MORE EVIDENCE TO BUILD THAN TO ANSWER: THE SUFFICIENCY-NECESSITY GAP IN TASK SYNTHESIS FOR MULTIMODAL SEARCH AGENT POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal deep search agents are increasingly post-trained on complex synthetic tasks designed to demand iterative retrieval and multi-step reasoning across visual and textual web evidence. Existing synthesis methods typically organize interconnected evidence into multi-step solution structures, deriving questions via structural constraints and solver-based filtering to promote task difficulty. However, evidence sufficient to construct an answerable query is not necessarily required to solve it: visually grounded facts may be recovered through text alone, intended premises can be bypassed, and intermediate answers may leak into the prompt. A task can therefore remain valid yet fail to supervise its intended capabilities—a mismatch we term the *sufficiency-necessity gap*. To bridge this gap, we introduce **EvidenceWeave**, a task synthesis framework that jointly addresses evidence access, premise dependence, and information exposure. By composing qualified atomic relations into multi-step dependencies while keeping intermediate answers implicit, EvidenceWeave ties evidence acquisition and intermediate reasoning to the information requested by the query. Extensive experiments show that our queries yield lower direct-answer recovery rates than the public queries evaluated and elicit visual evidence acquisition and use in correct teacher solutions. Fine-tuning Qwen3.5-27B on the resulting trajectories improves performance across five core benchmarks and leads to more varied and connected evidence use. These results support EvidenceWeave as an approach to narrowing the sufficiency-necessity gap by turning construction evidence into targeted supervision for the acquisition, interpretation, and use of multimodal evidence. Our code, data, and model will be publicly available.

1 INTRODUCTION

Multimodal deep search addresses complex information needs through iterative retrieval and reasoning over visual and textual sources (Geng et al., 2026; Huang et al., 2026; Liu et al., 2026). Answering such queries requires agents to connect visual observations with facts dispersed across sources and reason about their relationships, often bringing together evidence for interdependent subquestions to support a final answer (Zhang et al., 2026b;c; Ning et al., 2026).

Existing work develops these capabilities through post-training on synthetic search tasks, which provide scalable supervision with control over the evidence and reasoning behind a solution (Zhou et al., 2025; Chen et al., 2026a). Recent methods organize collected facts into linked entity paths, reasoning graphs, or exploration chains (Chu et al., 2026; Yao et al., 2026; Jiao et al., 2026). Within these structures, constraints govern reasoning depth and how facts are combined across sources and modalities, while visual grounding, leakage controls, and solver-based filtering promote answerability and substantive tool use (Chen et al., 2026b; Deng et al., 2026).

However, even with these safeguards, evidence sufficient to construct an answerable query may not be necessary to solve it. Alternative sources and clues in the query can bypass the intended derivation, as illustrated in Figure 1: a visually grounded fact may also be available in text, redundant

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.

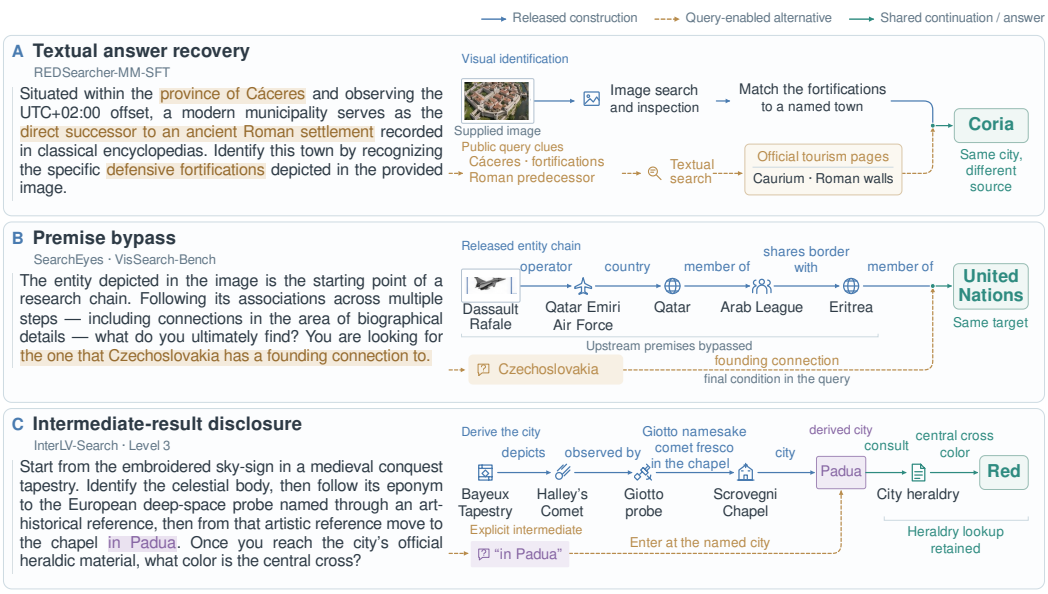


Figure 1: Three manifestations of the sufficiency-necessity gap in released tasks. (A) Textual clues recover a visually grounded answer in REDSearcher-MM-SFT (Chu et al., 2026). (B) An independent clue bypasses upstream premises in SearchEyes VisSearch-Bench (Jiao et al., 2026). (C) An explicitly named intermediate answer shortcuts the derivation in InterLV-Search (Hou et al., 2026).

descriptions can make an upstream premise unnecessary, and a disclosed intermediate answer can remove the need to derive it. We call this mismatch between constructive sufficiency and solution-side necessity the *sufficiency-necessity gap*. Such valid alternatives can enter training through correct teacher trajectories, yielding supervision that misses the intended capabilities.

To address this gap, we introduce **EvidenceWeave**, which builds queries from qualified evidence relations. Atomic QAs begin with sufficient, nonredundant conditions and are assessed for source support, premise dependence, and alternative access before their answers become premises for further questions. This composition allows a visual finding to determine what evidence to retrieve or interpret next. Because revealing that finding in the query would bypass its derivation, formulation replaces intermediate values with their defining relations, preserving the dependencies that connect evidence acquisition to the final information need.

Extensive experiments show that EvidenceWeave queries are less readily answered without search than the public queries evaluated and remain challenging for the base model even with tools. A stronger teacher nevertheless produces correct solutions involving visual evidence acquisition and use, so the retained queries offer successful demonstrations of the capabilities they target. Post-training on these trajectories improves performance across five core benchmarks and leads the model to more often use acquired images in subsequent inspection and search. These gains in both accuracy and evidence use support our approach to narrowing the sufficiency-necessity gap, linking task construction to the capabilities exercised in solving.

We make three contributions:

- We formulate the sufficiency-necessity gap in multimodal search task synthesis and identify three mechanisms through which tasks can lose intended requirements: alternative access, premise bypass, and intermediate-result disclosure.
- We introduce EvidenceWeave, a task synthesis framework that qualifies atomic relations under their premises and carries the resulting access and dependency requirements into composed queries without exposing intermediate answers.
- We demonstrate the value of the resulting supervision through lower direct-answer recovery, visual evidence use in correct teacher solutions, and improvements in both search performance and solving behavior after post-training.

2 Related Work

Multimodal information seeking. A central question in multimodal information seeking is how visual evidence contributes to successful search. Knowledge-intensive visual question answering connects visual content to external facts from structured resources or the web (Wang et al., 2018; Jiang et al., 2025). Learned search policies extend this setting to iterative retrieval, developing selective tool use and sustained interaction through supervised trajectories and reinforcement learning (Wu et al., 2026; Huang et al., 2026; Zhang et al., 2026b). Yet interaction length describes search effort without identifying the contribution of each modality. Benchmarks centered on visual webpage content and interleaved retrieval assess how agents interpret acquired evidence and use those observations to guide subsequent search (Zhang et al., 2026d; Hou et al., 2026). Together, these directions make the role of visual evidence central to multimodal search evaluation: a correct answer alone does not establish whether visual interpretation was needed to obtain it.

Synthetic data for agent learning. Synthetic supervision must provide verifiable tasks whose solutions exercise the capabilities being developed. Deriving tasks from tool executions grounds answers in observed outcomes and supports executable verification (Zhou et al., 2025; Chen et al., 2026a). For deep search, graph-based synthesis further controls intended difficulty through linked facts, visual–textual relations, and evidence dispersion (Chu et al., 2026; Yao et al., 2026; Jiao et al., 2026). These structures specify a supported route, while the question's clues and accessible sources determine what other routes a solver can take. Shortcut-aware approaches accordingly complement structural controls with visual grounding, concealed intermediate entities, or restrictions on identifying clues to limit solutions that bypass the intended search (Chen et al., 2026b; Deng et al., 2026). These controls address distinct sources of bypass, from identifying clues to information shared across sources. Their adequacy must be assessed on the composed question, where individually useful conditions can make one another redundant. The quality of a synthesized task therefore depends on whether its intended information requirements survive composition and final wording, beyond the validity and complexity of its supporting derivation.

Bias and robustness in visual question answering. The distinction between answer accuracy and visual reasoning is a longstanding concern in studies of shortcut learning (Geirhos et al., 2020). Diagnostic datasets vary answer priors between training and evaluation to expose reliance on question–answer correlations (Agrawal et al., 2018). Question-only debiasing addresses this reliance during learning by reducing the influence of language-biased examples (Cadene et al., 2019). These approaches target predictive regularities in the training distribution. External retrieval also allows a solver to obtain the relevant fact directly from text. Access comparisons in multimodal search show that textual clues and retrieval can resolve questions designed for visual search (Zeng et al., 2026). A solution obtained through textual retrieval can thus be factually grounded even when it bypasses online visual inspection. Evaluating visual dependence requires examining the information recoverable through alternative sources as well as the correlations learned from training data.

3 Problem Formulation

A multimodal search task $T = (q, a, \Omega)$ consists of a public query q (including supplied media), reference answer a, and tool environment Ω. Access levels $\lambda_0 \prec \lambda_t \prec \lambda_m$ allow public input and model knowledge, then additional textual retrieval, then online visual retrieval and inspection. Supplied media remain available throughout. Direct QAs use λ_0 without additional retrieval.

Construction evidence forms a finite graph $\mathcal{G} = (V, E)$. Object vertices V_{o} represent entities, contextual values, and visual records (visual items with source context). Fact vertices $V_{\mathrm{r}} = V \setminus V_{\mathrm{o}}$ record predicates and source support, with role-labelled edges $\{u, r\} \in E$ linking facts to objects. Each fact yields atomic question–answer pairs (q_i, a_i) by choosing given and target objects. For i in the finite QA index set $\mathcal{J}$, inputs $U_i \subseteq V_{\mathrm{o}}$ and target $v_i \in V_{\mathrm{o}} \setminus U_i$ are linked to the supporting fact r_i. The answer $a_i = \mathrm{val}(v_i)$ is a contextual name, value, or visual content. Composition follows object identity, not matching answer values, so identical names do not merge distinct entities.

Identifying conditions C_i, supplied or determined by the inputs, specify the relation and scope. With source evidence ε_i, the premises $I_i = C_i \cup \{(u, \mathrm{val}(u)) : u \in U_i\}$ must satisfy $I_i \cup \{q_i, \varepsilon_i\} \vdash_{\lambda_i} a_i$.

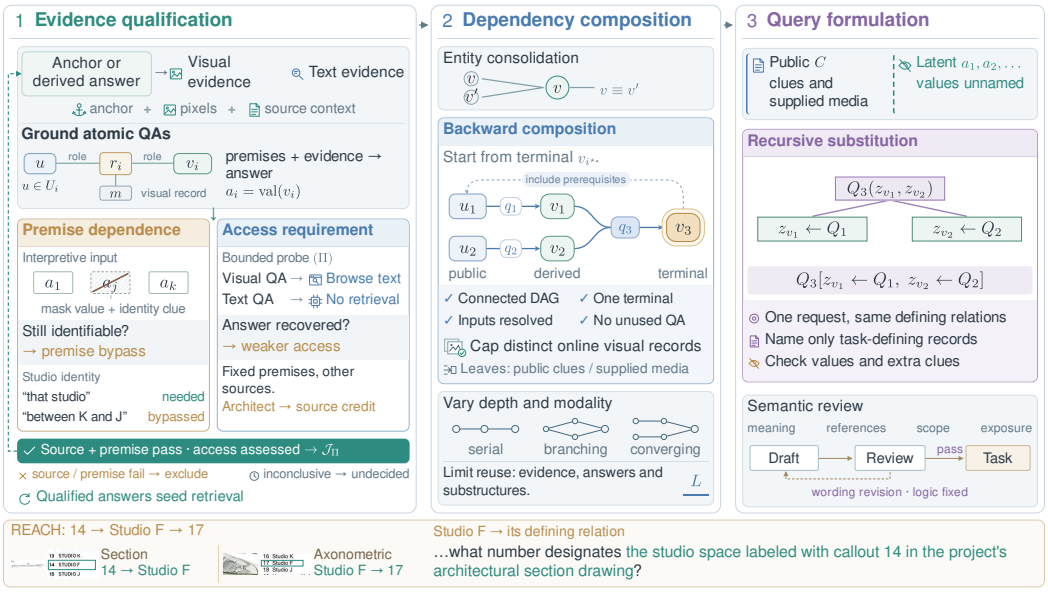


Figure 2: EvidenceWeave qualifies atomic relations and composes dependencies into queries with implicit intermediate values. The example links two drawings through an unnamed studio.

Here $\vdash_\lambda$ denotes answer resolution at access level λ, with λ_i assigned to the QA. This establishes answerability given the stipulated premises and supporting evidence, not the necessity of that access.

A selection $L \subseteq \mathcal{J}$ induces the dependency graph $D_L = (L, E_L)$, where $(j, i) \in E_L$ exactly when $v_j \in U_i$: one answer supplies another QA's input. An acyclic selection with supplied or derived inputs and one terminal $i^\star$ yields $a = a_{i^\star}$. Query formulation $q = \rho(L, C)$ expresses this information need under public conditions C, including supplied media. Preserving the selected relations gives *constructive sufficiency*: a supported route from C to the answer.

Here, access requirements govern fact retrieval, and dependency requirements govern upstream answers needed to locate or interpret evidence. Alternative access, redundant premises, or disclosure of intermediate answers or identifying clues can make these requirements avoidable: the *sufficiency-necessity gap*. Alternative sources and tool sequences are valid if they satisfy these requirements.

4 EvidenceWeave

To address the sufficiency-necessity gap, EvidenceWeave is designed to carry qualified access and dependency requirements into complete queries. It assesses atomic relations under their premises, then composes them while keeping intermediate answers implicit, as Figure 2 illustrates.

4.1 Evidence qualification

Anchored evidence relations. Retrieval is guided by an *anchor*, which is either an initial entity or a previously qualified answer. A multimodal model interprets each visual record using the anchor and source context to identify objects, while pixels must support the decisive relation. Extracted objects and facts enter the evidence graph. Construction seeks sufficient, nonredundant conditions to locate evidence and uniquely determine an answer at the intended granularity, excluding irrelevant inputs and descriptions that independently reveal an input or answer.

Conditional answerability. Premise checks vary supplied information, whereas access checks vary how evidence can be obtained. For an input $u \in U_i$ used to interpret a source, I_i^{-u} removes its value binding and q_i^{-u} its identifying phrase. With the target, source, and other descriptions fixed, the intended dependence is

$$I_i^{-u} \cup \{q_i^{-u}, \varepsilon_i\} \not\vdash_{\lambda_i} a_i. \tag{1}$$

Model-assisted review checks source support and whether remaining or equivalent clues bypass the removed premise. Inputs used to locate evidence instead receive contextual review: supplying the source cannot test their necessity for retrieval, since the locating step has already been completed.

The review and access checks follow a common protocol Π specifying the models, access conditions, budgets, and answer adjudication. Within this protocol, access probes isolate the need for a source by retaining premise values while withholding the nonpublic source itself. Visual QAs permit textual search and browsing beyond the original page (λ_t), while textual QAs permit only public input and model knowledge (λ_0).

An equivalent answer counts as a probe match but may be a guess. Source-supported alternatives establish how it is recovered. After affirmative source and premise reviews, a completed non-match retains the proposed level, whereas a match supports a weaker level for composition. This need not be the minimum possible access, and unavailable judgments cannot justify stronger requirements. Failed finite probes do not establish necessity.

Conditional evidence availability. The QA set qualified under Π, indexed by $\mathcal{J}_\Pi$, determines whether a visual record's identifying premises are supplied or derivable. If so, the record can serve as online evidence. Otherwise, it remains a candidate for supplied media. Its role can change as further relations qualify, but access cannot depend on an answer obtainable only from that record.

4.2 Dependency composition

Composition works backward from a terminal QA, consolidating equivalent object references and selecting producers until all inputs reach public conditions or supplied media. A visual finding can thus identify the subject of a later text-retrieval question.

Let $F \subseteq V_\mathrm{o}$ contain objects explicitly supplied under C, including input visual records whose depicted facts still require interpretation. A selection $L \subseteq \mathcal{J}_\Pi$ forms an acyclic D_L with one terminal QA. Each nonpublic input has one selected producer, other inputs belong to F, and derived objects cannot have public value bindings. These constraints place each QA on a path to the answer, with inputs resolvable from C or preceding QAs, preserving a supported derivation.

Selection caps distinct online visual records, excluding supplied media, varies depth and modality transitions, and limits evidence and answer reuse. Appendix D.2 formalizes these constraints.

4.3 Query formulation

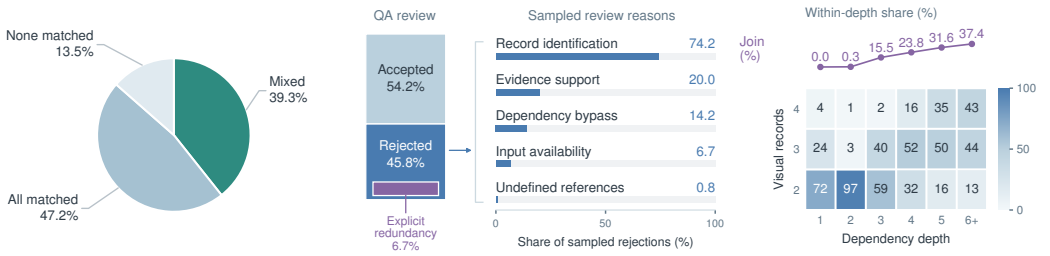
Query formulation replaces intermediate values with defining relations. For $i \in L$, Q_i describes the answer sought by q_i, using public inputs from C and variables z_u for other inputs $u \in U_i \setminus F$. Substituting descriptions from predecessors $p_L(i) = \{j \in L : (j,i) \in E_L\}$ gives the expanded description $R_L(i)$ and terminal description Z_L:

$$R_L(i) = Q_i\big[z_{v_j} \leftarrow R_L(j)\big]_{j \in p_L(i)}, \qquad Z_L = R_L(i^\star), \tag{2}$$

with $R_L(i) = Q_i$ when $p_L(i)$ is empty. Unique producers and acyclicity make substitution well-defined and terminating, including for branching dependencies. The writer renders this structure as $q = \rho(L, C)$, preserving defining relations while masking intermediate values.

In the example, "What number denotes Studio F in the axonometric?" becomes "In the axonometric, what number denotes the studio identified by callout 14 in the section drawing?" The defining relation replaces the studio identity without changing the requested number.

To preserve these dependencies in the final wording, source descriptions must identify the intended record without revealing intermediate answers or providing clues that bypass their defining relations. A final semantic review checks this balance alongside reference clarity and answer scope, revising the wording while holding evidence and task logic fixed and rejecting unresolved record ambiguity. Because atomic probes receive premise values that the complete task leaves implicit, direct-answer and text-search probes are also applied to the completed query to check whether its composition or wording makes those premises avoidable. The algorithms and protocols are detailed in Appendix C.



(a) Recovery within QA groups. (b) Review decisions and concerns. (c) Composed task structure.

Figure 3: Task construction diagnostics. (a) Textual recovery in QA groups with shared context and premises, retaining supplied media. (b) Source-and-premise reviews and overlapping rejection concerns. (c) Visual-record distributions and multi-premise join rates within each dependency depth.

4.4 Analysis of Task Construction

We apply EvidenceWeave to construct approximately 10,000 queries and analyze the evidence qualification and dependency composition recorded during synthesis. Figure 3 summarizes how these stages address the gap between source support and solving requirements. Even within one visual context, facts differ in their accessibility: in Figure 3a, 39.3% of groups sharing visual context and premises contain both matched and non-matched answers under textual access. This variation supports checking access for individual conditional QAs rather than assigning a shared level to the whole record.

Those checks also depend on premises that identify the intended evidence. As Figure 3b shows, sampled model-review concerns more often involve record identification than evidence support (74% versus 20%). Photographs of the same event can yield different answers, so a well-defined task needs premises that distinguish records without revealing upstream results.

Composition then connects qualified visual findings to further information needs. Figure 3c shows that deeper tasks combine serial and converging dependencies. Across tasks, visual QA outputs supply downstream premises in 62.1% of occurrences, and 25.8% of edges leaving visual QAs lead to textual-retrieval or direct QAs. Visual findings thus support later retrieval and inference as well as final answers. Together, these observations support our response to the sufficiency-necessity gap: qualifying access and premise requirements, then carrying them into task dependencies.

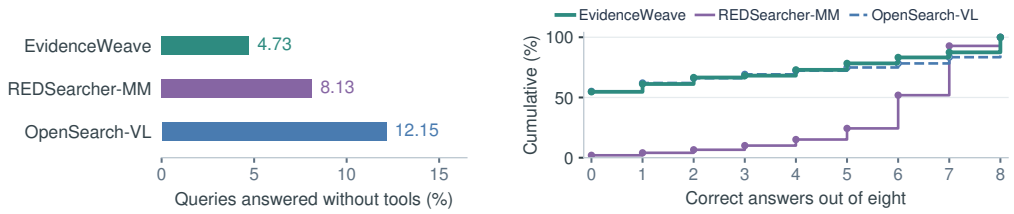
5 Experiments

Addressing problems in query construction alone does not establish that the intended capabilities are exercised during solving. We therefore assess our approach to the sufficiency-necessity gap by tracing these capabilities from completed queries to teacher supervision and post-training. The experiments examine resistance to direct answer recovery, evidence use in teacher solutions, and subsequent changes in model performance and behavior.

5.1 Experimental Setup

Query assessment. Using Qwen3.5-27B (Qwen Team, 2026a), our training backbone, we screen out queries answerable without search and assess the remainder's difficulty. We make 4 direct-answer attempts without tools, retaining supplied media, and discard any query answered correctly at least once. The remainder receive 8 full-tool ReAct attempts. Fewer successes indicate greater difficulty under this solver and budget.

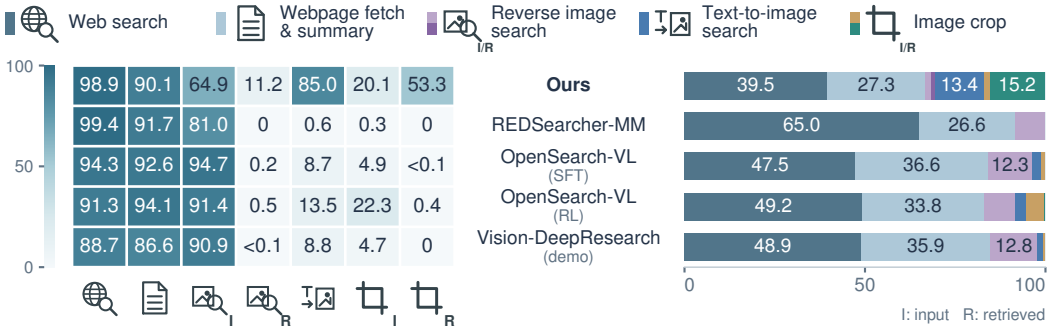
Teacher trajectory synthesis. We sample trajectories for retained queries using Qwen3.7-Plus, selecting those leading to correct answers and using visual tools. The same procedure is applied to queries from REDSearcher-MM (Chu et al., 2026), OpenSearch-VL (Chen et al., 2026b), and Vision-DeepResearch (Huang et al., 2026) to compare evidence use.



(a) Direct-answer recovery.

(b) ReAct success after filtering.

Figure 4: Assessment with Qwen3.5-27B. (a) Queries solved within four direct attempts. (b) Share with at most k correct answers in eight ReAct attempts, among queries with no direct success.



(a) Trajectory coverage (%).

(b) Call composition (%).

Figure 5: Tool use in all teacher attempts. (a) Trajectory coverage, conditional on supplied images for input-image operations. (b) Shares of all calls. I and R mark input and retrieved ancestry.

Model training. Using approximately 19K filtered trajectories generated exclusively from our own query-synthesis pipeline, we perform supervised fine-tuning (SFT) of Qwen3.5-27B to obtain EvidenceWeave-27B. Detailed training configurations are provided in Appendix E.3.

Evaluation. Our query synthesis method aims to strengthen long-horizon multimodal search and evidence integration. Evaluation covers MM-BrowseComp (Li et al., 2026), MMSearch-Plus (Tao et al., 2026), BrowseComp-V^3 (Zhang et al., 2026a), VisBrowse (Zhang et al., 2026d), and BrowseComp-VL (Geng et al., 2026). To test generalization to simpler factual queries, we also include MMSearch (Jiang et al., 2025), SimpleVQA (Cheng et al., 2025), LiveVQA (Fu et al., 2025), and FVQA (Wang et al., 2018). Both models use 4 samples per task with shared tools, budgets, and answer adjudication.

5.2 Query Quality and Teacher Supervision

A query offers limited supervision for search if the model can answer it without acquiring further evidence. We test this possibility through direct answering with Qwen3.5-27B, using only public input and model knowledge. The model solves only 4.73% of EvidenceWeave queries, fewer than either public source assessed, as shown in Figure 4a. After filtering out these queries, over half of those retained resist all eight ReAct attempts. The distribution in Figure 4b indicates difficulty comparable to OpenSearch-VL and greater than REDSearcher-MM, leaving substantial scope for the base model to learn from stronger teacher solutions.

Retained queries then proceed to teacher trajectory synthesis to assess whether this challenge can yield supervision that exercises the targeted capabilities. The teacher answers correctly in 77.7% of the analyzed attempts, and 96.1% of these correct trajectories use visual tools. These solutions demonstrate visual evidence use on tasks that challenge the base model.

The public sources analyzed focus on questions with supplied images. Our queries extend this setting to tasks in which the need for visual evidence emerges during search, with 59.23% of queries

Table 1: Performance comparison on multimodal deep search and factual search benchmarks (%). Bold indicates the best score within each model group. † denotes published results under the original evaluation settings. Unmarked scores are evaluated in this work.

Model	Multimodal deep search					Factual search			
	MMBC	MMS+	BC-V^3	VisB	BC-VL	MMS	Simple VQA	Live VQA	FVQA
LARGER MODELS AND API REFERENCES									
Qwen3.5-397B-A17B	46.0	42.1	31.0	52.7	49.4	78.3	76.4	85.2	77.1
Qwen3.7-Plus	37.1	46.3	42.3	59.8	**55.6**	80.7	77.7	84.6	79.7
Qwen3.8-Max	**53.2**	**59.1**	**56.3**	**66.9**	–	**84.0**	**83.8**	**88.8**	**85.6**
POST-TRAINED SEARCH AGENTS AND BASE REFERENCE (27–32B)									
WebWatcher-32B†	–	–	–	–	27.0	55.3	59.0	58.7	–
SenseNova-MARS-32B†	–	–	–	–	–	74.3	74.1	60.8	72.6
Vision-DeepResearch-30B-A3B†	–	28.5	–	–	53.7	69.6	–	77.6	74.2
OpenSearch-VL-30B-A3B†	–	–	–	–	41.1	68.7	74.9	67.4	73.2
REDSearcher-MM-SFT-30B-A3B†	25.3	20.2	–	–	55.3	70.3	–	78.5	–
REDSearcher-MM-RL-30B-A3B†	23.5	26.6	–	–	**57.2**	72.9	–	79.3	–
MM-DeepResearch-32B†	–	–	–	–	43.0	69.0	67.6	68.0	70.1
MTA-DeepSearch-32B†	–	31.9	–	–	53.8	**82.4**	–	–	76.0
LMM-Searcher-30B†	30.1	34.8	–	48.3	–	72.3	–	–	–
Qwen3.5-27B	31.3	37.6	31.7	48.5	47.9	79.3	74.5	84.5	75.7
EvidenceWeave-27B	**38.4**	**42.8**	**41.0**	**55.6**	54.4	78.0	**75.6**	**85.7**	**77.2**

in this analysis beginning without images, compared with almost none in the public sources. This limits the applicability of input-image operations and partly explains their call shares in Figure 5b, which includes calls from queries without images.

Beyond these input-image operations, text-to-image search, retrieved-image cropping, and reverse search using retrieved images are more widespread in our teacher attempts than in those generated from public-source queries, as Figure 5a illustrates. These operations show how visual evidence is acquired and inspected during actual solving, connecting the capabilities targeted in construction to the behavior elicited by the completed queries. Together with the visual-tool use observed in correct solutions, this coverage supports our approach to the sufficiency-necessity gap: turning evidence sufficient for task construction into supervision that exercises the intended capabilities.

5.3 POST-TRAINING PERFORMANCE AND BEHAVIOR

Using the resulting training corpus, we fine-tune Qwen3.5-27B to assess whether its supervision improves performance and carries the targeted evidence-use behaviors into new tasks.

As illustrated in Table 1, EvidenceWeave-27B improves over the base model by 5.2–9.3 percentage points across all five core benchmarks. Performance on the additional benchmarks remains broadly stable, with small gains or declines, suggesting preserved generalization.

To connect these gains to the capabilities targeted by our queries, we examine post-training changes in evidence acquisition and use on MM-BrowseComp, MMSearch-Plus, BrowseComp-V^3, and VisBrowse. The analysis considers tool selection, output reuse, the onset of visual-tool use, and the prevalence of these behaviors across tasks, distinguishing frequency from tool combinations.

A broader repertoire beyond image lookup. As illustrated in Figure 6a, post-training increases the use of Webpage fetch & summary, image crop, and both image search tools across all four benchmarks, while web search changes less. This selective expansion indicates that the model more often engages the retrieval and inspection capabilities targeted by our queries, giving it more ways to acquire and examine evidence when resolving an information need.

Images connect acquisition, inspection, and further retrieval. The acquired material also becomes input to subsequent operations. In Figure 6b, text-to-image search followed by cropping its output rises from 4.7% to 19.8% of tasks. Reverse search using a crop's output rises from 0.8% to 3.3%. Successive crops also increase. These patterns show that visual tools become connected steps in a solution: acquired images provide material for local inspection, and inspected regions become



(a) Tool-call means and ratios.

(b) Image reuse across operations.

(c) Cumulative visual-tool use.

(d) Visual-tool repertoire per task.

Figure 6: Post-training changes in tool use. (a) EvidenceWeave/base call ratios (linear bars) and per-task means (base → EvidenceWeave). (b) Task coverage of image-output reuse between tools. (c) Visual-tool coverage by turn. (d) Task flows between numbers of distinct visual tool types.

inputs to further search or refinement. This reuse is consistent with our aim to make acquired visual evidence support further progress toward the answer.

Visual operations enter the search earlier. The post-trained model uses visual tools on a larger share of tasks from the start of search, and this separation persists as turns accumulate, as shown in Figure 6c. Additional tasks begin using visual tools in later turns, while the early separation indicates that gathering visual evidence has already become a more common part of the initial search strategy.

The repertoire broadens across tasks. More tasks also draw on multiple visual tools, with the share using at least two types rising from 17.3% to 48.8%. The flows in Figure 6d show how this expansion occurs as tasks move toward larger repertoires. Because repeated calls to one tool do not increase this count, the shift shows that a wider repertoire is exercised across more tasks, complementing the changes in call frequency and connected operations.

Together, these results trace capabilities targeted during query construction into teacher supervision and model behavior on new tasks. Improved accuracy accompanies a more varied and connected use of visual evidence. This correspondence supports our approach to narrowing the sufficiency-necessity gap: bringing the evidence acquisition and use that motivate task construction into actual solving, so that supervision develops capabilities for diverse real-world multimodal search tasks.

6 Conclusion

Evidence sufficient to construct a multimodal search task need not be necessary to solve it, allowing valid solutions to bypass the capabilities intended for training. EvidenceWeave addresses this gap by combining access and premise qualification with query construction that preserves dependencies while keeping intermediate answers implicit. The resulting tasks elicit visual evidence use in correct solutions, and post-training improves multimodal search accuracy alongside more varied and connected evidence use. These findings support EvidenceWeave as an approach to narrowing the sufficiency-necessity gap, showing how evidence used to construct tasks can become effective supervision for the capabilities those tasks were intended to develop.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, clean and reformat data, implement methods, refine the conceptual framework and research hypotheses, provide feedback on research methodology and experimental design, support qualitative analysis of task examples and agent behavior, and interpret experimental results. We have not used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, write proofs, or assist with translation. None of the required-disclosure categories are designated as not applicable. Additionally, we used generative AI tools for writing and editing code, drafting and revising manuscript text, refining the visual presentation of figures, and checking references. The authors directed the research and reviewed and revised AI-assisted manuscript and figure content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies task synthesis for post-training multimodal search agents. Task construction and agent execution use publicly available visual and textual web information, and evaluation uses existing benchmarks. The resulting tasks may inherit biases from web sources and the models used for synthesis and assessment. Released artifacts will respect the licenses and redistribution terms of their sources, with access credentials and private configuration information excluded from public materials.

REPRODUCIBILITY STATEMENT

Section 4 describes the task synthesis framework, with algorithms and implementation details in Appendix C. Section 5.1 and Appendix E describe query assessment, teacher trajectory synthesis, model training, and evaluation. Supplementary evidence and illustrative cases are provided in Appendices F and G.2, and Appendix D.2 provides the formal analysis. Our code, data, and model will be publicly released, together with configuration records and scripts for reproducing the reported tables and figures.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Alibaba Cloud. qwen3.7-plus. Alibaba Cloud Model Studio model documentation, 2026. URL https://docs.modelstudio.console.alibabacloud.com/en/model-studio/qwen3-7-plus.

ArchDaily. The Kennedy Center for the Performing Arts / Steven Holl Architects, September 2019. URL https://www.archdaily.com/924378.

Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing Unimodal Biases for Visual Question Answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/51d92be1c60d1db1d2e5e7a07da55b26-Paper.pdf.

Aili Chen, Chi Zhang, Junteng Liu, Jiangjie Chen, Chengyu Du, Yunji Li, Ming Zhong, Qin Wang, Zhengmao Zhu, Jiayuan Song, Ke Ji, Junxian He, Pengyu Zhao, and Yanghua Xiao. DIVE: Scaling Diversity in Agentic Task Synthesis for Generalizable Tool Use. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=4uM4P5NaLU.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. OpenSearch-VL: An Open Recipe for Frontier Multimodal Search Agents, 2026b. URL https://arxiv.org/abs/2605.05185.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Hangyuan Ji, Tongliang Li, Wenhao Huang, and Zhoujun Li. SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4637–4646, October 2025.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, Gao Huang, and Lewei Lu. SenseSearch: Empowering Vision-Language Models with High-Resolution Agentic Search-Reasoning via Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26508–26517, June 2026.

Zheng Chu, Xiao Wang, Jack Hong, Huiming Fan, Yuqi Huang, Yue Yang, Guohai Xu, Chenxiao Zhao, Cheng Xiang, Shengchao Hu, Dongdong Kuang, Ming Liu, Bing Qin, and Xing Yu. REDSearcher: A Scalable and Cost-Efficient Framework for Long-Horizon Search Agents, 2026. URL https://arxiv.org/abs/2602.14234.

Jia Deng, Yimeng Chen, Xiaoqing Xiang, Ziyang Zeng, Shuo Tang, Wayne Xin Zhao, Feng Chang, Chuan Hao, Yuan Wei, Ran Tao, Bryan Dai, and Ji-Rong Wen. FORT-Searcher: Synthesizing Shortcut-Resistant Search Tasks for Training Deep Search Agents, 2026. URL https://arxiv.org/abs/2606.12087.

Yifan Du, Zikang Liu, Jinbiao Peng, Jie Wu, Junyi Li, Jinyang Li, Wayne Xin Zhao, and Ji-Rong Wen. Towards long-horizon agentic multimodal search, 2026. URL https://arxiv.org/abs/2604.12890.

Mingyang Fu, Yuyang Peng, Dongping Chen, Zetong Zhou, Benlin Liu, Yao Wan, Zhou Zhao, Philip S Yu, and Ranjay Krishna. Seeking and Updating with Live Visual Knowledge. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2513. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6c7c9811d06b41b320b69abf37234f84-Paper-Datasets_and_Benchmarks_Track.pdf.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, Huifeng Yin, Yong Jiang, Pengjun Xie, Fei Huang, Huaxiu Yao, Yi R. Fung, and Jingren Zhou. WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 61240–61271, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/63ca1cca58101545e937302be7050098-Paper-Conference.pdf.

Bohan Hou, Jiuning Gu, Jiayan Guo, Ronghao Dang, Sicong Leng, Xin Li, Xuemeng Song, and Jianfei Yang. InterLV-Search: Benchmarking Interleaved Multimodal Agentic Search, 2026. URL https://arxiv.org/abs/2605.07510.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Yao Hu, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-DeepResearch: Incentivizing DeepResearch Capability in Multimodal Large Language Models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2Cgy0P2v8s.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, jiayi lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Gao Peng, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Unveiling the Potential of Large Models as Multi-modal Search Engines. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025,

pp. 1662–1699, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/04e6525463afaa460b5a2e6868b18f18-Paper-Conference.pdf.

Zhengbo Jiao, Yiming Cheng, Yilei Jiang, Kaituo Feng, Rui Huang, Tianyi Jiang, Juanxi Tian, Jiapeng li, Qunzhong Wang, Tailai Chen, Qianshan Wei, Chuan Xiao, Shanyu Rong, Yangfu Li, Yanhan Zhou, Yunpu Ma, Yifan Zhang, and Xiangyu Yue. SearchEyes: Towards Frontier Multimodal Deep Search Intelligence via Search World Simulation, 2026. URL https://arxiv.org/abs/2607.05943.

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, 2026. URL https://arxiv.org/abs/2508.13186.

Yikun Liu, Yuan Liu, Haicheng Wang, Zhongyin Zhao, Le Tian, Xiao Zhou, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. POINTS-Seeker: An Open Recipe for Multimodal Search Agents with Visual Memory Management, 2026. URL https://arxiv.org/abs/2604.14029.

Xuying Ning, Dongqi Fu, Tianxin Wei, Mengting Ai, Jiaru Zou, Ting-Wei Li, Hanghang Tong, Yada Zhu, Hendrik Hamann, and Jingrui He. MC-Search: Evaluating and Enhancing Multimodal Agentic Search with Structured Long Reasoning Chains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JEGDp1E4OH.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

XIANGYU PENG, Can Qin, An Yan, Xinyi Yang, Ran Xu, Zeyuan Chen, and Chien-Sheng Wu. MTA-Agent: An Open Recipe for Multimodal Deep Search Agents. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=6rtp2PaElk.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A New Bar for Coding and Cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Suzanne Stephens. The Reach at the Kennedy Center by Steven Holl Architects and BNIM. *Architectural Record*, October 2019. URL https://www.architecturalrecord.com/articles/14281.

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. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 70385–70422, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7291251c42879a93141d1340a96c6a22-Paper-Conference.pdf.

Turismo de Extremadura. Coria. Official regional tourism portal, n.d. URL https://www.turismoextremadura.com/es/explora/Coria/. Accessed September 26, 2026.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. FVQA: Fact-Based Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427, Oct 2018. ISSN 1939-3539. doi: 10.1109/tpami.2017.2754246. URL http://dx.doi.org/10.1109/TPAMI.2017.2754246.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun MA, and Ziwei Liu. MMSearch-r1: Incentivizing LMMs to search. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pp. 2456–2487, San Diego, California, United States, jul 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.114. URL https://aclanthology.org/2026.acl-long.114/.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. MM-DeepResearch: A Simple and Effective Multimodal Agentic Search Baseline. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XZAOCytvKQ.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023.

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, 2026. URL https://arxiv.org/abs/2602.02185.

Huanyao Zhang, Jiepeng Zhou, Bo Li, Bowen Zhou, Yanzhe Shan, Haishan Lu, Zhiyong Cao, Jiaoyang Chen, Yuqian Han, Zinan Sheng, Zhengwei Tao, Hao Liang, Jialong Wu, Yang Shi, Yuanpeng He, Jiaye Lin, Qintong Zhang, Guochen Yan, Runhao Zhao, Zhengpin Li, Xiaohan Yu, Lang Mei, Chong Chen, Wentao Zhang, and Bin Cui. BrowseComp-V^3: A Visual, Vertical, and Verifiable Benchmark for Multimodal Browsing Agents, 2026a. URL https://arxiv.org/abs/2602.12876.

Ruiyang Zhang, Qianguo Sun, Chao Song, Yiyan Qi, and Zhedong Zheng. VSearcher: Long-Horizon Multimodal Search Agent via Reinforcement Learning, 2026b. URL https://arxiv.org/abs/2603.02795.

Zhengbo Zhang, Changtao Miao, Jinbo Su, Zhaowen Zhou, Chunxia Zhang, Xukai Wang, Ruiqi Liu, Kaiyuan Zheng, Jiansheng Cai, Bo Zhang, Zhe Li, Shiming Xiang, and Ying Yan. VisualSeeker: Towards Visual-Native Multimodal Agentic Search via Active Visual Reasoning, 2026c. URL https://arxiv.org/abs/2606.15231.

Zhengbo Zhang, Jinbo Su, Zhaowen Zhou, Changtao Miao, Yuhan Hong, Qimeng Wu, Yumeng Liu, Feier Wu, Yihe Tian, Yuhao Liang, Zitong Shan, Wanke Xia, Yi-Fan Zhang, Bo Zhang, Zhe Li, Shiming Xiang, and Ying Yan. VisBrowse-Bench: Benchmarking Visual-Native Search for Multimodal Browsing Agents, 2026d. URL https://arxiv.org/abs/2603.16289.

Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. Self-Challenging Language Model Agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113959–113991. Curran Associates, Inc., 2025. doi: 10.52202/085713-3801. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a5a305fac88fb4ae40969cfec5eef48d-Paper-Conference.pdf.

A NOTATION

Table 2 summarizes the notation used in the main text and the supplementary method analysis. Object identities and their answer values are distinguished throughout.

Table 2: Notation for tasks, evidence relations, and query composition.

Symbol	Meaning
Tasks and information requirements	
$T = (q, a, \Omega)$	Query (including supplied media), reference answer, and tool environment.
$\lambda_0 \prec \lambda_t \prec \lambda_m$	Public input and model knowledge; additional textual retrieval; additional online visual access.
$\vdash_\lambda$	Resolution of a question under access level λ.
Π	Assessment protocol: models, access conditions, budgets, and answer adjudication.
$\ell_{\mathrm{src}}, \ell_{\mathrm{dep}}, \ell_{\mathrm{disc}}$	Indicators of alternative access, premise bypass, and intermediate-result disclosure.
Evidence and atomic QAs	
$\mathcal{G} = (V, E)$	Evidence graph with vertices V and role-labelled edges E.
$V_{\mathrm{o}}, V_{\mathrm{r}}$	Object vertices and supported fact vertices, respectively.
$i \in \mathcal{J}, (q_i, a_i)$	Atomic QA index and question–answer pair supported by fact vertex r_i.
U_i, v_i, val	Input objects, target object, and contextual value/content function; $a_i = \mathrm{val}(v_i)$.
C_i, I_i	Identifying conditions and their union with input object–value bindings.
I_i^{-u}, q_i^{-u}	Premises and question after withholding input u's value binding and corresponding identifying phrase.
ε_i, λ_i	Supporting evidence and assigned access level after assessment.
$w_\Pi(i)$	Weaker-access probe outcome: match, judged non-match, or inconclusive.
$\mathcal{J}_\Pi$	Indices of QAs satisfying grounding, premise, and access assessments.
$K, \mathrm{cl}_\Pi(K)$	Objects with available values and their least closure under qualified QAs whose inputs are all available.
$\mathcal{J}_\Pi^v(m)$	Qualified QAs supported by visual record m and assigned access level λ_m.
$M_{\mathrm{online}}(K)$	Visual records outside K eligible for online acquisition from its available bindings.
Composition and query formulation	
$L, D_L = (L, E_L)$	Selected QA indices and their dependency graph; $(j, i) \in E_L$ iff $v_j \in U_i$.
C, F	Public conditions (including supplied media) and objects with explicitly supplied values.
$O_L, \partial L$	Derived objects and inputs not produced by the selected QAs.
$M_v(L), v_{\min}, v_{\max}$	Records supporting selected online visual-access QAs, excluding supplied media, and count bounds.
$i^\star, a = a_{i^\star}$	Terminal QA index and complete-task answer.
$Q_i, z_u, p_L(i)$	Answer description, unresolved input variable, and immediate predecessor QAs.
$R_L(i), Z_L$	Recursively expanded description and the terminal description.
$\mathfrak{L}_\Pi(\mathcal{G}, F)$	Structurally feasible QA selections, with C and visual-record bounds held fixed.
$q = \rho(L, C)$	Query formulation under public conditions C, including supplied media.

B LIMITATIONS

Evidence qualification relies on finite solver probes and model-assisted review. A failed probe does not prove that alternative access is impossible, and an answer match can reflect a guess rather than

a source-supported solution. Premise review can miss an identifying clue, while composition and wording can introduce bypasses that were absent from individual atomic QAs. EvidenceWeave therefore narrows the sufficiency-necessity gap under the assessed conditions rather than certifying universal necessity.

The post-training comparison evaluates the supervision produced by the complete framework. It does not isolate the training contribution of each construction mechanism. The behavior analysis uses one evaluation attempt per model and task, so it characterizes the observed strategies rather than variation across repeated generations or independent training runs. Although performance gains accompany changes in visual evidence acquisition and reuse, this comparison does not establish that increased visual-tool use caused those gains. Variation in online retrieval conditions may also affect both tool use and task outcomes.

C Task Synthesis Details

C.1 Evidence extraction and object identity

Records are interpreted with their anchors and source context to extract atomic QAs. Inputs that do not affect the target and descriptions that bypass an input are removed. Equivalent object references are consolidated only when identity is supported, allowing a shared entity to connect facts across sources. Only answers from qualified QAs extend exploration.

C.2 Evidence qualification

Figure 7 summarizes the qualification interventions. Premise assessment removes an interpretive input and its identifying phrase while retaining the source and other descriptions. Locating premises instead receive contextual review. Access probes retain premise values and supplied media but withhold the nonpublic source, permitting alternative sources under the tested access level.

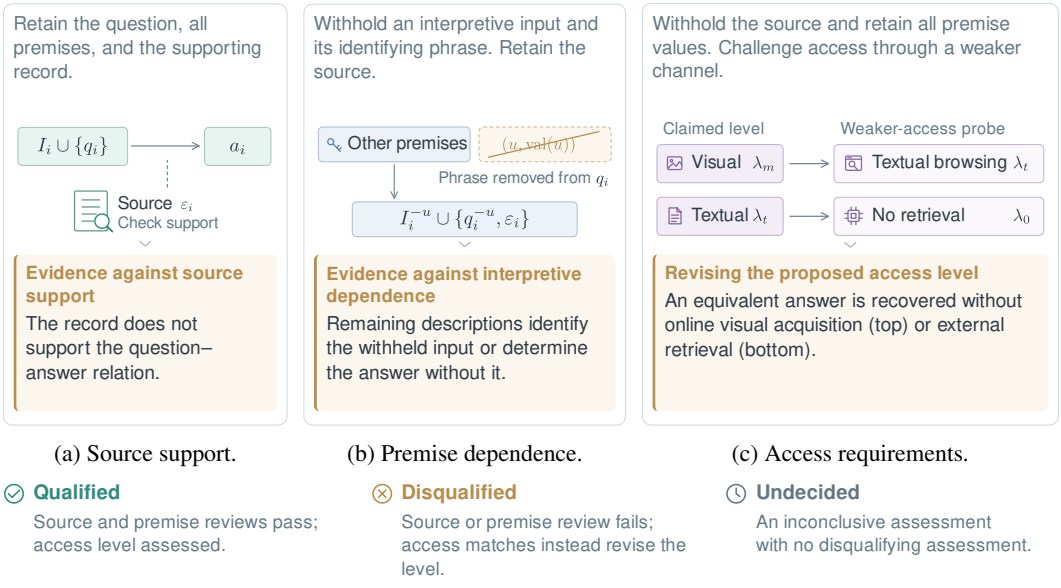


Figure 7: Information interventions used to qualify atomic QAs. (a) Source review checks the answer against the supporting record and premises. (b) Removing an interpretive input and its identifying phrase tests premise dependence. (c) Weaker-access probes test answer recovery while retaining supplied media and model knowledge. Textual retrieval may use other sources. Locating premises receive contextual review rather than input-removal tests.

Source and premise reviews yield affirmative, negative, or inconclusive judgments. Access outcomes are recorded as $w_\Pi(i) \in \{0, 1, \perp\}$: an equivalent answer, a completed non-match, or an unavailable

Algorithm 1 Backward completion of a terminal QA

Require: Evidence graph $\mathcal{G}$, qualified QAs $\mathcal{J}_\Pi$, public conditions C and bindings F
Require: Terminal index $i^\star$; online visual-record bounds $v_{\min}, v_{\max}$
Ensure: A feasible selection ending at $i^\star$, or FAILURE

function COMPLETE(L)
  Form D_L, O_L, and ∂L from the selected QAs
  if $D_L \notin$ DAG **or** $|O_L| \neq |L|$ **or** $O_L \cap F \neq \varnothing$ **then**
    return FAILURE
  end if
  if $\partial L \subseteq F$ **then**
    if $L \in \mathfrak{L}_\Pi(\mathcal{G}, F)$ **then**
      return L
    end if
    return FAILURE
  end if
  Choose an unresolved input $u \in \partial L \setminus F$
  for each $j \in \mathcal{J}_\Pi$ with $v_j = u$ **do**
    $L' \leftarrow$ COMPLETE($L \cup \{j\}$)
    if $L' \neq$ FAILURE **then**
      return L'
    end if
  end for
  return FAILURE
end function
return COMPLETE($\{i^\star\}$)

judgment. Visual QAs are probed under textual retrieval (λ_t), and textual QAs under public input and model knowledge (λ_0).

Negative reviews exclude a relation. After affirmative reviews, a non-match retains the proposed access level and a match assigns the weaker level. Unavailable outcomes remain undecided. Policy refusals do not justify stronger requirements or enter match/non-match statistics without a valid answer judgment. Retained QAs form $\mathcal{J}_\Pi$.

Online availability follows the closure in Equation 3: a record is eligible once the premises identifying it are supplied or derivable through qualified QAs. Otherwise, it remains a candidate for supplied media. Supplied records are excluded from online acquisition requirements.

C.3 DEPENDENCY COMPOSITION

Algorithm 1 completes a terminal QA backward until every input has a selected producer or public binding. Branches requiring the same object share its producer, preserving object identity across the selected dependencies.

Each recursive call adds a producer from the finite QA set. Cycles and multiple producers are rejected, and all added branches lead to the chosen terminal. Distinct online visual records are counted after excluding supplied media. The implementation samples terminal QAs and alternative producers within fixed search budgets, then ranks feasible selections by depth, visual-record count, modality transitions, topology, and source diversity. Selection also limits evidence and answer reuse. Structural properties are derived in Appendix D.2.

C.4 QUERY FORMULATION

Algorithm 2 substitutes defining relations for intermediate values in topological order. Shared predecessors retain the same object identity even when their descriptions appear in multiple branches.

The writer may combine clauses while preserving their antecedents and keeping intermediate values implicit. Source attributes are included only as public conditions or through retained defining relations. Semantic review checks grounding, answer scope, reference resolution, and intermediate-value disclosure.

Algorithm 2 Query formulation of a feasible selection

Require: Evidence graph $\mathcal{G}$ and qualified QAs $\mathcal{J}_\Pi$
Require: Feasible selection L, public conditions C, and explicit bindings F
Ensure: Candidate query q and answer $a = a_{i^\star}$ for semantic review
Let $i^\star$ be the unique sink of D_L
for each $i \in L$ in a topological order of D_L **do**
    Form Q_i, instantiating public inputs and using z_u for each $u \in U_i \setminus F$
    $R_L(i) \leftarrow Q_i$
    for each $j \in p_L(i)$ **do**
        $R_L(i) \leftarrow R_L(i)[z_{v_j} \leftarrow R_L(j)]$
    end for
end for
$Z_L \leftarrow R_L(i^\star)$
Express Z_L under C as a coherent information need q with one terminal request
return $(q, a_{i^\star})$

Complete queries receive direct-answer probes followed, when unsuccessful, by textual search without online visual acquisition. At most one wording revision is allowed before renewed checks, with evidence, relations, and the terminal answer fixed. These construction-time checks precede the separate four-attempt direct and eight-attempt ReAct assessment of the training backbone.

D Structural Properties and Information Requirements

D.1 Evidence representation

The evidence graph records a supported relation before choosing which of its objects to ask about. Its incidence representation connects each fact vertex $r \in V_{\mathrm{r}}$ to the participating objects in V_{o}, with edge annotations retaining multiple roles when an object participates more than once. A conditional QA specifies $U_i \cup \{v_i\} \subseteq N_{\mathcal{G}}(r_i)$ and $v_i \notin U_i$, where $N_{\mathcal{G}}(r_i)$ denotes the adjacent vertices. Different choices of inputs and target can thus yield different questions about the same fact.

These questions induce dependencies through object identity, rather than through equality of answer strings. The evidence graph may contain cycles or alternative readings of a fact, while a feasible QA selection induces an acyclic dependency graph. A fact's presence in the graph establishes source support during construction. Whether its value is explicitly provided to the solver depends on C. Supplying a visual record provides content to inspect, not explicit value bindings for every fact depicted within it. A supporting record also need not be an input object of every QA it supports.

D.2 Structural properties of dependency composition

For a fixed evidence graph, let $K \subseteq V_{\mathrm{o}}$ contain the objects whose values are available to the solver. With identifying conditions supplied or determined by inputs, the QA closure $\mathrm{cl}_\Pi(K)$ is the least superset of K closed under the qualified QAs. Its iterative construction, which determines the reachable premise values, is

$$K_0 = K, \qquad K_{t+1} = K_t \cup \{v_i : i \in \mathcal{J}_\Pi,\ U_i \subseteq K_t\}, \qquad \mathrm{cl}_\Pi(K) = \bigcup_{t \geq 0} K_t. \tag{3}$$

Each strict update adds an object from the finite set V_{o}, so the sequence stabilizes after at most $|V_{\mathrm{o}} \setminus K|$ strict updates. Its limit is closed under every qualified QA. Conversely, induction on t shows that any closed superset of K contains every K_t, establishing that the limit is the least such set. Requiring all inputs rules out unsupported circular derivations.

For a visual record m, let $\mathcal{J}^v_\Pi(m)$ index qualified QAs supported by m and assigned online visual access λ_m, with source identification assessed during review. Its eligibility for online acquisition is

$$m \in M_{\mathrm{online}}(K) \quad \Longleftrightarrow \quad m \notin K \ \wedge\ \exists i \in \mathcal{J}^v_\Pi(m) : U_i \subseteq \mathrm{cl}_\Pi(K). \tag{4}$$

A record without such a QA remains a candidate for supplied media.

For public conditions C and their explicit bindings F, a selection L derives $O_L = \{v_i : i \in L\}$ and has unresolved boundary $\partial L = (\bigcup_{i \in L} U_i) \setminus O_L$. Let $M_v(L)$ collect distinct records supporting its

online visual-access QAs, excluding records supplied in C. With lower and upper bounds $v_{\min}$ and $v_{\max}$, the feasible family is

$$\mathfrak{L}_{\Pi}(\mathcal{G}, F) = \left\{ L \subseteq \mathcal{J}_{\Pi} \;\middle|\; \begin{array}{c} D_L \in \mathsf{DAG}, \quad |\operatorname{sink}(D_L)| = 1, \\ |O_L| = |L|, \quad O_L \cap F = \varnothing, \\ \partial L \subseteq F, \quad v_{\min} \le |M_v(L)| \le v_{\max} \end{array} \right\}, \tag{5}$$

where DAG is the class of directed acyclic graphs and $\operatorname{sink}(D_L)$ contains QAs with no downstream dependent.

For a feasible selection L, the same argument has a directed interpretation. Consider any topological order of D_L. Every input of its first QA belongs to F. Thereafter, each input is either public or produced earlier in that order. The condition $|O_L| = |L|$ supplies a unique producer for every derived object, while $O_L \cap F = \varnothing$ prevents a public value from being introduced again as an unknown. Finally, every vertex in a finite DAG reaches a sink. Since D_L has only one sink, every selected QA contributes to a path ending at the terminal question.

These properties make recursive substitution in Equation 2 well-defined. Each nonpublic input variable has one producer, and expansion terminates without leaving free input variables. Shared predecessors may contribute to several branches without being treated as distinct facts. This is a structural guarantee about the selected derivation. It does not establish that alternative sources cannot answer the task, or that its eventual wording preserves every dependency. Information interventions and semantic review assess those additional requirements.

D.3 Information requirements and alternative solutions

A shortcut witness identifies a particular way to avoid an intended requirement. Its scope depends on the information supplied during assessment: an atomic probe receives premise values that can remain latent in the complete query, whereas a task-level probe must begin with the public query alone and obtain any premises withheld from it.

For access requirements, let $\mathbf{1}[\cdot]$ denote the indicator function. With the nonpublic supporting source withheld, an equivalent answer obtainable below λ_i yields the witness

$$\ell_{\mathrm{src}}(i) = \mathbf{1}[\exists \lambda \prec \lambda_i : \; I_i \cup \{q_i\} \vdash_{\lambda} a_i] \,. \tag{6}$$

This includes a visually grounded fact recovered through textual browsing, as well as a textual fact already available from model knowledge. For a premise used to determine the answer within a source, a bypass assessment instead holds that source available and removes the input binding and its identifying phrase. Using I_i^{-u} and q_i^{-u} from Equation 1, answer recovery gives

$$\ell_{\mathrm{dep}}(u, i) = \mathbf{1}\big[I_i^{-u} \cup \{q_i^{-u}, \varepsilon_i\} \;\vdash_{\lambda_i}\; a_i\big] \,. \tag{7}$$

A remaining description that identifies the withheld input is likewise evidence against this source-conditioned dependence. The same intervention does not establish whether the input was needed to locate the source. At the level of the realized query, $\ell_{\mathrm{disc}}(j; q) = 1$ when the public input directly states or independently identifies the answer a_j of a nonterminal QA $j \in L$ without requiring its defining relation, and is zero otherwise. Together, these witnesses concern loss of access, a required premise, or a latent intermediate result. Finite protocols can detect such violations, but an unsuccessful probe does not certify that none exists.

E Experimental Protocols

E.1 Query assessment

Qwen3.5-27B makes 4 direct-answer attempts with supplied media retained. Queries answered correctly at least once are removed before 8 full-tool ReAct attempts (Yao et al., 2023). Query wording, reference answers, and supplied media remain fixed. The reported success count is the number of correct answers among these attempts. OpenSearch-VL uses its SFT queries in both stages.

E.2 Teacher synthesis and data selection

Qwen3.7-Plus (Alibaba Cloud, 2026) generates teacher trajectories. The behavior comparison pools all attempts from four synthesis batches before filtering. The training corpus selects correct-answer trajectories with visual tool calls and nonempty image references from six batches of our own pipeline, including those four. The public comparison uses REDSearcher-MM, OpenSearch-VL SFT and RL queries, and pooled Vision-DeepResearch generations.

Image origin is inherited through cropping: descendants of supplied images remain input images, and descendants of search results remain retrieved images. Mixed-target reverse searches are assigned to input images. Input-image search and crop coverage use only trajectories with supplied images, while call shares use all calls. Unresolved origins remain in the total call denominator. Input-image availability is counted per query, independently of repeated attempts.

E.3 Model training configuration

Selected trajectories pass checks for image-placeholder consistency and tool-call structure, then are split by query into 90% training and 10% validation. Training runs for 3 epochs and 405 updates, with validation and checkpoint saving once per epoch. Evaluation uses the final checkpoint.

The global batch size is 128, using 64 workers with sequence parallelism of 8 and 8 data-parallel replicas. Each replica uses microbatches of 1 and 16 gradient-accumulation steps, with final partial batches dropped. Table 3 collects the remaining training settings.

All other training options use the framework defaults.

Table 3: Hyperparameters for supervised fine-tuning of Qwen3.5-27B.

Parameter	Value
Trainable parameters	All parameters (27B)
Visual encoder	Trainable
Optimizer	AdamW
Learning rate	1×10^{-6}
Weight decay	0
Learning-rate schedule	Cosine decay
Warmup ratio	0.1
Maximum sequence length	256k tokens
Numerical precision	BF16 mixed precision
Gradient clipping	1.0
Random seed	42

E.4 Published evaluation settings

Published scores follow the original reports of WebWatcher (Geng et al., 2026), SenseNova-MARS (Chng et al., 2026), Vision-DeepResearch (Huang et al., 2026), REDSearcher-MM (Chu et al., 2026), OpenSearch-VL (Chen et al., 2026b), MM-DeepResearch (Yao et al., 2026), MTA-DeepSearch (PENG et al., 2026), and LMM-Searcher (Du et al., 2026). REDSearcher-MM reports separate SFT and RL checkpoints. Other reference models include Qwen3.8-Max (Qwen Team, 2026b).

Vision-DeepResearch uses FVQA and LiveVQA subsets and single-image MMSearch-Plus. LMM-Searcher also uses single-image MMSearch-Plus, with a 100-turn budget and context management. MM-DeepResearch evaluates an English SimpleVQA subset. WebWatcher averages repeated evaluations and uses a BC-VL subset.

F Supplementary Figures

F.1 Access transitions in constructed dependencies

Figure 8 gives the access-level composition of producer–consumer QA pairs in the constructed tasks.

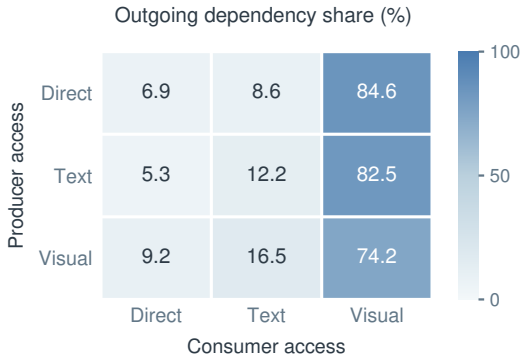
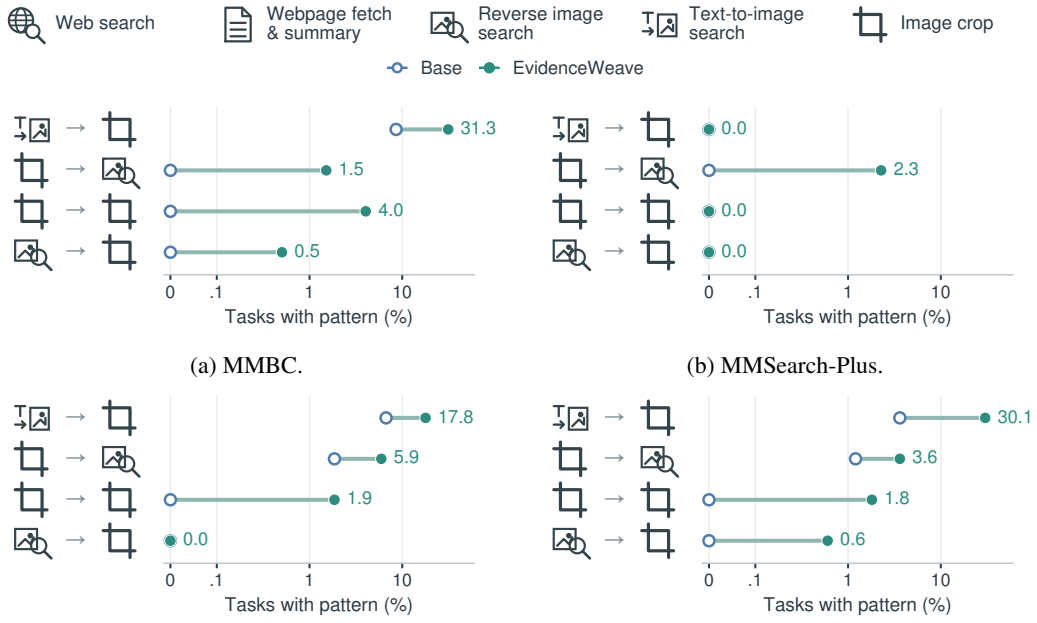


Figure 8: Access requirements along constructed QA dependencies. Rows identify the producer QA's assigned access level and columns the consumer's. Values are percentages of outgoing edges within each row. Direct access retains supplied media but permits no additional retrieval.

F.2 Tool organization across core benchmarks

Figures 9 and 10 provide benchmark-specific views of image reuse and the onset of visual-tool use.



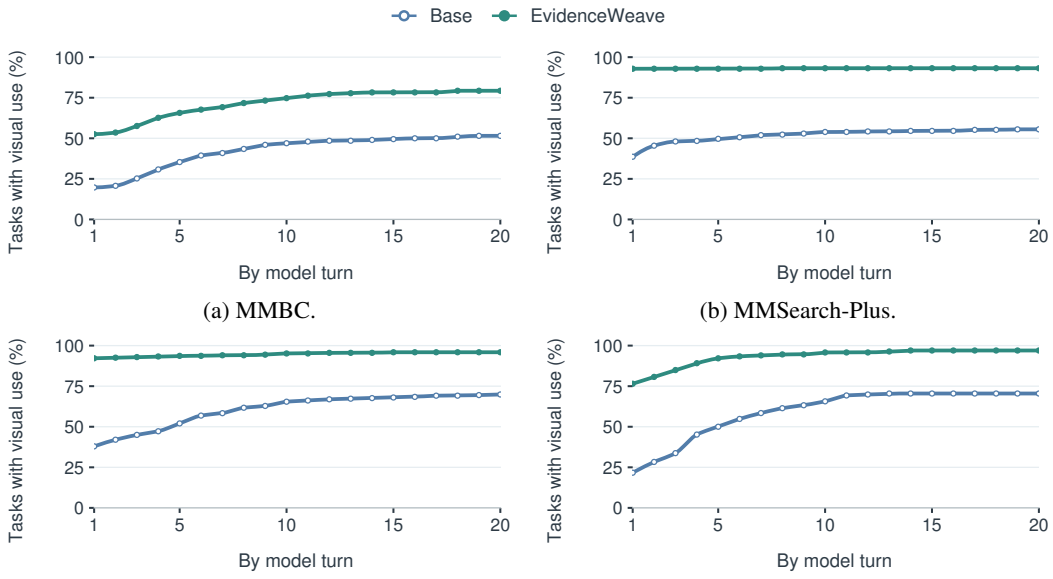
(a) MMBC.

(b) MMSearch-Plus.

(c) BrowseComp-V^3.

(d) VisBrowse.

Figure 9: Image-linked tool use before and after post-training on each core benchmark. Points give the percentage of tasks exhibiting each pattern. Icon arrows denote reuse of an image output, and patterns may overlap within a task. The horizontal scale is linear up to 0.1% and logarithmic above it, preserving zero coverage.

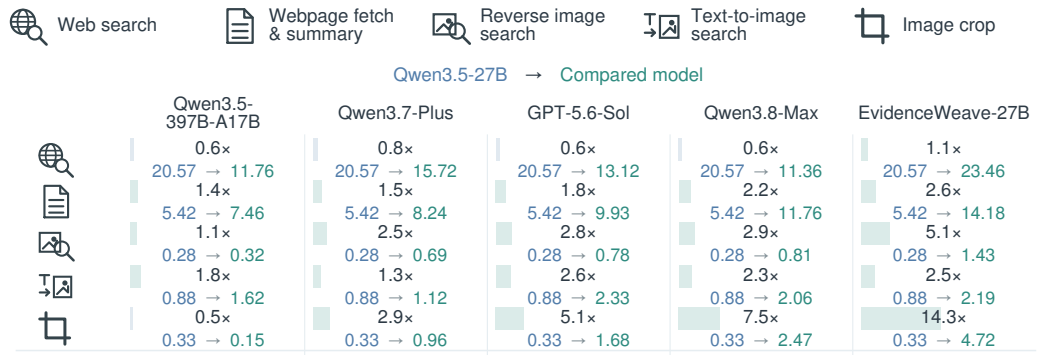


(a) MMBC. (b) MMSearch-Plus.

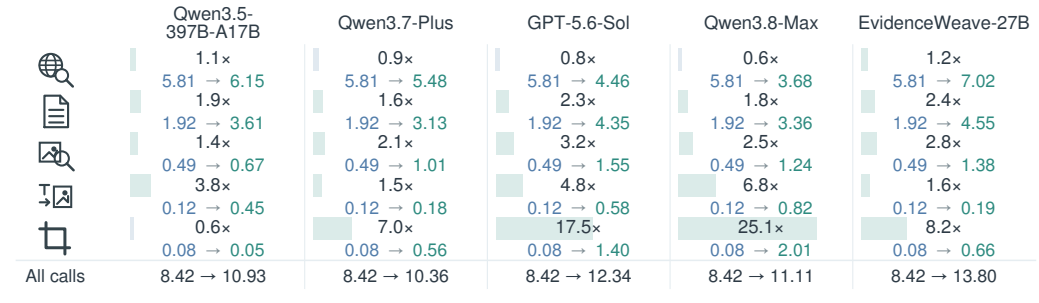
(c) BrowseComp-V^3. (d) VisBrowse.

Figure 10: Onset of visual-tool use before and after post-training on each core benchmark. Each curve gives the percentage of tasks with a visual call by the indicated model-output turn, including tasks without visual calls in the denominator. Markers are observed values, joined by monotone interpolation.

F.3 Reference-model tool use

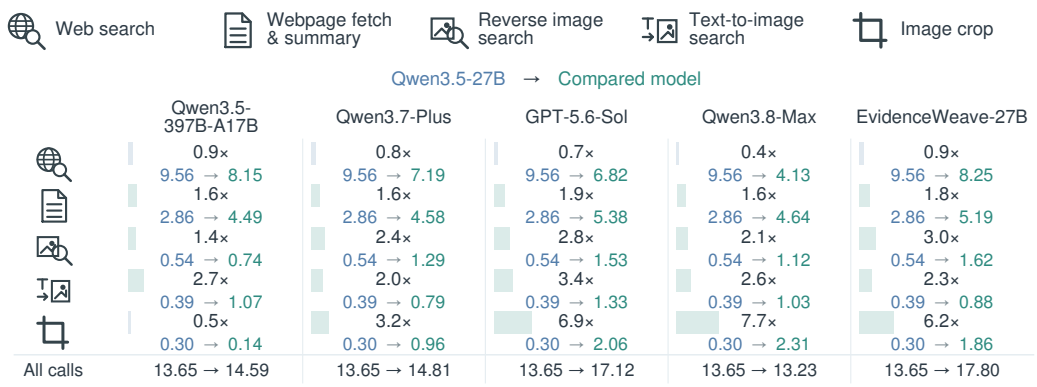


(a) MMBC.

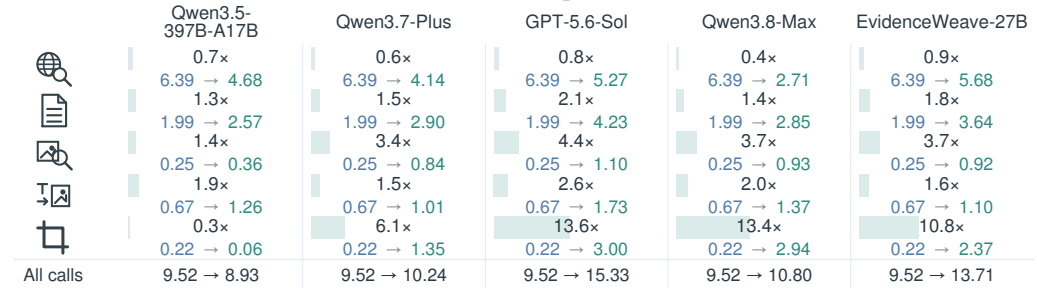


(b) MMSearch-Plus.

Figure 11: Reference-model tool use. Cells show base → compared-model mean calls and their ratios. Bars share a linear ratio scale from 0 to 30.



(c) BrowseComp-V^3.



(d) VisBrowse.

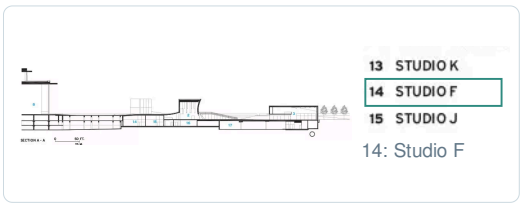
Figure 11: Reference-model tool use (continued): BrowseComp-V^3 and VisBrowse, with the same encoding and scale. All panels use full reported evaluation runs. The final row sums the five tool means.

G Source Examples and Case Analysis

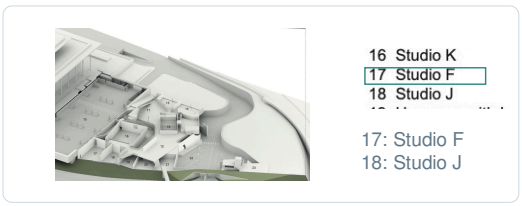
Quoted queries retain their original wording and are labelled by source. The accompanying text examines their information requirements. The final subsection presents constructed illustrations rather than recorded outcomes.

G.1 Queries from our construction records

Figure 12 reproduces the REACH section drawing from Architectural Record (Stephens, 2019) and axonometric drawing from ArchDaily (ArchDaily, 2019), both credited to Steven Holl Architects / BNIM. The legends connect section callout 14 to Studio F and then to axonometric label 17. Queries and reference answers are reproduced from construction records, with labels summarizing their construction-time assessments.

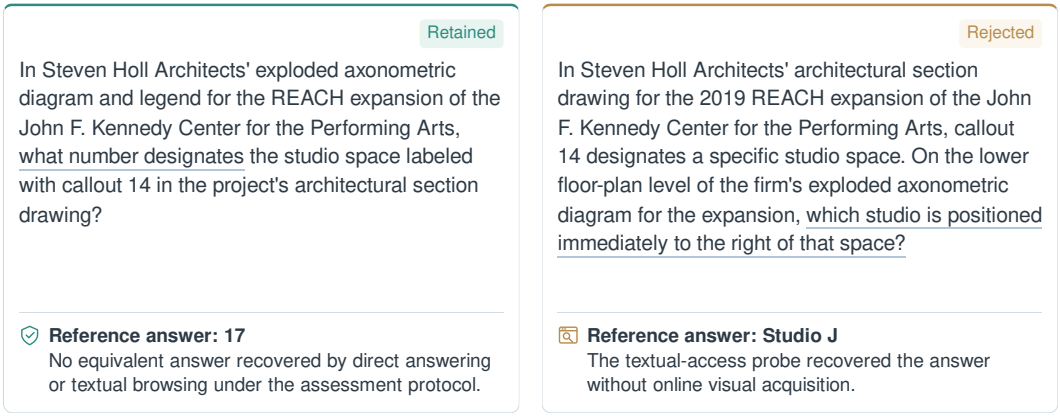


(a) Section drawing and studio identity.



(b) Axonometric drawing and target label.

Figure 12: Architectural evidence connecting section callout 14 to Studio F and then to axonometric label 17. Drawings: Steven Holl Architects / BNIM.



(c) Retained number lookup.

(d) Rejected spatial lookup.

Figure 12: Queries from the same architectural evidence (continued). Verbatim queries and reference answers are reproduced from construction records, with no input media supplied. Labels summarize construction-time assessments.

The retained query keeps Studio F implicit and asks for axonometric label 17, which neither direct answering nor textual browsing recovered during construction. The alternative asks for the studio immediately to its right. Textual browsing recovered its reference answer, Studio J, despite the shared visual evidence. In both queries, the architect name identifies the source without revealing the studio that connects the drawings.

G.2 Queries from public releases

The released queries in Figure 1 illustrate three ways that an answerable task can lose its intended evidence requirements: a visual fact recovered through text, a premise bypassed by another clue, and an intermediate answer disclosed in the query.

G.2.1 Visual evidence recoverable through text

Released query: REDSearcher-MM-SFT.

> Situated within the province of Cáceres and observing the UTC+02:00 offset, a modern municipality serves as the direct successor to an ancient Roman settlement recorded in classical encyclopedias. Identify this town by recognizing the specific defensive fortifications depicted in the provided image.

The REDSearcher-MM-SFT-5K example (Chu et al., 2026) is entry 1646 (zero-based) in the released processed data. Its trajectory identifies the city through image search, but the public clues already specify Cáceres, an ancient Roman predecessor, and defensive fortifications. A textual lookup of official tourism pages identifies Caurium and its Roman walls (Turismo de Extremadura, n.d.), recovering the relevant fact without online visual access.

G.2.2 A premise bypassed by an independent clue

Released query: SearchEyes VisSearch-Bench.

> The entity depicted in the image is the starting point of a research chain. Following its associations across multiple steps — including connections in the area of biographical details — what do you ultimately find? You are looking for the one that Czechoslovakia has a founding connection to.

The SearchEyes VisSearch-Bench example (Jiao et al., 2026), identified by the entity pair Q190973–Q1065, constructs a visual chain ending at the United Nations. A final public clue gives Czechoslovakia's founding relationship to the target organization, providing a direct textual identification route that bypasses the earlier chain.

G.2.3 AN INTERMEDIATE ANSWER DISCLOSED IN THE QUERY

Released query: InterLV-Search Level 3.

> Start from the embroidered sky-sign in a medieval conquest tapestry. Identify the celestial body, then follow its eponym to the European deep-space probe named through an art-historical reference, then from that artistic reference move to the chapel in Padua. Once you reach the city's official heraldic material, what color is the central cross?

The InterLV-Search Level 3 example (ID 6) (Hou et al., 2026) in Figure 1 has a derivation that reaches Padua through Giotto and the Scrovegni Chapel before asking about the city's heraldry. Yet the public wording already states "the chapel in Padua," revealing the city that the preceding tapestry–comet–probe–chapel chain was meant to establish. The remaining lookup can still be difficult, although identifying the city through the earlier chain is no longer required.

G.3 AUTHOR-CONSTRUCTED ILLUSTRATIONS

Figure 2 also includes two constructed explanatory contrasts: the architect lookup and the phrase "listed between K and J." Source credit supplies the former answer, while the latter identifies Studio F from the legend even after its identifying premise is withheld. These constructed contrasts illustrate review criteria rather than measured filtering outcomes.

H RESOURCES AND LICENSES

Table 4 lists the licenses and service terms of the models, datasets, and software used in this work.

Table 4: Licenses and service terms for the models, datasets, and software used in this work.

Resource	Use	License or service terms
Datasets		
MM-BrowseComp (Li et al., 2026)	Evaluation	MIT (dataset card)
MMSearch-Plus (Tao et al., 2026)	Evaluation	CC BY 4.0 (dataset card)
BrowseComp-V^3 (Zhang et al., 2026a)	Evaluation	CC BY 4.0 (dataset)
VisBrowse (Zhang et al., 2026d)	Evaluation	Apache-2.0 (repository)
BrowseComp-VL (Geng et al., 2026)	Evaluation	Apache-2.0 (repository)
MMSearch (Jiang et al., 2025)	Evaluation	MIT (dataset card)
SimpleVQA (Cheng et al., 2025)	Evaluation	Apache-2.0 (dataset card)
LiveVQA benchmark (Fu et al., 2025)	Evaluation	Apache-2.0 (benchmark card)
FVQA (Wang et al., 2018)	Evaluation	Not specified in the release repository
REDSearcher-MM SFT queries (Chu et al., 2026)	Query and teacher analysis	Apache-2.0 (dataset card)
OpenSearch-VL SFT and RL queries (Chen et al., 2026b)	Query and teacher analysis	Apache-2.0 (project)
Vision-DeepResearch public queries (Huang et al., 2026)	Teacher analysis	CC BY-NC-SA 4.0 (dataset cards)
Models		
Qwen3.5-27B	Query assessment and post-training backbone	Apache-2.0 (weights)
Qwen3.5-397B-A17B	Reference agent	Apache-2.0 (weights)
Qwen3.7-Plus	Teacher synthesis and reference agent	Alibaba Cloud Model Studio service agreements
Qwen3.8-Max	Reference agent	Alibaba Cloud Model Studio service agreements
GPT-5.6-Sol (OpenAI, 2026)	Reference tool-use analysis	OpenAI Services Agreement
DeepSeek API	Answer adjudication	DeepSeek Terms of Use
Software and retrieval services		

Table 4: Licenses and service terms (continued).

Resource	Use	License or service terms
Pillow	Image processing and cropping	MIT-CMU
Flask	Local service interfaces	BSD-3-Clause
PyYAML	Pipeline configuration	MIT
Serper	Web and image search	Serper Terms of Service
Jina Reader	Webpage extraction	Jina service terms