic mainly appear in models with stronger state maintenance capabilities, and become more pronounced toward Mid/High; conversely, mid-tier models can suffer a clear backlash on High. For example, Gemini-3-pro-preview continues to improve on Mid/High from 61.11%→69.44% and 56.25%→65.62%; but while GPT-4o improves on Low/Mid (50.00%→62.50%, 30.56%→38.89%), it drops sharply on High (46.88%→28.12%). This phenomenon is highly consistent with our implementation constraints: in Agentic, the model cannot re-watch the video during research, so the subsequent process must rely solely on cues obtained from the initial viewing to guide multi-round search and reasoning, lacking opportunities for online correction via revisiting the video. Once early visual anchors drift, web noise can amplify the deviation in later searches, making High-difficulty questions more prone to goal drift; in contrast, Workflow externalizes video cues into an intermediate textual representation, effectively providing a repeatedly accessible external memory that reduces drift risk in long-horizon decision-making.

Table 3: Performance comparison across different domains under the **Workflow** and the **Agentic** settings.

Model	Setting	Domain (%)						
		History	Geography	Culture	Economy	Technology	Daily Life	Ave.
#Samples		11	10	15	16	14	33	100
Qwen3-Omni-30B-A3B	Workflow	36.36	30.00	26.67	43.75	50.00	36.36	37.00
	Agentic	54.55	40.00	26.67	43.75	35.71	33.33	37.00
InternVL3.5-14B	Workflow	9.09	50.00	20.00	25.00	21.43	30.30	27.00
	Agentic	36.36	40.00	26.67	31.25	28.57	24.24	30.00
MiniCPM-V 4.5	Workflow	27.27	10.00	46.67	25.00	14.29	24.24	25.00
	Agentic	9.09	10.00	26.67	12.50	14.29	18.18	16.00
Gemini-3-pro-preview	Workflow	72.73	70.00	80.00	62.50	64.29	69.70	69.00
	Agentic	81.82	50.00	86.67	68.75	85.71	78.79	76.00
GPT-4o	Workflow	63.64	40.00	33.33	43.75	42.86	39.39	42.00
	Agentic	63.64	20.00	53.33	50.00	35.71	39.39	43.00
Human		36.36	34.00	49.33	56.25	60.00	52.73	50.40

Video-duration stratification further reinforces this explanation: as videos get longer, cues become more dispersed, and the state space grows, so Agentic places higher demands on a model's ability to retain and continuously leverage initial video cues over the long run, leading to strong polarization. As shown in Table 2, Gemini-3-pro-preview achieves substantial gains on Medium/Long (65.79%→84.21%, 50.00%→70.00%), indicating that strong models can preserve enough discriminative details from one viewing and keep using them in subsequent search; in contrast, Qwen3-Omni-30B-A3B and MiniCPM-V 4.5 drop markedly on Long (50.00%→20.00%, 30.00%→10.00%), suggesting that when a model struggles to stably preserve and continuously utilize key video anchors across multiple rounds of search and reasoning, directly using the video does not translate into better search decisions and instead makes the system more likely to destabilize over long chains. In other words, long videos amplify the key trade-off between the two paradigms: Workflow may lose details but is more controllable, whereas Agentic preserves details more faithfully but relies more on long-term consistency.

Domain stratification highlights the respective advantages of the two paradigms across domains. As shown in Table 3, in Technology, Gemini-3-pro-preview shows the largest Agentic-over-Workflow improvement (64.29%→85.71%), indicating that such questions often require translating fine-grained visual cues into high-precision queries, and Workflow's one-shot summary is more likely to miss crucial search triggers. However, in Geography, both Gemini-3-pro-preview (70.00%→50.00) and GPT-4o (40.00%→20.00) degrade noticeably, suggesting that geographic search is more ambiguous and therefore depends more on clear and consistent visual anchors to keep the search target stable.

Table 4: Tool-use statistics under the **Workflow** and **Agentic** paradigms.

Model	Paradigm	`think`	`search`	time (s)
Qwen3-Omni-30B-A3B	Workflow	1.82	0.95	222.86
	Agentic	1.80	1.21	367.19
InternVL3.5-14B	Workflow	1.58	1.48	214.28
	Agentic	1.20	1.24	228.05
MiniCPM-V 4.5	Workflow	0.94	1.13	242.21
	Agentic	1.97	2.07	139.07
Gemini-3-pro-preview	Workflow	2.40	1.86	422.48
	Agentic	2.89	2.52	449.44
GPT-4o	Workflow	1.90	1.11	136.15
	Agentic	1.31	1.17	181.87

Across all three stratifications, we conclude that under current mainstream model capabilities, Agentic is not necessarily superior to Workflow: its effectiveness depends on whether the model can continually rely on initial video cues throughout multi-round search and reasoning; meanwhile, Workflow provides stable anchors for downstream search and reasoning through an explicit intermediate text, thereby more effectively reducing the risk that the search process drifts away from the correct target.

4.3 Tool-use Analysis

As shown in Table 4, the number of tool calls is not monotonically correlated with accuracy. What determines performance is not using tools more, but whether a model can turn a limited number of search calls into a high-quality evidence chain. Concretely, Gemini-3-pro-preview demonstrates stronger tool-use effectiveness: under the Agen-

Table 5: Error type distribution across different models under both the **Workflow** and the **Agentic**.

Model	Setting	Error Type Count								
		Categorical	Incomplete	Not Found	Numerical	Context	Semantic	Reasoning	Others	Total
Qwen3-Omni-30B-A3B	Workflow	22	16	6	7	11	0	0	1	63
	Agentic	16	20	11	11	3	1	1	0	63
InternVL3.5-14B	Workflow	25	18	20	9	1	0	0	0	73
	Agentic	29	17	15	7	0	1	1	0	70
MiniCPM-V 4.5	Workflow	25	23	13	10	2	1	1	0	75
	Agentic	36	13	14	12	4	1	1	3	84
Gemini-3-pro-preview	Workflow	5	10	4	9	0	2	0	1	31
	Agentic	7	8	1	6	1	0	1	0	24
GPT-4o	Workflow	22	11	8	10	2	2	0	3	58
	Agentic	18	20	9	8	1	1	0	0	57

tic setting, it averages 2.89/2.52 think/search calls (449s runtime) and achieves the best overall score (76%), indicating that additional retrieval and reflection are indeed converted into more reliable evidence integration. In contrast, Qwen3-Omni-30B-A3B's extra search overhead does not yield corresponding performance gains: its think usage is nearly identical between Agentic and Workflow (1.80 vs. 1.82), but Agentic has higher search usage and substantially longer runtime (1.21, 367s vs. 0.95, 223s), while accuracy remains unchanged (both 37%), suggesting that the extra searches mostly fall into low-yield exploration or failed evidence filtering. An even sharper contrast is MiniCPM-V 4.5: under Agentic it uses tools more frequently (think/search 1.97/2.07) and runs faster (139s), yet accuracy drops from 25% (Workflow) to 16%; combined with its across-the-board decline on High difficulty in Table 1 (12.50%→9.38%), this indicates that it provides a stable reference for downstream research via an explicit intermediate text representation, making search and reasoning more focused on relevant evidence and thus less likely to drift.

4.4 Error Analysis

To obtain a more fine-grained understanding of MLLMs' behaviors under different paradigms, we conduct a detailed analysis of their error trajectories across all experiments. We further break down the errors into eight categories, as shown in Table 5.

Perception Grounding Limitation. Table 5 indicates that Reasoning Error is nearly absent under both paradigms, with counts ranging from 0 to 1 in Workflow and likewise 0 to 1 in Agentic. In contrast, Categorical Error remains the dominant failure mode across all models, spanning 5–25 cases in Workflow and 7–36 cases in Agentic, and it increases for some models when switching to Agentic (MiniCPM-V 4.5 rises from 25 to 36). This pattern is closely tied to our setting: once the research phase cannot revisit the video, downstream retrieval and reasoning depend entirely on the visual anchors extracted from the first pass. When early perception deviates, there is no mechanism to re-localize key frames and correct the anchor, making error propagation along the evidence chain more likely.

Numerical Error Bottleneck. Table 5 shows that Gemini's overall error count is substantially lower in the Agentic setting (24 total errors, compared with 57 for GPT-4o), yet this advantage does not carry over to Numerical Error. Under Workflow, Gemini-3-pro-preview has 9 numerical errors, while Qwen3-Omni-30B-A3B, InternVL3.5-14B, MiniCPM-V 4.5, and GPT-4o have 7, 9, 10, and 10, respectively. Under Agentic, Gemini-3-pro-preview records 6 numerical errors, compared with 11, 7, 12, and 8 for the same four models. The numerical-error counts, therefore, remain tightly clustered across models, suggesting that numerical reliability constitutes a distinct and persistent weakness for current MLLMs.

5 Conclusion

We introduce VideoDR, a video deep research benchmark for evaluating multimodal deep research on the open web. It requires models to extract multi-frame visual anchors, turn them into searchable queries, retrieve open-web evidence, and perform multi-hop reasoning to produce verifiable factual answers. We benchmark mainstream multimodal models under Workflow and Agentic paradigms, with stratified analyses by difficulty, video duration, and semantic domain to characterize capability boundaries on video deep search.

6 Limitations

During the annotation process, while we enforced a strict quality control protocol and verified the uniqueness of final answers, the intermediate search queries and reasoning paths were derived from the subjective search behaviors of our expert annotators. In real-world scenarios, different users might employ diverse keywords or browsing strategies to locate the same evidence. Although this does not affect the objective verifiability of the VideoDR samples, the current benchmark primarily reflects a specific set of high-efficiency research trajectories. Future work could involve collecting a wider variety of human search logs to better model the diversity of user-agent interactions.

References

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025a. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv preprint arXiv:2503.10200*.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, and 1 others. 2025b. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, and 1 others. 2024. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024b. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Paolo Rota, Nicu Sebe, Zheng Liu, and Lizi Liao. 2025. Video-browsecomp: Benchmarking agentic video research on open web. *arXiv preprint arXiv:2512.23044*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, and 1 others. 2025. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, and 1 others. 2024. Neptune: The long orbit

to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

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

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and 1 others. 2025. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. 2025. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. 2025a. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. 2025b. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, and 1 others. 2025a. Vrbench: A benchmark for multi-step reasoning in long narrative videos. *arXiv preprint arXiv:2506.10857*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025b. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. 2025. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701.