UNIVERSE: A UNIFIED BENCHMARK FOR VIDEO DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Video deep research requires agents to coordinate video understanding with external information seeking, yet existing benchmarks cover only fragmented aspects of this capability and remain difficult to scale. Some provide a video and require research grounded in its content, while others require agents to discover relevant videos on the open web. Moreover, existing benchmarks largely emphasize using visual observations to initiate research, leaving the reverse process of using external knowledge to determine what should be observed underexplored. We introduce UNIVERSE, a unified benchmark that jointly evaluates supplied-video research and open-web video discovery. Beyond observation-to-research, UNIVERSE introduces knowledge-guided observation, where agents first resolve a target from indirect external clues and then locate it in video to obtain the requested visual evidence, together with image-based query clues that couple visual references with relational research. To enable scalable construction, we combine automated instance generation with multi-stage verification and shortcut filtering, yielding 800 high-quality questions across ten domains, substantially exceeding the scale of prior video deep research benchmarks. Evaluations of representative proprietary and open-weight multimodal models under a shared tool-enabled framework reveal substantial challenges in coordinating external research, video discovery, and visual observation, highlighting considerable room for progress in video deep research.

1 INTRODUCTION

Deep research tasks require agents to answer complex questions by iteratively searching for information, reasoning over retrieved evidence, and synthesizing findings across multiple sources (Wei et al., 2025; Wu et al., 2025a; Du et al., 2025). Recent work has extended this paradigm to multimodal settings, where both user queries and relevant web evidence may contain images and other visual content (Li et al., 2025; Geng et al., 2025). Multimodal deep research therefore couples visual understanding with external information seeking, enabling agents to use visual cues to guide search and integrate visual and textual evidence when answering complex questions (Wu et al., 2025b; Huang et al., 2026b).

More recently, research has begun to explore *video deep research*, extending multimodal information seeking from images to videos and requiring agents to coordinate video observation with external research. Existing benchmarks broadly fall into two paradigms. VideoDR, VideoSearch-QA, and Video-DR-Bench begin with a supplied video, where agents identify entities or events from visual clues and search the web for additional information needed to answer the question (Liu et al., 2026; Gao et al., 2026; Fang et al., 2026). By contrast, Video-BrowseComp begins with a research question and requires agents to discover relevant videos on the open web before inspecting their contents for answer evidence (Liang et al., 2025). Despite this progress, several important capabilities remain insufficiently covered. **Supplied-video research and open-web video discovery remain separated.** Evaluating either setting in isolation provides only a partial view of video research capability: in realistic scenarios, an agent should be able both to investigate video content supplied by a user and to actively acquire relevant video evidence when none is given. **Knowledge-guided observation remains underexplored.** Existing supplied-video tasks largely follow an *observation-to-research* pattern, where visual content provides clues about a target and external search subsequently identifies the target or retrieves related facts. Much less attention has been paid to the reverse direction,

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.

where external knowledge must first determine *what in the video should be examined*. This leaves a key capability inadequately tested: resolving a relevant person, object, or scene from external information and then locating that target in the video to obtain the required visual evidence. **Image-based query clues receive limited coverage.** Existing benchmarks primarily focus on video and text, although an image may itself indirectly specify a research target through a related entity. Without such queries, agents' ability to connect image interpretation, external knowledge, and video evidence remains insufficiently assessed. Beyond these gaps in capability coverage, **benchmark construction itself remains difficult to scale**. Each instance must consistently connect a real-world research target, external factual evidence, and specific video evidence, while ensuring that the answer cannot be recovered by bypassing the intended research or observation process. Existing video deep research benchmarks therefore rely substantially on manual annotation or human–AI collaborative curation, making large-scale construction costly while typically limiting benchmarks to only a few hundred questions (Liu et al., 2026; Gao et al., 2026; Fang et al., 2026; Liang et al., 2025).

To address these limitations, we introduce UNIVERSE, a benchmark for **UNI**fied **V**id**E**o deep **R**e**SE**arch. UNIVERSE unifies supplied-video research and open-web video discovery through three complementary task configurations. In *observation-to-research* (O2R), a supplied video provides the visual evidence needed to resolve a research anchor before external information is retrieved to answer the question. In *knowledge-guided observation* (KGO), this dependency is reversed: indirect external clues first determine the relevant anchor, which must then be located in the supplied video to obtain the requested visual evidence. In *open-web video discovery* (OVD), no video is supplied; the agent must resolve the relevant research context, discover suitable footage on the open web, and inspect it for the answer. Across KGO and OVD, we further introduce *image-based query clues*, in which an image indirectly specifies the research anchor through relational knowledge, requiring image interpretation to participate directly in target resolution and subsequent video research.

To make such evaluation scalable, we develop a structured construction pipeline that automates the core stages of instance generation by coupling external relational knowledge with grounded video evidence. Generated candidates undergo multi-stage evidence verification, task-specific shortcut filtering, and final human review, reducing dependence on manual question authoring while preserving factual consistency, evidence sufficiency, and the intended research dependencies. This pipeline enables us to construct 800 high-quality questions over 800 distinct research anchors and 782 distinct videos across ten content domains, substantially exceeding the scale of the existing video deep research benchmarks discussed above. The benchmark includes 300 questions with supplied videos and 500 requiring open-web video discovery, with 161 questions incorporating image-based query clues. We evaluate eight representative proprietary and open-weight multimodal models under a shared tool-enabled framework and analyze their performance across research requirements, tool-access conditions, and input modalities. The results reveal substantial remaining challenges in coordinating external information seeking, target resolution, video discovery, and visual observation.

Our main contributions are as follows:

- We formulate a unified setting for video deep research that jointly evaluates supplied-video research and open-web video discovery, while extending evaluation to knowledge-guided observation and image-based query clues.
- We develop a scalable benchmark-construction pipeline that couples external relational knowledge with grounded video evidence and combines automated construction with multi-stage verification and targeted shortcut filtering. This enables UNIVERSE to comprise 800 high-quality questions across ten content domains, substantially exceeding the scale of prior video deep research benchmarks.
- We evaluate representative proprietary and open-weight multimodal models under a shared tool-enabled framework and systematically analyze their performance across research requirements, tool-access conditions, and input modalities, characterizing the strengths and limitations of current video deep research agents.

2 Related Work

Deep Research Agents. Early foundations of deep research agents connect language-model reasoning with browsing, source collection, and structured synthesis to answer questions using external

information (Nakano et al., 2021; Yao et al., 2022; Shao et al., 2024). Subsequent approaches train autonomous information-seeking policies using supervised or reinforcement learning (Jin et al., 2025; Zheng et al., 2025; Wu et al., 2025a). Evaluations cover complementary objectives, from locating hard-to-find facts with short, verifiable answers to assessing research-report quality and citation support (Wei et al., 2025; Du et al., 2025). Multimodal research extends these capabilities to visual queries and image–text evidence, enabling visually grounded entity identification and repeated searches across sources (Jiang et al., 2024; Wu et al., 2025b; Geng et al., 2025; Huang et al., 2026b). Corresponding benchmarks test the extraction and integration of visual and textual evidence, as well as citation-grounded report generation (Li et al., 2025; Tao et al., 2025; Huang et al., 2026a). VideoDR, VideoSearch-QA, and Video-DR-Bench investigate web research grounded in supplied videos, linking cross-frame visual clues and entity identification to knowledge beyond the recording (Liu et al., 2026; Gao et al., 2026; Fang et al., 2026). Video-BrowseComp evaluates the complementary setting of open-web video discovery, with tasks spanning explicit and implicit retrieval as well as reasoning across multiple video sources (Liang et al., 2025). Across these settings, videos play complementary roles as clues for acquiring external knowledge and as sources of observable answer evidence, motivating evaluation of both roles within a common research framework. UNIVERSE unifies these settings and systematically evaluates how external relations and image-based query clues determine what agents should observe in supplied or discoverable videos.

Video Understanding. Video-language models develop video understanding through visual alignment and instruction tuning (Maaz et al., 2023; Lin et al., 2023; Zhang et al., 2024). Efficient long-video representations and reinforcement learning further address context costs and video reasoning capabilities (Shen et al., 2024; Feng et al., 2025; Park et al., 2025). Benchmarks assess complementary capabilities in temporal reasoning (Xiao et al., 2021; Li et al., 2023; Liu et al., 2024), broad multimodal understanding (Fu et al., 2024; 2026), and long-context comprehension (Wu et al., 2024; Zhou et al., 2024). Knowledge-intensive evaluations further require models to apply expert knowledge to specialized video content or transfer knowledge learned from videos to new problems (Zhao et al., 2025; Hu et al., 2025). Agentic methods complement these advances through query-adaptive frame selection, hierarchical representations, and iterative temporal grounding and inspection, dynamically choosing where and at what granularity to examine the video (Wang et al., 2024a;b; Yuan et al., 2025; Li et al., 2026). These studies establish capabilities for temporal reasoning, knowledge use, and selective observation within supplied recordings. UNIVERSE additionally evaluates external information seeking to resolve research targets, determine the observations needed for an answer, and discover relevant video evidence on the open web.

3 UNIVERSE

3.1 Data Overview

UNIVERSE evaluates whether agents can coordinate external information seeking with video observation to answer research questions. Each instance consists of an input $x = (q, I, v_{\text{in}})$ and a reference answer a, where q is a textual question, I contains optional image-based query clues, and v_{in} is either a supplied video or $\varnothing$ when relevant footage must be discovered from the open web.

We consider three task configurations under two video-access settings. In *observation-to-research* (O2R), visual evidence from a supplied video identifies a relevant real-world entity from which external research proceeds. In *knowledge-guided observation* (KGO), external information first resolves the relevant entity, after which the requested visual evidence is obtained from the supplied video. In *open-web video discovery* (OVD), no video is supplied; the agent must resolve the research context, discover suitable footage, and inspect it for the answer.

3.2 Question Construction

Our construction couples three components: a *research anchor* e, external knowledge associated with e, and a supporting video $v^\star$. The anchor is the central real-world referent connecting the external and visual sides of a question; external knowledge provides factual attributes and relations involving e, while $v^\star$ provides the required visual evidence.

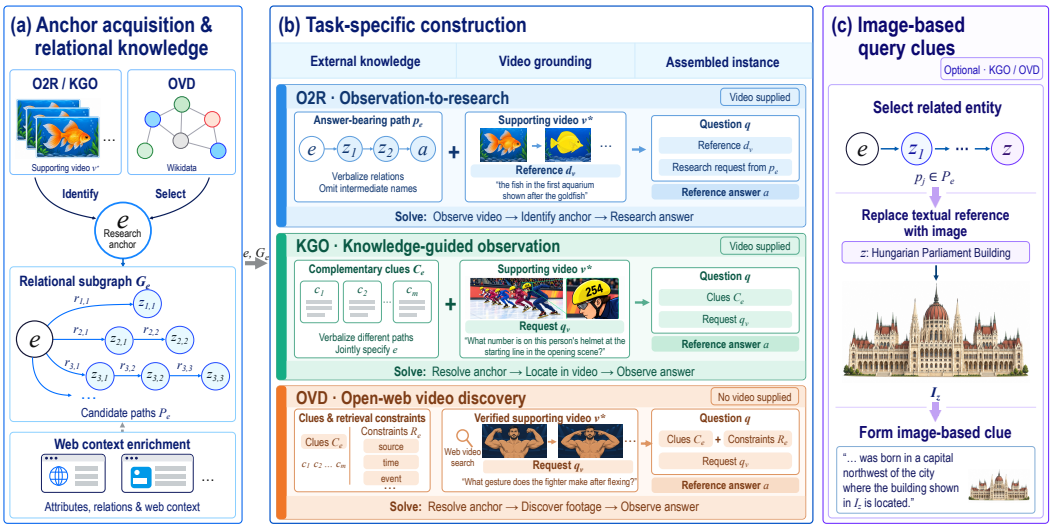


Figure 1: Overview of UNIVERSE question construction: (a) research-anchor acquisition and relational-knowledge construction, (b) task-specific construction for O2R, KGO, and OVD, and (c) optional image-based query clues for KGO and OVD.

3.2.1 Anchor Acquisition and Relational Knowledge Construction

For the supplied-video configurations, we identify a real-world entity appearing in $v^\star$ and resolve it to a Wikidata item; for OVD, we select e directly from Wikidata. In both cases, e serves as the root for collecting anchor-related external knowledge.

Starting from e, we construct a relational subgraph

$$G_e = (V_e, E_e),$$

where $e \in V_e$, and each edge $(u, r, z) \in E_e$ represents a retained Wikidata relation r from u to an entity or attribute value z. We collect usable attributes and relations, selectively expand entity-valued neighbors to introduce multi-hop structure, and augment the resulting subgraph with facts and descriptions from web sources.

3.2.2 Task-specific Construction

Observation-to-research. For an anchor e appearing in $v^\star$, we construct a video-dependent reference d_v from an action, appearance, or scene involving e, without explicitly naming it, such that resolving e requires inspecting the video.

We define candidate relational paths rooted at e as

$$P_e = \left\{ \left(e \xrightarrow{r_1} z_1 \xrightarrow{r_2} \cdots \xrightarrow{r_\ell} z_\ell \right) \;\middle|\; (z_{i-1}, r_i, z_i) \in E_e,\ z_0 = e \right\},$$

and select an answer-bearing path

$$p_e = \left(e \xrightarrow{r_1} z_1 \xrightarrow{r_2} \cdots \xrightarrow{r_L} a \right) \in P_e,$$

whose endpoint a determines the reference answer. We verbalize this path as a research question, referring to intermediate entities through their relations rather than explicitly naming them to preserve the intended research steps.

The resulting information dependency is

$$v^\star \xrightarrow[d_v]{\text{video observation}} e \xrightarrow[p_e]{\text{external research}} a.$$

Knowledge-guided observation. Knowledge-guided observation reverses this information flow. We sample multiple paths from P_e,

$$p_j = \left(e \xrightarrow{r_{j,1}} z_{j,1} \xrightarrow{r_{j,2}} \cdots \xrightarrow{r_{j,\ell_j}} z_{j,\ell_j}\right) \in P_e, \qquad j = 1, \ldots, m,$$

and verbalize them into indirect clues $C_e = \{c_1, \ldots, c_m\}$ that jointly specify e without naming it. Each path may be one- or multi-hop. We select complementary paths from different relational branches so that individual clues are insufficient to uniquely identify e, while their combination supports anchor resolution; we favor non-unique attributes, reduce repeated predicates, and use multi-hop relations when appropriate.

We then construct a video-dependent request q_v about an observable property of e in $v^\star$, such as its appearance, action, or surrounding scene. When possible, we select video content containing multiple plausible entities with different values for the queried property, making the answer depend on first resolving the intended anchor.

The resulting information dependency is

$$C_e \xrightarrow{\text{external research}} e \xrightarrow[q_v]{\text{video observation}} a.$$

Open-web video discovery. OVD uses C_e, when present, to specify the research anchor; for instances without anchor-identifying clues, the anchor is directly specified in the question. Unlike KGO, the relevant footage must then be discovered from the open web. We therefore construct anchor-related retrieval constraints R_e, which may specify source, time, topic, metadata, visible content, or other contextual properties.

Let $\mathcal{V}_{\text{web}}$ denote the space of web-accessible videos. We define

$$V_e = \{v \in \mathcal{V}_{\text{web}} \mid v \text{ is associated with } e \text{ and satisfies } R_e\}$$

as the videos matching the intended research context. V_e is used only during construction and is not provided to the evaluated agent; membership does not imply that a video contains sufficient evidence for the final answer.

During construction, we identify a supporting video $v^\star \in V_e$ containing the required visual evidence and use it to formulate q_v and the reference answer a. The supporting video $v^\star$ serves as a verified witness rather than a unique retrieval target: at evaluation time, the agent may inspect any relevant web videos and must determine which provide useful evidence.

The resulting information dependency is

$$C_e \xrightarrow{\text{external research}} e \xrightarrow[R_e,\, q_v]{\text{video discovery and observation}} a.$$

3.2.3 Image-based Query Clues

For KGO and OVD, we further construct image-based variants of the anchor clues in C_e. Given a selected path $p_j \in P_e$, we choose a related entity z on the path with a suitable image I_z and replace its explicit textual reference in c_j with I_z, while retaining the relations connecting z to the anchor. The agent must therefore identify the depicted entity and follow the relational clue to resolve e.

To avoid shortcutting the intended research process, we prefer suitable entities farther from e along the path, filter images containing direct identity cues, and remove redundant textual clues that disclose the same entity through the corresponding relational branch.

Following automated construction, all candidate instances undergo multi-stage evidence verification, task-specific shortcut filtering, and final human review, as detailed in Appendix B.

3.3 Dataset Statistics

UNIVERSE contains 800 questions, including 150 O2R, 150 KGO, and 500 OVD instances. Across the benchmark, the questions involve 800 distinct Wikidata research anchors and 782 distinct videos. Image-based clues are included in 161 questions, comprising 93 KGO instances (62.0%) and 68 OVD instances (13.6%).

Table 1: Performance comparison on UNIVERSE. We report Accuracy and Pass@3 for observation-to-research (O2R), knowledge-guided observation (KGO), open-web video discovery (OVD), and the overall benchmark.

Model	O2R		KGO		OVD		Overall	
	Acc.	Pass@3	Acc.	Pass@3	Acc.	Pass@3	Acc.	Pass@3
w/o Tools								
Gemini-2.5-Pro	58.44	68.67	52.44	67.33	15.87	25.80	30.71	41.62
Gemini-3-Pro	57.56	70.00	53.56	68.00	16.80	28.60	31.33	43.75
Gemini-3.6-Flash	66.89	74.00	58.44	68.67	23.00	31.40	37.88	46.38
Kimi-K2.5	48.00	58.67	48.89	61.33	16.47	27.40	28.46	39.62
Kimi-K2.6	37.56	48.67	37.11	46.67	14.00	20.40	22.75	30.62
InternVL3.5-30B-A3B	10.22	14.67	28.44	38.67	9.13	16.20	12.96	20.12
Qwen3-VL-30B-A3B-Thinking	15.78	22.67	31.33	43.33	9.40	17.00	14.71	23.00
Qwen3.5-35B-A3B	20.89	30.00	37.33	55.33	10.40	19.40	17.42	28.12
w/ Tools								
Gemini-2.5-Pro	63.78	74.00	59.33	76.00	16.13	30.60	33.17	47.25
Gemini-3-Pro	64.00	76.67	60.67	77.33	23.00	43.40	37.75	56.00
Gemini-3.6-Flash	77.56	81.33	65.78	79.33	47.33	63.20	56.46	69.62
Kimi-K2.5	71.11	81.33	54.67	68.67	35.67	53.80	45.88	61.75
Kimi-K2.6	68.89	77.33	51.33	64.00	30.53	48.80	41.62	57.00
InternVL3.5-30B-A3B	9.33	14.00	27.56	38.67	10.53	19.40	13.50	22.00
Qwen3-VL-30B-A3B-Thinking	28.44	42.00	36.00	48.67	10.27	21.60	18.50	30.50
Qwen3.5-35B-A3B	47.33	67.33	42.22	62.00	20.93	37.60	29.88	47.75

The benchmark spans ten content domains, as shown in Figure 5. Sports and film/screen entertainment are the two largest categories, accounting for 37.9% and 24.8% of the questions, respectively, while the remaining questions cover a broad range of topics including music, lifestyle, public affairs, history, technology, gaming, science, and business. Research-anchor distributions also differ across task configurations. O2R covers a relatively diverse set of anchor types, whereas KGO is primarily centered on people (68.0%) and OVD on events or event series (76.0%).

Table 2 further summarizes the characteristics of individual instances. Videos have an overall mean duration of 8.88 minutes, with task-level means of 7.60, 8.69, and 9.32 minutes for O2R, KGO, and OVD, respectively. For OVD, these durations refer to the verified supporting videos used during construction rather than videos supplied to the agent at evaluation time. Question length varies substantially across configurations, averaging 45.2 words for O2R, 144.6 for KGO, and 86.8 for OVD, while reference answers remain concise, with a median length of two words and 88.4% containing at most five words. Additional statistics on relational structure and video evidence are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

We evaluate eight representative multimodal models: Gemini-2.5-Pro (Comanici et al., 2025), Gemini-3-Pro (Google DeepMind, 2026b), Gemini-3.6-Flash (Google DeepMind, 2026a), Kimi-K2.5 (Kimi Team et al., 2026), Kimi-K2.6 (Moonshot AI, 2026), Qwen3.5-35B-A3B (Qwen Team, 2026), Qwen3-VL-30B-A3B-Thinking (Bai et al., 2025), and InternVL3.5-30B-A3B (Wang et al., 2025). Each model is evaluated on all 800 questions under both *tool-enabled* and *no-tool* settings. The tool-enabled setting follows the shared research-agent framework in Appendix C, using the seven tools in Table 3, with at most one tool call per turn and 50 model turns per trajectory; VIDEO-SUMMARY uses Gemini-3-Pro as a fixed backend across models. The no-tool setting retains the original questions and supplied visual inputs while disabling external retrieval and online-video inspection. We independently sample three trajectories per question and use Gemini-3.6-Flash as

the multimodal judge for final-answer correctness. We report **Accuracy** over all trajectories and **Pass@3**, indicating whether at least one of the three trajectories answers a question correctly.

4.2 Main Results

Table 1 presents the main results on UNIVERSE. We make two key observations.

UNIVERSE remains challenging for current multimodal models. With tools enabled, Gemini-3.6-Flash achieves the strongest overall performance, with 56.46 Accuracy and 69.62 Pass@3. Nevertheless, its accuracy remains below 60%, and performance on open-web video discovery reaches only 47.33. The remaining models achieve overall accuracies below 50%, with substantial gaps across model families. These results show that UNIVERSE poses considerable challenges even to strong multimodal models equipped with external research capabilities.

The three task configurations pose distinct research challenges. OVD is consistently the most challenging setting across models. For example, Gemini-3.6-Flash achieves 77.56 Accuracy on O2R and 65.78 on KGO, but only 47.33 on OVD. This pattern is consistent with the additional need in OVD to discover suitable footage and determine which candidate videos contain useful evidence. Among the two supplied-video tasks, the higher-performing tool-enabled models generally achieve higher Accuracy on O2R than on KGO. Unlike O2R, KGO requires the agent to resolve the target from indirect external clues before conducting target-specific video observation, introducing an additional coordination challenge between research and observation.

4.3 Tool Utilization and Effectiveness

Figure 2(a) shows that the benefit of tool access varies substantially across both models and task configurations. On O2R, several models obtain large improvements, including Kimi-K2.6 (+31.3 points), Qwen3.5-35B-A3B (+26.4), and Kimi-K2.5 (+23.1). By contrast, gains on KGO are generally more moderate, with the largest improvement reaching +14.2 points. OVD exhibits particularly heterogeneous tool benefits: Gemini-3.6-Flash, Kimi-K2.5, and Kimi-K2.6 improve by +24.3, +19.2, and +16.5 points, respectively, whereas Gemini-2.5-Pro and Qwen3-VL-30B-A3B-Thinking gain less than one point. InternVL3.5-30B-A3B shows little change across all three tasks. These results indicate that access to the same tool environment does not translate into uniform performance gains across models.

Figures 2(b–d) further examine whether these differences are associated with how actively models engage with the available tools. Across the eight models, mean tool calls per trajectory are positively correlated with Accuracy gains on all three tasks, with Pearson correlations of $r = 0.766$, 0.742, and 0.744 for O2R, KGO, and OVD, respectively. InternVL3.5 makes very few tool calls and obtains little benefit from tool access, whereas Kimi-K2.6 is consistently among the most tool-active models and achieves large gains across all three tasks. The relationship is not one-to-one, however. Qwen3.5 obtains a large gain on O2R with only moderate tool usage, while Kimi-K2.5 makes substantially more calls on KGO but achieves a comparatively small improvement. On OVD, Gemini-2.5-Pro similarly shows nontrivial tool interaction with almost no gain. Together, these results suggest that active tool engagement is associated with greater benefit from tool access, but tool-call frequency alone does not determine effectiveness. We further analyze how models allocate their calls across different tools in Appendix D.1.

4.4 Impact of Image-Based Query Clues

Under the tool-enabled setting, to isolate the effect of image-based query clues, we construct a paired *Text Oracle* variant for each KGO and OVD instance containing an image-based clue. Specifically, we replace the image with the textual identity of the entity that it substitutes during question construction, while keeping the remaining relational clues, video evidence, and requested observation unchanged. This removes the need to resolve the visual reference itself, while preserving the downstream research and video reasoning required by the original question.

Figure 3 shows a consistent performance gap between the original image-clue questions and their text-oracle counterparts. Replacing the image reference with its textual identity improves Accuracy

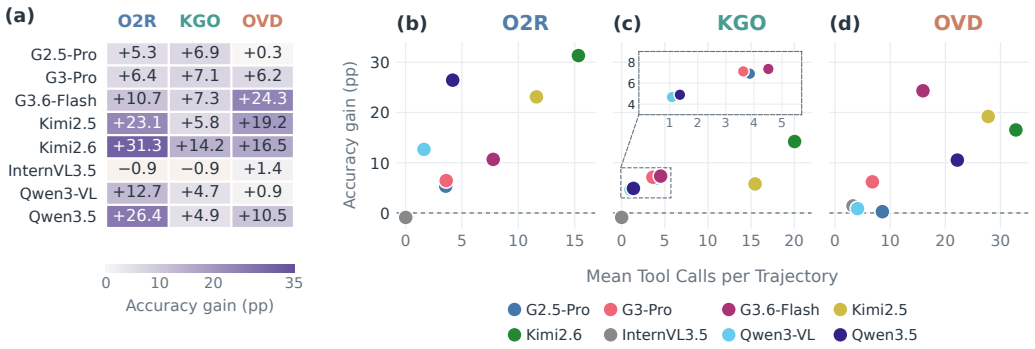


Figure 2: Tool engagement and effectiveness across models and tasks. (a) Accuracy gains from tool access for each model–task pair. (b–d) Relationship between mean tool calls per trajectory and Accuracy gains on O2R, KGO, and OVD. Each point represents one model.

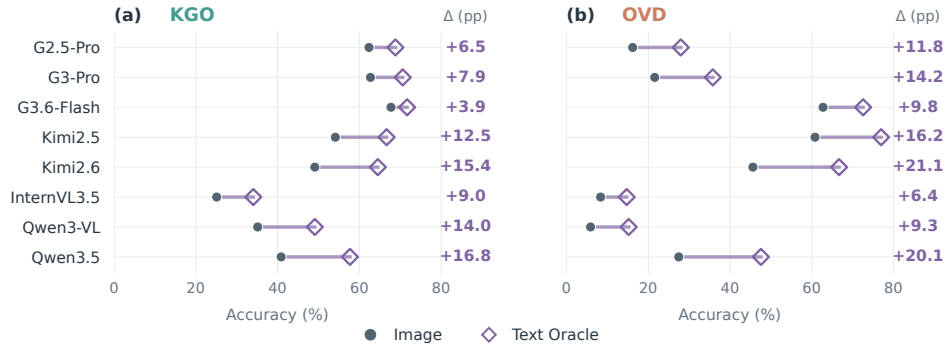


Figure 3: Impact of image-based query clues on KGO and OVD. Circles show Accuracy on the original image-clue questions, while diamonds show Accuracy after replacing the image clue with its underlying textual entity reference (*Text Oracle*). Δ denotes the resulting Accuracy improvement in percentage points.

for every model on both tasks, with average gains of 10.8 percentage points on KGO and 13.6 points on OVD. The magnitude of this effect is substantial for several models. On KGO, Qwen3.5-35B-A3B, Kimi-K2.6, and Qwen3-VL-30B-A3B-Thinking improve by 16.8, 15.4, and 14.0 points, respectively. On OVD, the gains reach 21.1 points for Kimi-K2.6 and 20.1 points for Qwen3.5-35B-A3B.

These results indicate that image-based references introduce a distinct difficulty beyond the subsequent relational reasoning and video processing. The effect is particularly pronounced in OVD, where resolving the visual clue precedes not only anchor identification but also the discovery of relevant video evidence on the open web. An incorrect interpretation of the image can therefore misdirect the subsequent research process before video inspection begins. More broadly, the consistent oracle gains show that image clues are not merely auxiliary multimodal inputs: successfully grounding them into the research context remains an important bottleneck for current multimodal research agents.

4.5 Failure Modes

We further analyze incorrect trajectories in the tool-enabled setting, pooling O2R, KGO, and OVD to characterize the overall failure profile of each model. We use Gemini-3.6-Flash as an LLM annotator to assign each failed trajectory a single primary failure mode based on the full interaction trace and the taxonomy defined below. We consider six categories: *Anchor Resolution*, where the agent fails to resolve the intended research anchor or resolves it incorrectly; *Evidence Retrieval*, where required facts, sources, or video evidence are not successfully acquired; *Visual Understand-*

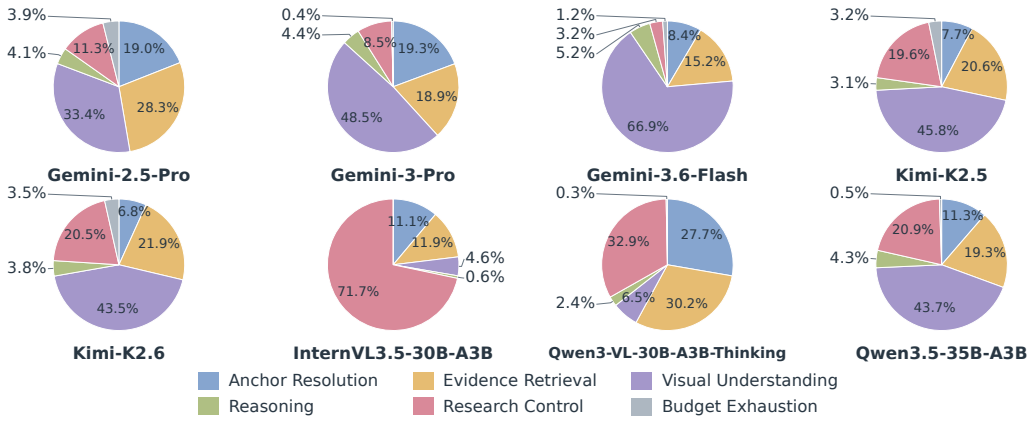


Figure 4: Distribution of primary failure modes across incorrect tool-enabled trajectories, pooling O2R, KGO, and OVD for each model.

ing, where available image or video evidence is incorrectly interpreted; *Reasoning*, where acquired evidence is incorrectly connected, filtered, or synthesized; *Research Control*, covering ineffective tool use, repeated unproductive actions, insufficient verification, or premature termination; and *Budget Exhaustion*, where the interaction limit is reached without a more specific unrecovered failure accounting for the incorrect answer.

Figure 4 reveals substantially different failure profiles across models. Visual Understanding is the largest category for Gemini-3-Pro, Gemini-3.6-Flash, Kimi-K2.5, Kimi-K2.6, and Qwen3.5-35B-A3B, accounting for 43.5–66.9% of their failed trajectories. For these models, interpreting visual evidence constitutes a major bottleneck despite notable differences in their overall performance and tool use. Gemini-2.5-Pro exhibits a more distributed profile, with substantial failures in Visual Understanding (33.4%), Evidence Retrieval (28.3%), and Anchor Resolution (19.0%). By contrast, Qwen3-VL-30B-A3B-Thinking and InternVL3.5-30B-A3B are dominated by failures in the research process itself. For Qwen3-VL, Anchor Resolution (27.7%), Evidence Retrieval (30.2%), and Research Control (32.9%) account for most failures, while Visual Understanding contributes only 6.5%. InternVL3.5 shows an even stronger concentration in Research Control (71.7%), consistent with its limited engagement with the tool environment observed in the preceding analysis. These contrasting profiles indicate that models differ not only in how often they fail, but also in whether their main difficulties lie in resolving and acquiring research evidence or in interpreting the visual evidence involved in the task.

Notably, Reasoning and Budget Exhaustion account for relatively small fractions across models. For Reasoning, this partly reflects our attribution scheme: the category is reserved for cases where the necessary evidence has already been obtained but is incorrectly combined or interpreted, while failures involving missing or incorrect evidence are attributed to the corresponding failure mode. Budget Exhaustion is also uncommon, indicating that the interaction limit itself is rarely the primary factor behind an incorrect trajectory. Taken together, these results reveal qualitatively different bottlenecks across current video deep research agents, spanning anchor resolution, evidence acquisition, research control, and visual understanding.

5 Conclusion

We introduce UNIVERSE, a unified benchmark for video deep research that spans supplied-video and open-web settings through three complementary task configurations: observation-to-research, knowledge-guided observation, and open-web video discovery, with image-based query clues further coupling visual grounding and external research. Evaluations of eight representative multimodal models under a shared tool-enabled framework show that current agents still struggle to coordinate external information seeking with video observation, especially when relevant footage must be discovered from the open web. Further analyses reveal substantial variation in tool utilization, persistent difficulty in grounding image-based clues, and qualitatively different failure patterns across models.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, cleaning and reformatting datasets, and assisting with the interpretation of experimental results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, or support qualitative or thematic data analysis. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. We have reviewed all AI-assisted work. AI-assisted benchmark data were subject to the same multi-stage quality-control procedure described in Section 3.2 and Appendix B, including evidence-based verification, targeted shortcut filtering, and final human review. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

UNIVERSE is constructed from publicly accessible web resources, including Wikidata, web pages, images, and videos, and is intended solely for research on multimodal information-seeking and video understanding. The benchmark focuses on information that is publicly observable or retrievable and does not intentionally introduce private or sensitive personal information. Third-party web content is not redistributed where redistribution rights are not available; instead, we provide references to the original sources where appropriate.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025. doi: 10.48550/arXiv.2507.06261. URL https://arxiv.org/abs/2507.06261.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint arXiv:2506.11763*, 2025. doi: 10.48550/arXiv.2506.11763. URL https://arxiv.org/abs/2506.11763.

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, Shaohui Lin, Wanli Ouyang, Shaosheng Cao, and Feng Zhao. Video-DeepResearch: Towards the Next-Generation Multimodal Deepresearch Agent. *arXiv preprint arXiv:2608.03979*, 2026. doi: 10.48550/arXiv.2608.03979. URL https://arxiv.org/abs/2608.03979.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing Video Reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025. doi: 10.48550/arXiv.2503.21776. URL https://arxiv.org/abs/2503.21776.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The

First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*, 2024. doi: 10.48550/arXiv.2405.21075. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding. *arXiv preprint arXiv:2604.05015*, 2026. doi: 10.48550/arXiv.2604.05015. URL https://arxiv.org/abs/2604.05015.

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. doi: 10.48550/arXiv.2607.02927. 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. doi: 10.48550/arXiv.2508.05748. URL https://arxiv.org/abs/2508.05748.

Google DeepMind. Gemini 3.6 Flash. Official model card, July 2026a. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/. Published July 21, 2026. Accessed 2026-09-22.

Google DeepMind. Gemini 3 Pro: Model Card. Official model card, May 2026b. URL https://deepmind.google/models/model-cards/gemini-3-pro/. May 2026 revision; model released November 2025. Accessed 2026-09-22.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. *arXiv preprint arXiv:2501.13826*, 2025. doi: 10.48550/arXiv.2501.13826. URL https://arxiv.org/abs/2501.13826.

Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, Chaofan Tao, Yan Xu, Dimitrios Dimitriadis, Tuo Zhang, and Mi Zhang. MMDeepResearch-Bench: A Benchmark for Multimodal Deep Research Agents. *arXiv preprint arXiv:2601.12346*, 2026a. doi: 10.48550/arXiv.2601.12346. URL https://arxiv.org/abs/2601.12346.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Xu Tang, Yao Hu, Shaohui Lin, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-DeepResearch: Incentivizing DeepResearch Capability in Multimodal Large Language Models. *arXiv preprint arXiv:2601.22060*, 2026b. doi: 10.48550/arXiv.2601.22060. URL https://arxiv.org/abs/2601.22060.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Benchmarking the Potential of Large Models as Multi-modal Search Engines. *arXiv preprint arXiv:2409.12959*, 2024. doi: 10.48550/arXiv.2409.12959. URL https://arxiv.org/abs/2409.12959.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516*, 2025. doi: 10.48550/arXiv.2503.09516. URL https://arxiv.org/abs/2503.09516.

Kimi Team et al. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*, 2026. doi: 10.48550/arXiv.2602.02276. URL https://arxiv.org/abs/2602.02276.

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. doi: 10.48550/arXiv.2601.15724. URL https://arxiv.org/abs/2601.15724.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *arXiv preprint arXiv:2311.17005*, 2023. doi: 10.48550/arXiv.2311.17005. URL https://arxiv.org/abs/2311.17005.

Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, Chuanhao Li, Jiayi Tian, Chenchen Zhang, Tianhao Peng, Yancheng He, Jihao Gu, Hui Huang, Donghao Zhou, Yuanxing Zhang, Jian Yang, Ge Zhang, Wenhao Huang, Zhaoxiang Zhang, Qiangpeng Yang, and Shilei Wen. MM-BrowseComp: A Comprehensive Benchmark for Multimodal Browsing Agents. *arXiv preprint arXiv:2508.13186*, 2025. doi: 10.48550/arXiv.2508.13186. URL https://arxiv.org/abs/2508.13186.

Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Paolo Rota, Nicu Sebe, Zheng Liu, and Lizi Liao. Video-browsecomp: Benchmarking agentic video research on open web. *arXiv e-prints*, pp. arXiv–2512, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*, 2023. doi: 10.48550/arXiv.2311.10122. URL https://arxiv.org/abs/2311.10122.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Jisheng Dang, Rui Xu, Sen Hu, Jianheng Hou, Chengwei Qin, Xiaobin Hu, Kunyi Wang, Zhi Yang, Hao Peng, Hong Peng, Ronghao Chen, and Huacan Wang. Watching, Reasoning, and Searching: A Video Deep Research Benchmark on Open Web for Agentic Video Reasoning. *arXiv preprint arXiv:2601.06943*, 2026. doi: 10.48550/arXiv.2601.06943. URL https://arxiv.org/abs/2601.06943.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos? *arXiv preprint arXiv:2403.00476*, 2024. doi: 10.48550/arXiv.2403.00476. URL https://arxiv.org/abs/2403.00476.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*, 2023. doi: 10.48550/arXiv.2306.05424. URL https://arxiv.org/abs/2306.05424.

Moonshot AI. Kimi-K2.6. Official model card, Hugging Face, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6. Accessed 2026-09-22.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. doi: 10.48550/arXiv.2112.09332. URL https://arxiv.org/abs/2112.09332.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. DeepVideo-R1: Video Reinforcement Fine-Tuning via Difficulty-aware Regressive GRPO. *arXiv preprint arXiv:2506.07464*, 2025. doi: 10.48550/arXiv.2506.07464. URL https://arxiv.org/abs/2506.07464.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Citation recommended by the official Qwen3.5-35B-A3B model card. Accessed 2026-09-22.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. *arXiv preprint arXiv:2402.14207*, 2024. doi: 10.48550/arXiv.2402.14207. URL https://arxiv.org/abs/2402.14207.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. *arXiv preprint arXiv:2410.17434*, 2024. doi: 10.48550/arXiv.2410.17434. URL https://arxiv.org/abs/2410.17434.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. MMSearch-Plus: Benchmarking Provenance-Aware Search for Multimodal Browsing Agents. *arXiv preprint arXiv:2508.21475*, 2025. doi: 10.48550/arXiv.2508.21475. URL https://arxiv.org/abs/2508.21475.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form Video Understanding with Large Language Model as Agent. *arXiv preprint arXiv:2403.10517*, 2024a. doi: 10.48550/arXiv.2403.10517. URL https://arxiv.org/abs/2403.10517.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos. *arXiv preprint arXiv:2405.19209*, 2024b. doi: 10.48550/arXiv.2405.19209. URL https://arxiv.org/abs/2405.19209.

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. doi: 10.48550/arXiv.2504.12516. URL https://arxiv.org/abs/2504.12516.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. *arXiv preprint arXiv:2407.15754*, 2024. doi: 10.48550/arXiv.2407.15754. URL https://arxiv.org/abs/2407.15754.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebDancer: Towards Autonomous Information Seeking Agency. *arXiv preprint arXiv:2505.22648*, 2025a. doi: 10.48550/arXiv.2505.22648. URL https://arxiv.org/abs/2505.22648.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. MMSearch-R1: Incentivizing LMMs to Search. *arXiv preprint arXiv:2506.20670*, 2025b. doi: 10.48550/arXiv.2506.20670. URL https://arxiv.org/abs/2506.20670.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA:Next Phase of Question-Answering to Explaining Temporal Actions. *arXiv preprint arXiv:2105.08276*, 2021. doi: 10.48550/arXiv.2105.08276. URL https://arxiv.org/abs/2105.08276.

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. doi: 10.48550/arXiv.2210.03629. URL https://arxiv.org/abs/2210.03629.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. VideoExplorer: Think With Videos For Agentic Long-Video Understanding. *arXiv preprint arXiv:2506.10821*, 2025. doi: 10.48550/arXiv.2506.10821. URL https://arxiv.org/abs/2506.10821v6. Version 6; earlier versions titled VideoDeepResearch.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *arXiv preprint arXiv:2410.02713*, 2024. doi: 10.48550/arXiv.2410.02713. URL https://arxiv.org/abs/2410.02713.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding. *arXiv preprint arXiv:2501.12380*, 2025. doi: 10.48550/arXiv.2501.12380. URL https://arxiv.org/abs/2501.12380.

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. doi: 10.48550/arXiv.2504.03160. URL https://arxiv.org/abs/2504.03160.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. *arXiv preprint arXiv:2406.04264*, 2024. doi: 10.48550/arXiv.2406.04264. URL https://arxiv.org/abs/2406.04264.

A Additional Dataset Statistics

Table 2: Dataset statistics by task. Video durations refer to supplied videos for O2R/KGO and verified supporting videos for OVD. Overall video counts are deduplicated across tasks. Lengths are measured in whitespace-delimited words.

Statistic	O2R	KGO	OVD	Overall
Questions	150	150	500	800
Questions with image clues	0	93	68	161
Distinct videos	148	143	500	782
Mean video duration (min)	7.60	8.69	9.32	8.88
Mean question length (words)	45.2	144.6	86.8	89.8
Mean answer length (words)	2.0	2.8	2.9	2.7

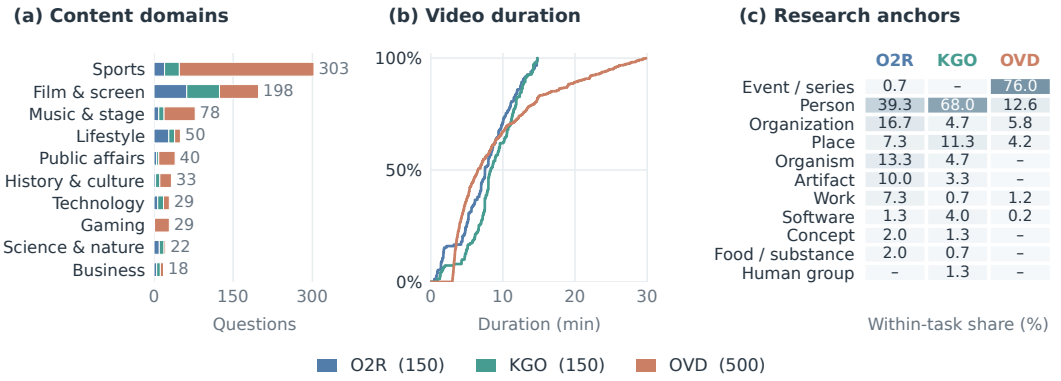


Figure 5: Dataset composition: (a) question counts by content domain and task; (b) empirical cumulative distributions of video duration; (c) research-anchor types as within-task percentages (dashes indicate zero). Category labels are machine-assisted and abbreviated.

A.1 Relational Structure

UNIVERSE exhibits distinct relational structures across its three task configurations. O2R questions use answer-bearing paths p_e with an average length of 2.92 hops and a maximum length of 3 hops. This indicates that the external research targets in O2R are predominantly defined through multi-hop relational chains rather than direct attributes.

KGO and OVD differ more substantially in how their research anchors are specified. As shown in Figure 6(a), KGO questions generally contain more anchor-identifying clues C_e and span a broader range of clue counts, whereas OVD questions are concentrated at smaller values of $m = |C_e|$. Figure 6(b) further shows a difference in relational depth. KGO contains a larger proportion of multi-hop clue paths, while OVD is more heavily concentrated on one-hop relations.

These distributions reflect the different roles of external knowledge across the task configurations. KGO typically requires the target to be resolved from a richer combination of relational clues before video observation, whereas OVD uses a comparatively lighter relational specification before the agent proceeds to open-web video discovery. Together with the predominantly multi-hop answer-bearing paths in O2R, the statistics show that relational composition is present throughout UNIVERSE, but is instantiated differently according to each task's information dependency.

A.2 Video Evidence Types

UNIVERSE covers a diverse range of video evidence types, with distinct distributions across the three task configurations, as shown in Figure 7. O2R is dominated by action and interaction cues

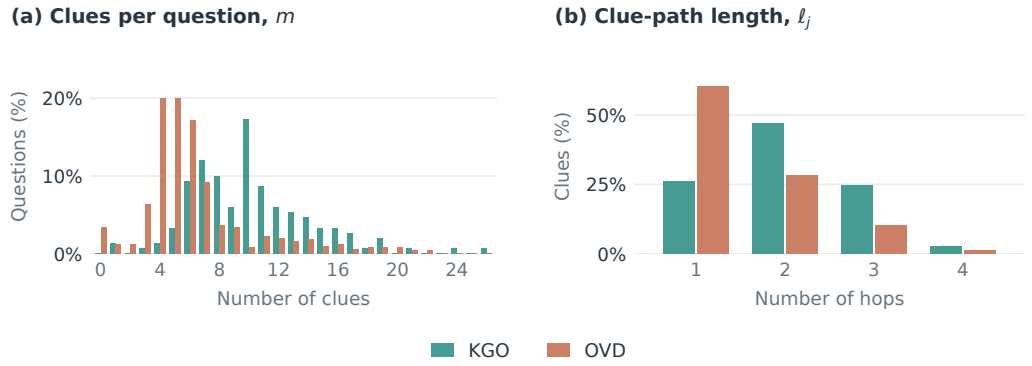


Figure 6: Relational structure of KGO and OVD questions. (a) Number of anchor-identifying clues $m = |C_e|$ per question. (b) Clue-path length ℓ_j, measured in relational hops.

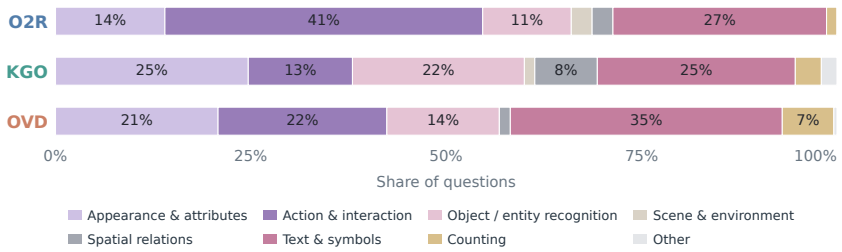


Figure 7: Distribution of primary video evidence types across tasks. For O2R, the category describes the anchor-identifying reference d_v; for KGO and OVD, it describes the requested observation q_v. Percentages are normalized within each task.

(41%), followed by text and symbols (27%). KGO exhibits a more balanced distribution, with appearance and attributes (25%), text and symbols (25%), and object/entity recognition (22%) accounting for similar shares. OVD relies most heavily on text and symbol evidence (35%), while action and interaction (22%) and appearance and attributes (21%) are also common. These distributions show that the three configurations differ not only in their research dependencies, but also in the forms of video evidence that must be interpreted.

B Quality Control

We apply a multi-stage quality-control procedure to ensure that each UNIVERSE instance is well formed, factually supported, and preserves the intended dependency between external research and video observation. The pipeline combines programmatic checks, model-based verification, targeted shortcut filtering, and a final human review.

Structural and construction consistency. We first remove malformed or internally inconsistent instances before performing more expensive verification. The checks cover missing question or answer fields, answer leakage in the question, invalid or unsuitable research anchors, inconsistent references, and violations of task-specific structural requirements. We additionally preserve the construction-side provenance of each instance, including its research anchor, selected relational paths or clues, supporting video evidence, and image-anchor identity when applicable. This provenance allows subsequent verification to operate on the exact evidence used to construct the question rather than reconstructing it from the final surface form.

Evidence-based verification. We then verify the factual and visual components of each instance against their construction-side evidence. For O2R, we check that the video-dependent reference

identifies the intended anchor and that the external relational path from the anchor supports the reference answer. For KGO, we verify the factual consistency of the anchor-identifying clues and confirm that, once the intended anchor is resolved, the supplied video supports the requested observation and answer. For OVD, we verify that the retrieval context is consistent with the supporting video used during construction and that the requested visual evidence is actually observable in that video. When image-based clues are present, we additionally verify the identity of the image anchor and the relational connection through which it specifies the research anchor.

Model-based verification follows a conservative acceptance rule. A verification unit is retained only when it receives repeated confirming judgments; contradictory or inconclusive cases are rejected. This separates evidence verification from the generation process and reduces the likelihood that unsupported relations, ambiguous references, or incorrect visual answers survive into the benchmark.

Targeted shortcut filtering. Correctness alone does not guarantee that a question evaluates the intended research capability. We therefore apply task-dependent probes that deliberately remove information that should be necessary for solving the question and test whether the answer can still be recovered. For supplied-video tasks, these probes test whether the target identity or final answer can be inferred without the intended video or anchor-resolution step. For OVD, we additionally probe whether video-dependent answers can be recovered from non-video web evidence when the question is intended to require observation of the relevant footage. Image-based clues are also screened to reduce cases in which the depicted entity can be trivially identified without research. Candidates for which these probes expose a bypass are filtered out. These checks are intended to remove common shortcut paths rather than to establish that no alternative solution strategy exists.

Human review. Finally, we manually review all retained instances. Reviewers inspect the final question and reference answer together with the construction-side provenance, including the research anchor, relevant relations or clues, supporting video evidence, and image identity when applicable. The review verifies that the question is clear and unambiguous, the reference answer is supported by the available evidence, the external and visual components are correctly connected, and no unintended information in the question directly reveals the target or answer. Instances with unresolved ambiguity, insufficient evidence, factual errors, or apparent shortcuts are corrected and re-checked or removed. Only instances that pass both the preceding verification pipeline and this final human review are included in the benchmark.

C Evaluation Details

We evaluate UNIVERSE with a shared tool-enabled agent framework. Given an instance $x = (q, I, v_{\text{in}})$, the evaluated model acts as a research agent that iteratively gathers evidence, reasons over intermediate observations, and decides when to produce the final answer. The framework provides a common interface for external information retrieval and video inspection, enabling the same agent formulation to cover both supplied-video research and open-web video discovery.

Tool environment. All models are equipped with the same set of tools, summarized in Table 3. Together, they support textual web search, visual retrieval, webpage reading, video discovery, and targeted video inspection. These tools expose complementary information-access primitives rather than task-specific workflows. Search and retrieval tools help the agent identify candidate entities, sources, and videos, while webpage and video inspection tools provide more focused evidence from selected resources. For online videos, the agent determines which footage to inspect and what information to query, while a shared video-understanding backend provides the resulting content observations.

Agentic research process. Evaluation follows an iterative reasoning–action–observation process. At turn t, the agent conditions on the task input and its interaction history,

$$h_t = (x, a_1, o_1, \ldots, a_{t-1}, o_{t-1}),$$

where a_i denotes an action and o_i the corresponding tool observation. Based on h_t, the agent may issue a new tool call or terminate with a final prediction. Newly retrieved textual or visual evidence is incorporated into subsequent reasoning, allowing the agent to refine entity hypotheses, reformulate

Table 3: Tools available to the research agent during evaluation.

Tool	Function
`WebSearch`	Searches the web for relevant pages, entities, facts, and candidate videos.
`ImageSearch`	Retrieves images associated with a textual query.
`ReverseImageSearch`	Retrieves visually related web evidence from a query image.
`ReverseFrameSearch`	Retrieves visually related web evidence from a selected frame of a supplied video.
`LinkSummary`	Extracts question-relevant information from a selected webpage.
`ChannelVideos`	Retrieves videos from an identified channel under metadata constraints.
`VideoSummary`	Answers a targeted question about the content of a selected online video.

searches, inspect candidate sources, and switch between external research and visual observation as needed.

The available video evidence depends on the task setting. In observation-to-research and knowledge-guided observation, the supplied video v_{in} is directly available to the agent throughout the interaction. In open-web video discovery, $v_{\text{in}} = \varnothing$, and video evidence must instead be acquired through web research. The agent may therefore explore and compare multiple candidate videos before identifying footage that provides sufficient evidence for the requested visual detail.

Prediction and evaluation. The evaluated prediction is the final answer $\hat{a}$. An independent multimodal judge evaluates its semantic correctness against the reference answer, using the corresponding video evidence when necessary to disambiguate visual descriptions. The research anchor is an intermediate variable in task construction and reasoning, not a separately scored prediction target; a valid solution need not explicitly submit a canonical anchor or reproduce the construction path.

D Additional Analysis

D.1 Tool Selection Across Tasks

Figure 8 shows that WebSearch dominates tool usage across all three tasks, while the secondary tool composition reflects different research requirements. On O2R, the main secondary tools are ReverseFrameSearch and LinkSummary, whereas KGO shows more frequent use of ReverseImageSearch, consistent with its image-based query clues. OVD exhibits a more diverse tool mix, with substantial use of ImageSearch, LinkSummary, ChannelVideos, and VideoSummary for locating and inspecting candidate videos. Tool preferences also vary across models; for example, Gemini-3.6-Flash allocates a relatively large share of calls to VideoSummary, while Qwen3-VL uses ChannelVideos more heavily. Together with Figure 2(b–d), these patterns show that models differ not only in tool-use intensity, but also in how they allocate calls across different research operations.

D.2 Performance by Video Evidence Type

We further examine tool-enabled model performance across the primary forms of video evidence required by different questions. Using the evidence taxonomy introduced in Appendix A.2, we group questions according to the primary visual information needed either to identify the research anchor in O2R or to answer the video-dependent request in KGO and OVD. Because the frequencies of different evidence types are imbalanced, we focus on recurring patterns across models rather than isolated results from low-frequency categories.

Figure 9 reveals substantial variation across evidence types, with several patterns recurring across models. In O2R, object/entity recognition is generally among the strongest categories for higher-performing models, whereas action-based identification tends to be more challenging. Text and symbol evidence is also handled relatively well by several models, suggesting that visually explicit textual cues can provide effective anchors for subsequent external research. By contrast, weaker models remain limited across nearly all evidence types, indicating that their difficulty cannot be attributed to a single form of visual grounding.

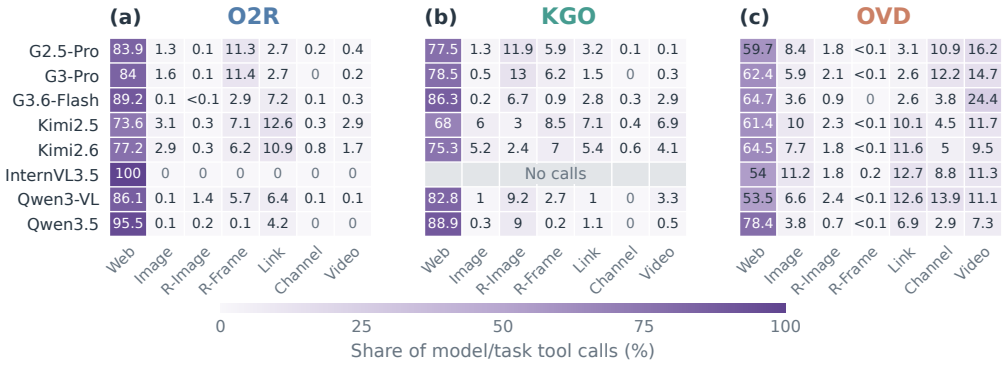


Figure 8: Tool-call composition across models and tasks. Each cell reports the percentage of tool calls assigned to a given tool within each model–task pair. InternVL3.5 makes only five tool calls on O2R and none on KGO. Tool labels abbreviate the names in Table 3. Model names are abbreviated, with G denoting Gemini.

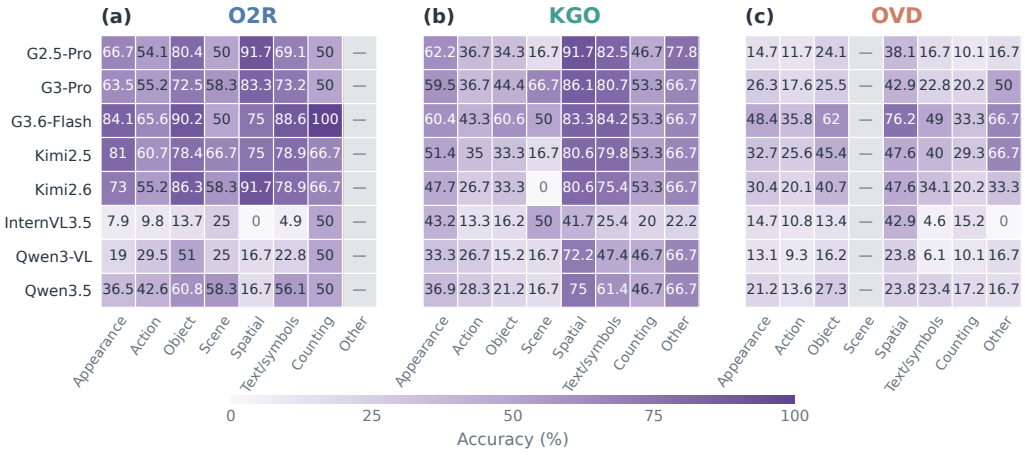


Figure 9: Tool-enabled Accuracy by primary video evidence type for O2R, KGO, and OVD. Each cell reports the Accuracy (%) of a model on questions assigned to the corresponding evidence category. Dashes indicate categories with no questions for the task.

KGO exhibits a different profile. Spatial relations and text/symbol evidence are among the more reliably solved categories for stronger models, while appearance-, action-, and object-based requests generally yield lower performance. This contrast reflects the changed role of video evidence in KGO: the video is no longer used primarily to establish the research anchor, but instead provides the final answer after the anchor has been resolved from external clues. Success therefore requires the agent to translate a resolved target into the specific visual observation requested by the question.

OVD remains difficult across nearly all evidence categories. Spatial evidence is comparatively easier for several stronger models, while action and counting questions remain particularly challenging. More importantly, even categories handled relatively well in the supplied-video settings exhibit substantially lower Accuracy in OVD. This suggests that OVD difficulty is not concentrated in a particular form of video understanding. Rather, interpreting the required video evidence is compounded by the preceding need to discover suitable footage on the open web and determine which retrieved videos provide useful evidence.

Overall, these results show that the three task configurations differ not only in how external research and video access are organized, but also in where models encounter difficulty when interpreting the required visual evidence. The variation across evidence types further shows that aggregate task Accuracy can obscure meaningful differences in the visual operations required by individual questions.

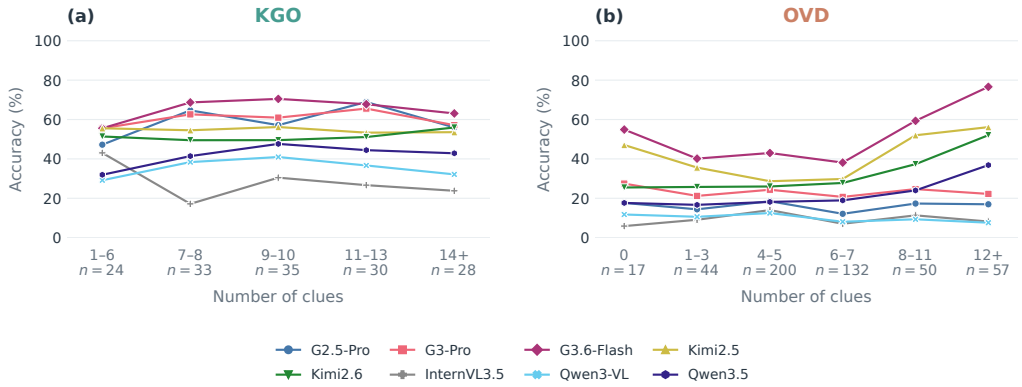


Figure 10: Tool-enabled Accuracy across different numbers of anchor-identifying clues for (a) KGO and (b) OVD. Questions are grouped into the displayed clue-count ranges, with n denoting the number of questions in each group. Each line corresponds to one evaluated model.

D.3 Performance Across Clue Counts

The number of anchor-identifying clues varies substantially across KGO and OVD, as characterized in Appendix A.1. We further examine whether this variation is associated with model performance. Figure 10 groups questions by the number of clues $m = |C_e|$ and reports tool-enabled Accuracy within each group.

On KGO, performance does not systematically deteriorate as the number of clues increases. Most models exhibit relatively moderate fluctuations across clue-count ranges, and several achieve higher Accuracy at intermediate counts than on questions with only 1–6 clues. This indicates that clue count alone is not a reliable proxy for KGO difficulty. While a larger clue set may introduce more relational information to integrate, it also provides additional evidence for resolving the research anchor before video observation.

OVD exhibits a more model-dependent pattern. For several stronger tool-using models, including Gemini-3.6-Flash, Kimi-K2.5, Kimi-K2.6, and Qwen3.5-35B-A3B, Accuracy increases substantially in the higher clue-count ranges, with particularly strong performance on questions containing 12 or more clues. In contrast, Gemini-2.5-Pro, Gemini-3-Pro, InternVL3.5-30B-A3B, and Qwen3-VL-30B-A3B-Thinking show much weaker variation across clue counts. This divergence is consistent with richer anchor specifications being useful only when the agent can effectively integrate the available clues into its subsequent web research and video discovery process.

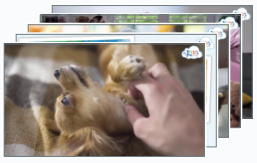
These results suggest that clue count alone is not a reliable indicator of difficulty. Instead, the observed patterns reflect substantial differences in how effectively models exploit the available relational context, particularly in OVD, where anchor resolution must be translated into subsequent web research and video discovery. Since the clue-count groups are not controlled for other instance characteristics, these trends should not be interpreted causally.

E Case Study

We present representative trajectories from Gemini-3-Pro and Gemini-3.6-Flash to illustrate how video deep research agents coordinate external research with visual evidence across the three task configurations. Each case traces the agent's reasoning, tool interactions, and evidence acquisition leading to the final answer. For O2R and KGO, the supplied video is directly available to the model during reasoning, whereas OVD requires the model to first discover relevant footage on the open web and then inspect it through VideoSummary. For clarity and ease of presentation, we condense tool-returned results to the information most relevant to each trajectory, while preserving the evidence used by the model in its subsequent reasoning and final answer. Stacked thumbnails denote videos, while timestamped frames highlight visual evidence examined during the trajectory.

O2R: Observation-to-research Gemini-3-Pro

User Input: Consider the species of fish shown in the first aquarium in this video that appears immediately after the goldfish. That species was first described under a different scientific name. What naturalist is credited as the author for the genus in which that original name was placed?



Supplied video

Step 1 O2R — Observe the video and start research

Reasoning

I first need to determine which fish the question refers to. The goldfish appear in the preceding shot, and the video then cuts to a different aquarium at 00:22. The fish there is bright yellow and appears to be a yellow tang. This gives me a candidate to investigate, but recognizing the fish is only the starting point: the requested answer is the author of the genus used in its original scientific name. I will first look up the yellow tang's scientific name and taxonomic history.



00:20 — the preceding goldfish shot



00:22 — the next aquarium

Tool call

```
WebSearch
query:  yellow tang scientific name history
```

Tool result

[PAGE] Yellow

A page about the color yellow, describing its position between green and orange in the spectrum.

[PAGE] Yellow Paint Colors

A paint-color page discussing shades of yellow. No fish identification or taxonomic history is provided.

Step 2 O2R — Check the visual identification

Reasoning
The first search returned pages about the color yellow rather than useful information about the fish. I should check the visual identification before building further research on it. The relevant frame shows a bright yellow fish in a coral-reef aquarium, so I will use those visible properties in an image search. Comparing the returned fish images with the supplied video should help establish whether the yellow-tang hypothesis is appropriate.

Tool call
`ImageSearch`
`query: bright yellow saltwater aquarium fish`

Tool result
[IMAGE] Yellow tang
A reference image of a yellow tang is returned.
[IMAGE] Hawaiian Yellow Tang
Another yellow-tang reference image is returned for comparison.

Step 3 O2R — Find the species' scientific name

Reasoning
The image results support identifying the fish as a yellow tang. I can now make the textual search more specific to this species. The question emphasizes that the fish was originally described under a different scientific name, so its current name may not be enough. I will search for the yellow tang's taxonomic history and original name, then use that information to determine which genus must be investigated.

Tool call
`WebSearch`
`query: yellow tang taxonomic history original name`

Tool result
[PAGE] Yellow tang
The yellow tang is identified as Zebrasoma flavescens, a species of marine ray-finned fish.
[PAGE] Creature Feature – Yellow Tang
The taxonomy entry lists the current scientific name as Zebrasoma flavescens and the current genus as Zebrasoma.

Step 4 — O2R — Trace the original classification

Reasoning

The search establishes the current scientific name as Zebrasoma flavescens. However, the question asks about the genus in the original name, so simply finding who named Zebrasoma would not resolve it. I need the original description or a historical synonym for this species. I will search using the scientific name together with those terms to recover the earlier taxonomic combination.

Tool call

```
WebSearch
query:  Zebrasoma flavescens original description synonym
```

Tool result

[PAGE] Zebrasoma flavescens Species — Tropical Fish Hobbyist

The species profile for Zebrasoma flavescens states that the fish was originally described in the genus Acanthurus.

Step 5 — O2R — Find the authority for the original genus

Reasoning

The original combination places the yellow tang in Acanthurus, rather than its current genus Zebrasoma. The remaining question is therefore who first proposed Acanthurus. I should look up the genus authority specifically: the author associated with the species description is a different piece of information from the author of its original genus. A targeted search for Acanthurus and its author should supply the final link.

Tool call

```
WebSearch
query:  Acanthurus genus author naturalist
```

Tool result

[PAGE] Acanthurus

Acanthurus was first proposed as a subgenus of Chaetodon in 1775 by the naturalist Peter Forsskål.

Step 6 — O2R — Combine the visual and taxonomic evidence

Reasoning

The relevant video shot identifies the fish as a yellow tang. Research connects that species to the original combination Acanthurus flavescens, and the final search identifies Peter Forsskål as the authority who proposed Acanthurus. This completes the connection from the fish visible after the goldfish shot to the naturalist requested in the question.

Final answer: Peter Forsskål

Figure 11: Observation-to-research case. Video observation identifies the fish; successive searches trace its original genus to the naturalist who named it.

KGO: Knowledge-guided observation Gemini-3.6-Flash

User Input: Consider an individual born in the late 1980s who was educated at a Shanghai university belonging to a Shanghai-Hong Kong university alliance and possessing its own medical school. This person competes in a cue sport overseen by the primary professional billiards and snooker association, took part in a major multi-sport Asian event during the 2010s that included rugby sevens, and was born in a city that is a twin town of the city shown in [IMG1]. In the video, in the overhead shot where the referee is seen re-spotting a ball on the table, what is the frame score displayed for this person?



Supplied video



query image

Step 1 KGO — Identify the city in the image clue

Reasoning

The clues describe a professional cue-sport player born in the late 1980s, educated at a Shanghai university, and connected to an Asian multi-sport event. His birthplace is specified through a sister-city relationship with IMG1, whose city is not named in the question. I will first use reverse image search to identify that city, then look up its twin towns to narrow the birthplace constraint.

Tool call

```
ReverseImageSearch
image_id: 1
```

Tool result

[IMAGE] Novo Mesto aerial view

The image is matched to an aerial photograph of Novo Mesto.

[IMAGE] Novo Mesto, Slovenia

A second image-search result also identifies Novo Mesto, Slovenia.

Step 2 KGO — Recover the birthplace relation

Reasoning

The reverse image search identifies IMG1 as Novo Mesto, Slovenia. The question refers to a player born in one of its twin towns, so this city name gives me a concrete starting point for the geographic clue. I will search its sister-city relationships before combining the resulting birthplace with the sporting and university clues.

Tool call

```
WebSearch
query: "Novo Mesto" "sister cities" OR "twin towns" OR "twin city"
```

Tool result

[PAGE] Novo Mesto, Slovenia

Novo Mesto and Yixing, China, established a sister-city relationship on November 9, 1995.

Step 3 KGO — Check the player's university connection

Reasoning

The search links Novo Mesto to Yixing. This brings Ding Junhui to mind: a Chinese snooker player from Yixing, born in 1987. He is a candidate consistent with the birthplace and age clues. I will now check whether he studied at a Shanghai university, as required by the question, before reading the corresponding score in the supplied video.

Tool call

```
WebSearch
query:  "Ding Junhui" "Shanghai" university OR "Fudan" OR "Jiao Tong"
```

Tool result

[PAGE] Ding Junhui

Ding enrolled at Shanghai Jiao Tong University in 2006 to study Business Administration and Management.

[PAGE] Top 10 sports stars courted by China's best universities

The article also reports Ding Junhui's admission to Shanghai Jiao Tong University.

Step 4 KGO — Read the target player's frame tally

Reasoning

The university search supports Ding Junhui as the target. In the supplied video, the overhead shot around 00:24–00:25 shows the referee re-spotting a ball. The scoreboard reads DING [0] 3 (17) 8 [8] O'SULLIVAN. The yellow boxes show points in the current frame, while the central 3 (17) 8 gives the match tally in frames. Ding's tally is therefore 3.



00:24 — the referee re-spots a ball



00:25 — inset: scoreboard detail

Final answer: 3

Figure 12: Knowledge-guided observation case. The city image supplies a birthplace constraint through a sister-city relation; the agent identifies Ding Junhui and reads his tally of frames won from the supplied video.

OVD: Open-web video discovery Gemini-3-Pro

User Input: Consider the UFC's first visit to Russia, staged in Moscow's vast Olympiysky arena complex and drawing an announced crowd of 22,603. During the official weigh-in for this event, immediately after flexing on the scale, what specific gesture does fighter Rustam Khabilov make?

Step 1 OVD — Identify the event from its description

Reasoning

No video is supplied, so I must first identify the event and then find footage of the relevant weigh-in. The description provides a strong combination of clues: the UFC's first visit to Russia, Moscow's Olympiysky arena, and an announced attendance of 22,603. I will start with the first-event clue to establish the event name and date. That should give me a concrete search context for Rustam Khabilov's appearance on the scale.

Tool call

```
WebSearch
query: UFC first event in Russia Moscow
```

Tool result

[PAGE] UFC Fight Night: Hunt vs. Oleinik

UFC Fight Night: Hunt vs. Oleinik took place on September 15, 2018, at Olimpiyskiy Stadium in Moscow, Russia.

[PAGE] UFC to hold debut event in Russia on September 15

The UFC announcement describes its debut event in Russia, scheduled for September 15 at Moscow's Olimpiyskiy Arena.

Step 2 OVD — Connect the named fighter to this event

Reasoning

The event is UFC Fight Night: Hunt vs. Oleinik in Moscow on September 15, 2018. I now need footage of this particular event's weigh-in, not a different appearance by Rustam Khabilov. I will search using the resolved event name, his name, and the weigh-in context. This should confirm the connection and help narrow the search toward the recording that contains his time on the scale.

Tool call

```
WebSearch
query: "UFC Fight Night: Hunt vs. Oleinik" Rustam Khabilov
weigh-in
```

Tool result

[PAGE] UFC Fight Night: Hunt vs. Oleinik

The event is also called UFC Fight Night 136. The listed results include Rustam Khabilov against Kajan Johnson.

[PAGE] UFC Fight Night: Hunt vs. Oleinik

UFCStats lists Hunt vs. Oleinik on September 15, 2018, and includes the Khabilov–Johnson bout.

Step 3 OVD — Find a recording of the weigh-in

Reasoning

The event records connect Khabilov to the correct card, but they do not show the gesture. I still need a video of the weigh-in itself. I will search explicitly for footage of Khabilov weighing in at UFC Moscow. The search may include other events or appearances by the same fighter, so the title and description of any candidate video will matter when deciding which recording to inspect.

Tool call

```
WebSearch
query:  video of Rustam Khabilov weigh-in UFC Moscow
```

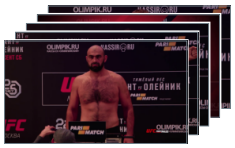
Tool result

[PAGE] UFC Fight Night 42 Weigh-Ins: Benson Henderson vs. Rustam

A YouTube result concerns UFC Fight Night 42 and the Benson Henderson–Rustam Khabilov weigh-in, a different event.

[PAGE] UFC Fight Night Moscow: Official Weigh-in

The video description reads: ”Watch the highlights from the UFC Fight Night: Hunt vs. Oleinik official weigh-ins.”



Retrieved video

UFC Fight Night Moscow: Official Weigh-in

https://www.youtube.com/watch?v=LkRQAfMUUf8

Step 4 OVD — Ask the video tool about the precise action

Reasoning

The result titled "UFC Fight Night Moscow: Official Weigh-in" explicitly names Hunt vs. Oleinik in its description, so it matches the event I have identified. The search has provided a relevant recording, but I still need to determine what Khabilov does immediately after flexing on the scale. I will pass this URL to VideoSummary with the question's precise timing condition. The tool should report the gesture at that moment, which will provide the visual detail needed for the final answer.

Tool call

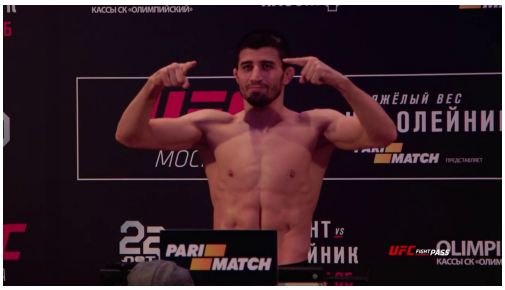
```
VideoSummary
question: What specific gesture does fighter Rustam Khabilov make
immediately after flexing on the scale?
url: https://www.youtube.com/watch?v=LkRQAfMUUf8
```

Tool result

Answer: He extends both index fingers.
Evidence: Immediately after flexing his biceps, Rustam Khabilov extends the index finger of each hand.



03:09 — flexing his biceps



03:11 — extending both index fingers

Step 5 OVD — Use the returned visual evidence to answer

Reasoning

The returned observation describes Khabilov extending the index finger of each hand immediately after flexing his biceps. The successive views show the flexed pose followed by both index fingers being extended, matching the temporal condition in the question. Since the recording is from the identified Hunt vs. Oleinik weigh-in and the observation concerns the requested moment, the answer is that he extends both index fingers.

Final answer: He extends both index fingers.

Figure 13: Open-web video discovery case. Event research leads to the relevant weigh-in footage, and a targeted video-tool call supplies the gesture used in the final answer.