GROUNDEDVDR: GROUNDING VIDEO-TO-WEB EVIDENCE DEPENDENCIES IN VIDEO DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large multimodal models have improved at identifying cross-frame visual evidence and conducting open-web retrieval, as reflected in recent video deep research (VDR) benchmarks. However, existing VDR benchmarks primarily evaluate tool-use trajectories and final responses without verifying whether models use temporal video information or retrieved web evidence. We introduce GROUNDEDVDR, an evidence-grounded benchmark with three features. First, it combines human review and model testing to exclude questions that can be answered using parametric knowledge alone. Second, it provides structured video-to-web evidence annotation that links related video frames with their corresponding web evidence, while recording candidate-linked hard negatives for diagnostic analysis. Third, it measures video-state localization, evidence retrieval, video-to-web evidence alignment, and final answer synthesis separately. Together, these features allow GROUNDEDVDR to evaluate both answer correctness and support from the required video and web evidence.

1 INTRODUCTION

Most existing video understanding benchmarks focus solely on reasoning over limited information within the video (Li et al., 2024; Wu et al., 2024; Fu et al., 2025; Zhou et al., 2025; Fang et al., 2024). Those questions can be easily answered from static frames or captions, rather than testing genuine deep research capabilities of video large multimodal models (Video-LMMs). In reality, understanding a video often requires jointly utilizing omni-modal evidence—such as audio, on-screen text, and temporal dynamics. Furthermore, the answers to complex questions frequently reside outside the video (e.g., on the open web).

Video Deep Research (VDR) addresses this limitation. A VDR agent must locate key evidence across the video, translate it into constraints for open-web retrieval, and synthesize a verifiable answer. This differs from traditional deep-research tasks, which rely solely on self-contained text queries (Wei et al., 2025; Wu et al., 2025; Li et al., 2025; Zheng et al., 2025). In VDR, the text query alone is insufficient because the agent must extract clues from the video to guide its research. The main challenge is therefore to ensure that video evidence determines what must be researched, rather than merely combining video analysis and web search.

Recent studies propose agent-based methods and benchmarks for VDR. VideoDR (Liu et al., 2026a) formalizes the Video Deep Research (VDR) setting. It requires models to extract cross-frame visual evidence and use it as constraints for open-web retrieval. Human annotators conduct ablation tests and discard questions that can be solved using only one modality. VideoSearcher (Gao et al., 2026) introduces the VideoSearch-QA benchmark to evaluate fine-grained tool usage. It requires agents to actively perform temporal localization, spatial zooming, and reverse image search to accurately capture visual evidence. Furthermore, Video-DeepResearch (Fang et al., 2026) identifies two failure modes in current multimodal agents: "modality bias" (i.e., a preference for text search over visual tools) and "parametric knowledge leakage" (i.e., using internal memory instead of the video and retrieval tools). Its VideoDR-Bench requires cross-frame visual grounding before web search. Together, these studies extend both VDR methods and their evaluation.

However, evaluating VDR systems remains difficult because conventional metrics are insufficient: final-answer accuracy does not distinguish evidence-based research from parametric recall, while

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.

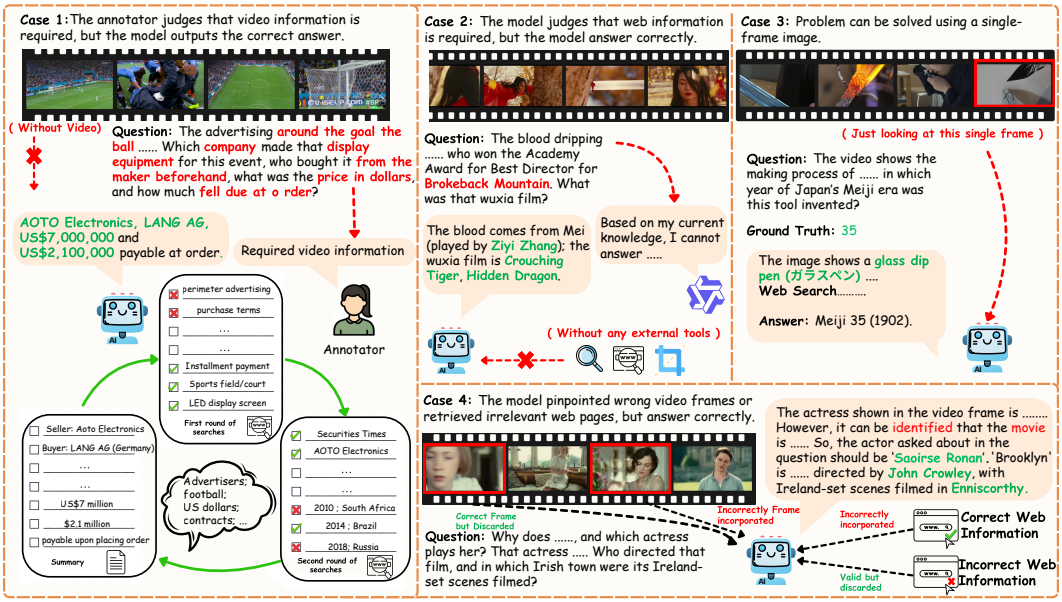

Figure 1: **Failure cases in video deep research (VDR).** Case 1 bypasses the video, and Case 2 bypasses web retrieval, illustrating variation in human and model judgments of modality dependence. Case 3 reduces the task to single-frame entity recognition rather than video-guided search. Case 4 produces the correct final answer from an incorrect video segment and irrelevant web evidence. Green text marks the correct final answers generated by the models.

tool-use statistics count actions without verifying their relevance. Specifically, existing VDR benchmarks share three drawbacks, as shown in Figure 1, with each rooted in a distinct challenge. The **first** challenge is deciding whether a question requires both video and web evidence. Existing benchmarks typically rely on either human annotators (Liu et al., 2026a) or model-assisted filtering and judging (Gao et al., 2026; Fang et al., 2026). These checks depend on the prior knowledge and capability of the evaluator. As shown in Figure 1, a model might bypass the intended video (Case 1) or web retrieval (Case 2) entirely.

The **second** challenge is ensuring that the video constrains the web search rather than merely providing a keyword. In some benchmarks, recognizing a single visual entity is enough to find the answer (Case 3).

The **third** challenge is the validity of the research process. An agent might locate the wrong video segment, retrieving irrelevant webpages and misinterpreting the evidence. Despite these errors, it can still guess the correct answer(Case 4). However, final-answer accuracy is misleading while simply counting tool uses is insufficient. These metrics cannot prove whether the agent actually relies on the correct evidence. In addition, they fail to reveal where the reasoning chain breaks down. To address these challenges, we introduce GROUNDEDVDR, an evidence-grounded benchmark that evaluates the video-to-web dependency. First, we combine human review with multi-model testing. We test multiple models without web search to rule out answers based on parametric knowledge alone. Moreover, we conduct controlled experiments by hiding specific video or web information. This ensures the necessity of both modalities is verified across diverse models, avoiding the bias from a single tester. Second, each annotation contains question-specific components: relevant time segments and their observed content, search constraints derived from those observations, supporting web evidence, and evidence-backed incorrect candidates. Third, the evaluation protocol reports video grounding, web evidence retrieval and verification, and final answer synthesis separately, allowing us to identify the stage at which an error occurs.

In this work, our contributions are fourfold:

- We rethink VDR benchmark construction as a necessity-testing problem. Human review and multi-model probing require each task to depend on both temporal video grounding and web retrieval, reducing parametric-memory shortcuts.
- We introduce GROUNDEDVDR, a benchmark with structured evidence annotations that link visual observations to web search constraints and supporting web claims.
- We develop a stage-wise evaluation protocol that reports video localization, search formulation, and evidence extraction separately to identify where errors occur.
- We evaluate open- and closed-source multimodal agents and analyze modality bias and parametric leakage in existing VDR benchmarks and evaluation protocols.

2 RELATED WORK

Video Understanding. Existing video benchmarks evaluate perception and temporal reasoning under closed-evidence settings (Li et al., 2024; Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025). Recent agentic approaches actively search within long videos using adaptive sampling and tool use (Zhang et al., 2025; Yang et al., 2025; 2026; Chen et al., 2025). Notably, LongVidSearch enforces retrieval necessity by requiring discontinuous evidence clips (Yu et al., 2026). However, all these settings assume the necessary information resides entirely within the video. In contrast, our protocol requires agents to extract temporal video states and use them as strict constraints for subsequent open-web research.

Deep Research Agents. Text-based deep-research systems traverse the web and combine information from multiple sources (Wei et al., 2025; Wu et al., 2025; Li et al., 2025; Zheng et al., 2025). Recent studies extend this setting to multimodal search with visual reasoning and multiple retrieval tools (Jiang et al., 2024; Geng et al., 2025). ALCE evaluates whether retrieved sources support generated claims (Gao et al., 2023), while Vision-DeepResearch shows that visual-search evaluations can contain text-only shortcuts and parametric knowledge leakage (Zeng et al., 2026). These studies focus on static images; we derive the search constraints from temporally distributed video evidence.

Video Deep Research. Several recent studies combine video understanding with open-web research. VideoDR defines the VDR task and uses manual filtering to test dependence on both modalities (Liu et al., 2026a), while Video-BrowseComp evaluates video-dependent open-web browsing (Liang et al., 2025). VideoSearcher combines temporal localization, spatial inspection, and multimodal search (Gao et al., 2026); Video-DeepResearch uses a perception-before-retrieval procedure to reduce modality bias (Fang et al., 2026). These studies cover multi-frame reasoning, open-web retrieval, and agent tool use. As shown in Table 1, GROUNDEDVDR additionally annotates the links among video evidence, search constraints, and claim-level web evidence, and tests whether each link is necessary.

3 GROUNDEDVDR

3.1 PRELIMINARIES

Task Definition. We formulate video deep research (VDR) as a cross-modal *information-completion* task. An instance provides a video V, a question q, and access to the open web W. The question specifies a required information set $R(q)$: the minimal knowledge needed to determine the final answer y. This set comprises a video-grounded component $R_V(q)$ and a web-derived component $R_W(q)$. A complete solution must retrieve and synthesize this information. Consequently, a VDR system outputs an answer alongside its supporting evidence, reported as a tuple $(\hat{y}, \hat{\mathcal{E}}_V, \hat{\mathcal{E}}_W)$, where the retrieved video evidence $\hat{\mathcal{E}}_V$ and web evidence $\hat{\mathcal{E}}_W$ must sufficiently ground the required components $R_V(q)$ and $R_W(q)$.

Evidence Representation. To evaluate whether a system resolves $R(q)$, we record the required evidence as structured annotations. First, the required information from video $R_V(q)$ is formalized as a set of required video states $\mathcal{S}_q^\star = \{s_1, \ldots, s_M\}$. Each state s_j is a semantic sub-fact supported solely by the video and grounded to one or more minimal temporal evidence units

Table 1: **Comparison of benchmark construction, necessity tests, and evidence annotations. Seed Direction** gives the order in which the task is constructed, and **Annotation (Anno.)** identifies manual or hybrid construction. **Video-only/web-only tests (V/W Tests)** check whether either modality alone is sufficient. **Multi-model** indicates testing across multiple models without search access. **Multi-hop** and **Multi-state** indicate multiple reasoning steps and temporally distinct video states, respectively. **Video-to-web links (V→W Links)** connect localized video evidence to web retrieval conditions. **Claim-level evidence (Claim Evid.)** records supporting text from web sources, and **hard negatives (HN)** are evidence-backed incorrect candidates. ✓, ◦, and ✗denote full, partial, and absent support; N/A denotes not applicable.

	Basic Construction		Necessity Tests				Evidence Annotations		
Benchmark	**Seed Direction**	**Anno.**	**V/W Tests**	**Multi-model**	**Multi-hop**	**Multi-state**	**V→W Links**	**Claim Evid.**	**HN**
Image Deep-Research Benchmarks									
BrowseComp-VL (Geng et al., 2025)	Text→Image	Hybrid	✗	✗	✓	N/A	◦	✗	✗
MMSearch* (Jiang et al., 2024)	Query-first	Manual	✗	✓	◦	N/A	◦	◦	✗
VDR-Bench (Zeng et al., 2026)	Image-first	Hybrid	◦	◦	✓	N/A	◦	◦	✗
Video Deep-Research Benchmarks									
VideoDR (Liu et al., 2026a)	Video-first	Manual	✓	✗	✓	✓	◦	◦	✗
VideoSearcher (Gao et al., 2026)	Video-first	Manual	✓	✗	✓	◦	◦	◦	✗
Video-DeepResearch (Fang et al., 2026)	Video-first	Hybrid	◦	◦	✓	◦	◦	◦	✗
VideoRover-Bench (Liu et al., 2026b)	Target→Video	Hybrid	◦	◦	◦	◦	◦	◦	✗
GROUNDEDVDR (Ours)	**Video/Evidence-first**	**Manual**	✓	✓	✓	✓	✓	✓	✓

*The full MMSearch benchmark is mixed, containing 171 image-query instances and 129 text-only instances. The comparison here concerns its image-query setting.

$\mathcal{U}(s_j) = \{u_{j,1}, \ldots, u_{j,n_j}\}$. This separates semantics from localization: evidence units identify *where* the information occurs, while states specify *what* video-side fact is established. Second, the web requirement $R_W(q)$ is formalized as a set of required semantic claims $\mathcal{C}_q^W = \{c_1^W, \ldots, c_K^W\}$. Each claim captures a necessary external fact and is supported by a specific web source. Finally, video-to-web links specify how the required video states determine which external claims must be established. We also annotate evidence-backed incorrect candidates to test whether a model binds valid web evidence to the wrong candidate.

3.2 BENCHMARK CONSTRUCTION

We construct GROUNDEDVDR through the four-stage pipeline in Figure 2. **Video Selection** first retains clips with temporally separated evidence, multiple entities or events, and natural links to stable external sources, while filtering single-scene, text-leaking, trivia-dominated, or weakly verifiable videos (Figure 2(1)). **Video Evidence Mining** then inspects the timeline and aggregates minimal observations across multiple moments into video-side states and constraints, rather than fitting evidence to a pre-written question (Figure 2(2)). **Video-to-Web Task Curation** uses these video constraints to guide web research and question formation; annotators collect claim-level supporting evidence and plausible alternative candidates, including candidate-linked hard negatives when available (Figure 2(3)). Finally, **Verification and Release** checks temporal necessity, video-to-web dependence, evidence validity, and shortcut resistance before an instance enters the benchmark (Figure 2(4)).

3.3 QUALITY ASSURANCE

Every question undergoes quality control before release. An instance is valid only when the joint video–web evidence determines the answer and neither modality alone is sufficient. This verification-and-release stage is summarized in Figure 2(4) and implemented through the checks below.

Multi-Model Parametric Probing. To detect answers based on internal knowledge, we test every candidate under video-only and web-only ablations using multiple vision-language models. We use pass@k sampling (e.g., $k = 8, T = 0.7$) rather than a single greedy output. A fixed LLM judge compares the generated attempts with the ground-truth answer and its predefined aliases. If *any* model produces the correct answer under either ablation, the question is flagged for human review.

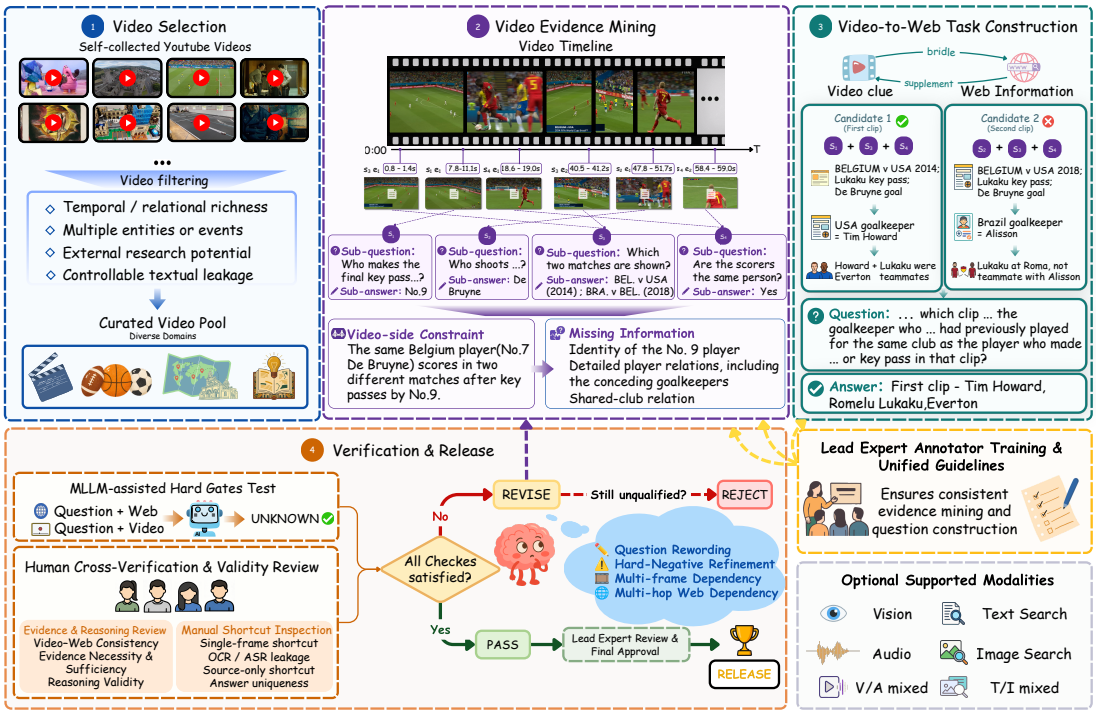

Figure 2: **Four-stage construction procedure for GROUNDEDVDR.** The stages are (1) video selection, (2) video evidence mining, (3) video-to-web task curation, and (4) verification and release. Questions are retained only when the annotated video evidence determines the web search and the supporting web claims pass automated checks and human review.

Independent Human Cross-Verification. Questions flagged by the probing phase, alongside all other candidate annotations, undergo blind review by independent annotators. For probed hits, reviewers inspect the generated evidence to determine whether the result stems from parametric leakage, optical character recognition (OCR), automatic speech recognition (ASR), accidental guessing, or judge error. Reviewers also check temporal necessity, evidence entailment, candidate naturalness, and whether a single webpage is sufficient. Any inter-annotator discrepancies are resolved by a third expert adjudicator.

Benchmark Profile. Figures 3 and 4 characterize the structure and capability composition of the released benchmark. Figure 3 shows that GROUNDEDVDR is predominantly multi-state, multi-claim, and temporally distributed, reflecting the intended video-to-web dependency rather than a single visual lookup followed by direct retrieval. Figure 4 provides a complementary view of the video-side reasoning requirements: primary VDR capabilities are typically supported by combinations of lower-level temporal, spatial, and relational capabilities, and many questions require several such capabilities jointly. Together, these profiles show that the benchmark emphasizes compositional video reasoning within a broader cross-modal research process.

3.4 EVALUATION METRICS AND PROTOCOL

Auditing the Research Process. We evaluate the submitted evidence against the structured video states and web claims defined in Section 3.1. For each question, **Video State** measures the proportion of its required video states that are supported by the submitted video evidence, while **Web Claim** measures the proportion of its required web claims that are supported by the submitted web evidence. Both metrics are computed per question and then macro-averaged across the benchmark, so that every question contributes equally regardless of how many states or claims it contains. We additionally characterize question-level grounding as **Complete**, **Partial**, or **None**, depending on whether all, some but not all, or none of the required states or claims are supported.

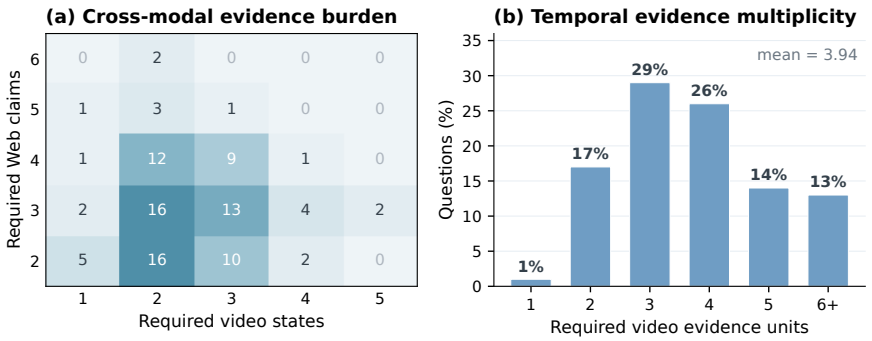

Figure 3: **Evidence Structure of GROUNDEDVDR.** (a) Joint distribution of the numbers of required video states and web claims; each cell reports the number of questions. (b) Distribution of temporally localized video evidence units required per question.

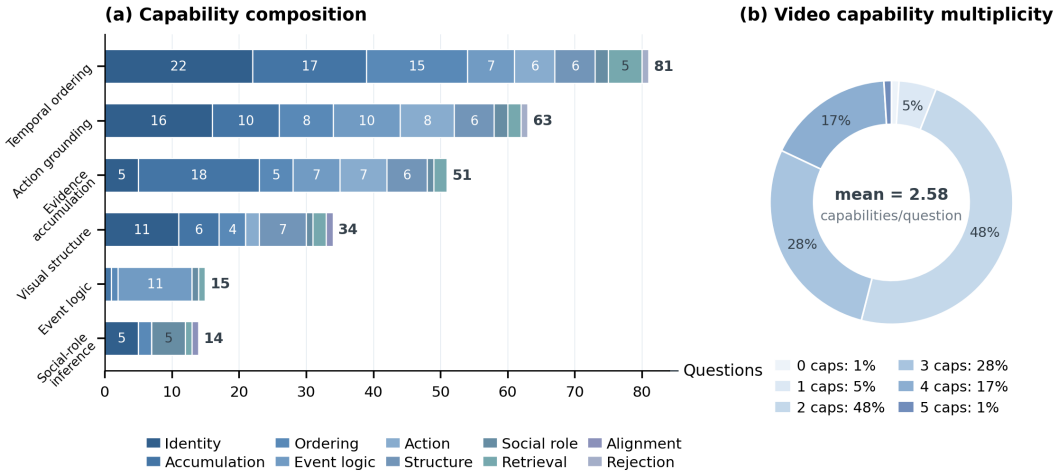

Figure 4: **Video capabilities required by GROUNDEDVDR.** (a) Supporting capabilities associated with each primary video deep research (VDR) capability. (b) Number of supporting video capabilities jointly required by each question.

Outcome Evaluation. We separately evaluate whether the final answer resolves the information requested by the question. For each question, **Field Acc.** measures the proportion of its required answer fields that are correctly resolved, and is likewise macro-averaged across questions. **Answer** is the stricter question-level measure: a question is counted as correct only when all of its required answer fields are correctly resolved. Field Acc. therefore gives partial credit to partially correct multi-field answers, whereas Answer measures complete answer resolution.

Semantic Judgement. Across all dimensions, we use a fixed LLM judge (GPT-5.6-Terra) with task-specific rubrics. The judge evaluates semantic support rather than requiring exact string matches or exact agreement with the annotated temporal boundaries. For video evidence, submitted temporal segments need not match the annotated time boundaries exactly, as long as they provide sufficient visual support for the required video state. For web evidence, alternative reliable sources and paraphrased text are accepted when they support the required claim; merely related pages and evidence attached to an incorrect candidate are rejected.

Table 2: **Main Results.** Answer metrics evaluate the final response, whereas Video and Web metrics measure recovery of the required semantic evidence. **Field Acc.** is the mean per-question proportion of correctly resolved required answer fields, while **Answer** requires all required answer fields of a question to be correct. **Video State** and **Web Claim** are the mean per-question recalls of required video states and Web claims, respectively. Best agent results in each column are shown in bold; the human-expert baseline is reported separately for reference.

Model	Field Acc.	Answer	Video State	Web Claim
Claude series				
Claude Opus 4.8	61.3	55.0	**43.5**	16.9
Claude Sonnet 5	61.6	55.0	31.2	17.9
GPT series				
GPT-5.6 Sol-High	**82.0**	**76.0**	37.6	**32.0**
GPT-5.6 Terra	55.5	47.0	20.5	10.6
Grok / Composer				
Grok-4.6	48.9	44.0	19.8	13.7
Composer-2.5	41.3	36.0	19.4	9.4
Qwen series				
Qwen3.8-Plus	62.3	54.0	33.8	19.2
Qwen3.8-Max	64.0	58.0	40.6	18.2
Human expert	92.8	90.0	77.2	63.4

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Platforms. We evaluate each model through an agent platform that supports video inspection, web browsing, and tool use. Every platform–model pair receives the same video, question, and task instructions, and is required to return a final answer together with supporting video and web evidence in a common structured format. We use each model's native agent platform where available, since tool-use behavior can depend on the surrounding agent system.

Models. We evaluate eight multi-modal agent configurations spanning several model families: Claude Opus 4.8 and Claude Sonnet 5 (Anthropic, 2026); GPT-5.6 Sol-High and GPT-5.6 Terra (OpenAI, 2026); Grok-4.6 (xAI, 2026) and Composer 2.5 (Anysphere, 2026); and Qwen3.8-Plus and Qwen3.8-Max (Qoder, 2026). All systems are evaluated on the same released 100-question benchmark.

Human-Expert Baseline. A human-expert baseline is evaluated under the same task definition, with access to the video and open web, and the exact same requirement to provide final answers alongside structured supporting evidence. To ensure a fair comparison against automated agents, human annotators are strictly prohibited from using conversational LLMs, AI agents, browser assistants, or automated research tools. Permitted utilities are strictly limited to standard web browsing , translation services, basic calculators, and manual video player controls. All submitted responses and reasoning summaries must be in English.

4.2 MAIN RESULTS

Table 2 summarizes final-answer performance together with video- and web-side evidence recovery. Among evaluated agents, GPT-5.6 Sol-High achieves the strongest answer performance and the

highest Web Claim recall, whereas Claude Opus 4.8 achieves the highest Video State recall. No single evaluated agent therefore dominates all stages of the task.

More broadly, the results reveal substantial variation between answer quality and evidence recovery across systems. Models with comparable final-answer performance can exhibit markedly different video- and web-grounding behavior, suggesting that answer accuracy alone does not characterize how successfully an agent establishes the cross-modal evidence dependency required by VDR. We analyze this gap further in Section 4.4.

4.3 Qualitative Analysis

We further inspect incorrect predictions from the strongest evaluated system to understand why locally plausible research trajectories still fail. Rather than treating failures as isolated perception, retrieval, or reasoning errors, we observe three recurring mechanisms in which an initially plausible decision changes which downstream evidence becomes relevant.

Correct Research on the Wrong Video-Grounded Branch. A recurring failure begins on the video side. The agent selects the wrong entity, event, or relation from the video, but then conducts coherent web research based on that choice. The retrieved sources may be valid and the subsequent reasoning internally consistent, yet they support the wrong candidate rather than the one implied by the required video evidence. As a result, better downstream retrieval may reinforce an incorrect research path instead of correcting it.

Incorrect Video-to-Web First-Hop Mapping. A second failure occurs even when the relevant video content is identified correctly. The error appears when the agent first connects that video evidence to an external entity, event, or record. A surface-level match—such as a shared name, visual concept, or related event—can lead the search to the wrong web target, after which the remaining research may again be internally consistent. For example, an agent may correctly identify the target object in the video but associate it with the wrong artist or work when beginning web search. These cases show that correctly understanding the video is not sufficient; the first video-to-web link must also be established correctly.

Selecting the Wrong Fact from Relevant Web Evidence. We also observe cases in which the agent reaches the correct entity and even the correct webpage, but extracts the wrong fact from the available evidence. This can happen when a page contains several closely related values or records, such as different dates, editions, events, or roles, and the agent does not use the full question context to distinguish between them. The failure therefore occurs after successful retrieval: the relevant source is available, but the selected fact does not satisfy the specific constraint posed by the question.

4.4 Discussion

Insight 1: Partial recovery does not readily translate into complete grounding. Table 3 shows that current agents frequently recover some of the required evidence without establishing the full evidence set. This pattern appears on both the video and web sides and is particularly pronounced for web evidence. The main difficulty is therefore not simply finding isolated relevant evidence, but satisfying all constraints required by a question within a coherent research trajectory.

Insight 2: Not all correct-answer paths are equally grounded. Answer correctness and evidence completeness do not always fail or succeed together. Our manual inspection of answer-correct but incompletely grounded cases reveals several distinct patterns. In some cases, the agent appears to reach the correct answer through a reasoning path that differs from the human-annotated one, while still establishing the same underlying semantic dependency. The intermediate representations, ordering of steps, or source choices may differ, yet the reasoning can converge on the same valid conclusion. This does not imply that the required evidence is redundant; rather, human annotators and model agents may realize the same minimal dependency through different reasoning paths. In other cases, however, the answer is obtained by bypassing a required modality or by directly reaching the final entity without establishing the necessary intermediate relations. These cases should not be treated equivalently: the former reflects a valid alternative realization of the same evidence dependency, whereas the latter remains incomplete grounding under the task definition.

Table 3: **Partial vs. Complete Evidence Grounding.** For each question, video states and Web claims are classified as **Complete** when all required units are supported, **Partial** when at least one but not all required units are supported, and **None** when no required unit is supported. All values are percentages of the 100 benchmark questions.

	Video Evidence			Web Evidence		
System	**Complete**	**Partial**	**None**	**Complete**	**Partial**	**None**
GPT-5.6 Sol-High	21	36	43	8	53	39
Claude Opus 4.8	24	39	37	4	34	62
Claude Sonnet 5	15	34	51	1	42	57
GPT-5.6 Terra	7	28	65	1	27	72
Grok-4.6	8	25	67	3	28	69
Composer 2.5	7	27	66	1	23	76
Qwen3.8-Plus	20	30	50	1	43	56
Qwen3.8-Max	**26**	31	43	2	40	58

Insight 3: Cross-stage constraint maintenance is a central bottleneck. The qualitative failures share a common structure: individual perception, retrieval, and reasoning steps can each appear locally plausible while the overall research trajectory is wrong. A video-side binding must continue to constrain which web branch is valid, and question-side semantics must continue to constrain which retrieved fact is selected. This suggests that progress in VDR requires stronger mechanisms for maintaining, validating, and revising constraints across stages, rather than improving individual stages only in isolation. Future VDR agents may therefore benefit from explicitly tracking intermediate entities and constraints, checking whether each new retrieval remains consistent with earlier video evidence, and revisiting earlier decisions when later evidence introduces a conflict.

Insight 4: Grounding evaluation should permit multiple valid research trajectories. A process-level benchmark should evaluate whether the required semantic information has been established, rather than whether a model reproduces an annotator's exact search sequence. GROUNDEDVDR therefore defines video states and web claims at the semantic level, independently of a particular URL, wording, or search order. This allows alternative research trajectories to receive credit when they establish the same necessary video-to-web dependency.

Insight 5: Strict evidence evaluation remains challenging. Evaluating research processes is inherently more ambiguous than evaluating final answers. Models may omit evidence they actually used, express an intermediate fact at a different semantic granularity, or support the same claim through a different valid source. Our state- and claim-level semantic evaluation accommodates variation in annotation wording, evidence granularity, and valid retrieval paths, providing a more robust basis for process-level evaluation than exact-match evidence checking.

5 CONCLUSION

We introduce GROUNDEDVDR, an evidence-grounded benchmark for Video Deep Research that explicitly connects temporally distributed video evidence with external web evidence through structured video states, source-independent web claims, and shortcut-resistant construction. Evaluations of eight multi-modal agents show that strong final-answer performance does not necessarily translate into complete evidence grounding: current systems often recover useful evidence fragments while failing to maintain the full set of constraints across video understanding, web retrieval, and evidence selection. Qualitative analysis further reveals that locally plausible reasoning can still follow an incorrect research trajectory, highlighting cross-stage constraint maintenance as a key challenge for future VDR systems. By making both outcomes and supporting research processes auditable, GROUNDEDVDR provides a testbed for evaluating this capability beyond answer accuracy alone.

6 AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the manuscript for clarity and readability. Additionally, we used generative AI tools for editing a research paper to improve readability. All AI-assisted content was critically reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. Claude models overview. https://www.anthropic.com/news/claude-opus-4-8, 2026. Accessed 2026-09-26.

Anysphere. Models and pricing. https://x.ai/news/composer-2-5, 2026. Accessed 2026-09-26.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of MLLM agents. *arXiv preprint arXiv:2503.10200*, 2025. URL https://arxiv.org/abs/2503.10200.

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. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Zhen Fang, Yu Zeng, Wenxuan Huang, Yiming Zhao, Shiting Huang, Tianfei Ren, Qi Lu, Qingnan Ren, Qisheng Su, Lionel Z. Wang, Qingyu Yin, Shuang Chen, Zehui Chen, Lin Chen, Zhenfei Yin, Yao Hu, Shaosheng Lin, Wanli Ouyang, Shaosheng Cao, and Feng Zhao. Videodeepresearch: Towards the next-generation multimodal deepresearch agent. *arXiv preprint arXiv:2608.03979*, 2026. URL https://arxiv.org/abs/2608.03979.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023. URL https://arxiv.org/abs/2305.14627.

Zhenkun Gao, Yicheng Bao, Jinlong Peng, Xueheng Li, Theo Huang, Bangwei Liu, Kunquan Li, Zhenye Gan, Tao Hu, Chengjun Xie, Mingqian Yang, Xuanhua He, Zhizhong Zhang, Xin Tan, Chengjie Wang, and Yuan Xie. VideoSearcher: Empowering video deep research with multi-tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026. URL https://arxiv.org/abs/2607.02927.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebWatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025. URL https://arxiv.org/abs/2508.05748.

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

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.

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

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 arXiv:2512.23044*, 2025. URL https://arxiv.org/abs/2512.23044.

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*, 2026a. URL https://arxiv.org/abs/2601.06943.

Wenqi Liu, Shijie Ma, Yunxiao Wang, Meng Liu, Qile Su, Han Liu, Bohan Hou, Xuanyu Zheng, Changyi Liu, Tianke Zhang, Haonan Fan, Kaiyu Jiang, Yingxin Li, Jiankang Chen, Xu Wang, Bin Wen, Tingting Gao, Han Li, Jianhua Yin, Yinwei Wei, and Xuemeng Song. Thinking beyond videos: Unifying video reasoning and deep research for open-world video agents. *arXiv preprint arXiv:2608.23329*, 2026b. URL https://arxiv.org/abs/2608.23329.

OpenAI. Models: Openai api documentation. https://openai.com/index/gpt-5-6/, 2026. Accessed 2026-09-26.

Qoder. Qwen3.8-Flash: Free for a limited time. https://qwen.ai/blog?id=qwen3.8-flash-next, 2026. Accessed 2026-09-26.

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

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

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.

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

xAI. Models and pricing. https://x.ai/news/grok-4-6, 2026. Accessed 2026-09-26.

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

Zuhao Yang, Kaichen Zhang, Sudong Wang, Keming Wu, Zhongyu Yang, Bo Li, Xiaojuan Qi, Shijian Lu, Xingxuan Li, and Lidong Bing. Paravt: Taming the tool prior paradox for parallel tool use in agentic video reinforcement learning. *arXiv preprint arXiv:2605.20342*, 2026.

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. URL https://arxiv.org/abs/2603.14468.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Xu Tang, Yao Hu, Philip Torr, Wanli Ouyang, and Shaosheng Cao. Vision-DeepResearch Benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026. URL https://arxiv.org/abs/2602.02185.

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. *arXiv preprint arXiv:2505.18079*, 2025.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

A Annotation Object

The following compact schema illustrates the intended gold object. It omits implementation-only metadata.

```
{
  "video": {"video_id": "...", "domain": "..."},
  "task_question": "...",
  "final_answer_gt": {"answer_text": "...", "answer_type": "entity"},
  "state_pool_gt": [{
    "state_id": "s1",
    "sub_question": "...",
    "sub_answer": "observable video fact",
    "key_evidence_units": [{
      "canonical_interval": [12.4, 16.8],
      "description": "observable fact only",
      "linked_observation_ids": ["obs_07"]
    }]
  }],
  "candidate_set": [{
    "candidate_id": "c1", "is_correct": true,
    "linked_state_ids": ["s1", "s2"]
  }],
  "web_evidence_gt": [{
    "url": "...", "claim_text": "...",
    "support_target": "candidate_disambiguation"
  }],
  "hard_negative_web": [{
    "linked_candidate_id": "c2",
    "plausibility_reason": "...", "exclusion_reason": "..."
  }]
}
```

B Submission Checklist for the Final Study

Before converting this draft into a submission, we require:

- a frozen accepted set with train/development/test separation and immutable manifests;
- independent double review, adjudication, agreement, compensation, and time reporting;
- source-known, single-frame, OCR/title, state-deletion, web-only, and video-only tests;
- verified snapshots or fetch-time entailment for every required web claim;
- group-aware strict temporal scoring and rejection of contradictory answer strings;
- at least two model families, natural and controlled budgets, repeated runs, confidence intervals, and paired tests;
- complete provenance, licensing, privacy, web-drift, and takedown documentation;
- removal or resolution of every **[TBD]** marker in the main paper.

C CONSTRUCTION DIAGNOSTICS

We report two development-stage diagnostics to illustrate the filtering effect of our construction protocol. These statistics refer to early development pools and are not part of the final benchmark results.

Table 4: **Development-stage construction diagnostics.** The blueprint pool and retrospective audit precede the frozen benchmark release and are reported only to characterize the construction process.

Diagnostic	**Result**
Candidate video packages	64
Provisional question blueprints	77
Videos yielding no viable item	23/64 (35.9%)
Earlier questions retrospectively audited	29
Passed temporal necessity	23/29
Passed retrieval necessity	22/29
Passed both necessity criteria	22/29

The early blueprint pool shows that a substantial fraction of candidate videos did not support a natural and verifiable VDR question under our construction criteria. Separately, the retrospective audit found that some structurally valid earlier questions still failed temporal or retrieval necessity. These observations motivated the explicit necessity checks used in the final construction pipeline.

D EVALUATION PROMPT

To standardize the reasoning steps and output formats of the evaluated agents on Benchmark, we employ a unified system prompt. The exact prompt provided to the models is detailed below:

```
You are solving a Video Deep Research benchmark. Each item provides a video and
a question. Use the video and any available external research tools to answer the
question.

You may choose your own research and reasoning strategy, including the order in
which you inspect the video and search external sources.

## Data access

From each question annotation, read only:
- `sample_id`
- `task_question`
- `video.video_id`
- `video.source_uri`

Do not read or use ground-truth or annotation-only fields such as
`final_answer_gt`, `candidate_set`, `state_pool_gt`, `web_evidence_gt`,
`hard_negative_web`, or `reasoning_checklist_gt`.

## Required output

Return:
1. the final answer;
2. the video evidence used to support it;
3. the web evidence used to support it;
4. a brief reasoning summary;
5. confidence and any limitations.

Return every answer part requested by the question in the same order.
```

```
In `final_answer.answer_text`, include only the requested final answer parts.
Separate multiple answer parts with literal semicolons (;).
If a requested part cannot be determined, write `UNKNOWN` in that position.

## Video evidence

Organize video evidence as states containing one or more evidence units.

Each evidence unit must:
- refer to one continuous video interval;
- use valid start and end times in seconds;
- describe only what is directly visible or audible in that interval;
- use one modality: `visual`, `audio`, or `audiovisual`.

On-screen text counts as `visual`.

Do not invent timestamps or describe information that is not observable in the
specified interval.

## Web evidence

For every web source actually used, provide:
- the opened URL;
- page title when available;
- a verbatim quote from the page;
- a concise `support_claim` describing what that quote supports.

Quotes must be copied from the opened page. Do not invent URLs, titles, quotes,
or source content.

If multiple discontinuous quote segments are used, separate them with `;`.
A single verbatim segment itself must not contain `;`.

## Output format

Write one JSON object per line to `vdr_predictions.jsonl`:

{
  "sample_id": "<sample id>",
  "final_answer": {
    "answer_text": "<requested final answer only>"
  },
  "used_video_evidence": [
    {
      "state_id": "s1",
      "sub_question": "<video-side question>",
      "sub_answer": "<video-side conclusion>",
      "evidence_units": [
        {
          "evidence_id": "s1_e1",
          "start_sec": 12.4,
          "end_sec": 15.1,
          "description": "<local observable fact>",
          "modality": "visual"
        }
      ]
    }
  ],
  "used_web_evidence": [
    {
      "evidence_id": "w1",
      "url": "<opened URL>",
      "title": "<page title>",
```

```
      "quote": "<verbatim quote>",
      "support_claim": "<fact supported by this quote>"
    }
  ],
  "reasoning": "<brief reasoning summary>",
  "confidence": 0.0,
  "limitations": "<uncertainties or access limitations>"
}
```