VISDOCAGENTBENCH: BENCHMARKING AGENTS FOR VISUALLY RICH DOCUMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Visually rich documents encode relevance through language, layout, structured visual elements, and corpus context, yet retrieval is typically evaluated by one-shot query–page matching. Agentic-search benchmarks usually score downstream question answering or report generation, leaving document ranking under iterative evidence acquisition underexplored. We introduce VISDOCAGENTBENCH, a closed-corpus benchmark comparing static and agentic retrieval under a shared ranked-output contract. It contains 2,375 pages from 100 documents and 120 unique-target queries balanced across direct, one-bridge, and two-bridge evidence structures. Relation-preserving construction yields semantic, relational, and visual queries, followed by full-document review and hard-negative validation. A strong late-interaction visual retriever reaches 97.50% Recall@1 on direct items but 2.50% on two-bridge items, exposing the limits of query–target matching when relevance depends on corpus context. Agents recover much of this loss, but planner choice and retrieval representation remain decisive. Every planner performs better with visual retrieval, whose best R@1 reaches 67.50% versus 37.50% for OCR-text. Ablations identify iterative search and page inspection as consequential capabilities, and providing the complete support context improves ranking on both routes. Trace analysis localizes the remaining losses to target discovery, candidate examination, and evidence-role integration. These findings motivate retrieval agents that combine modality-preserving discovery with evidence-directed verification.

1 INTRODUCTION

Visual retrieval has expanded from matching text to photographs and other natural images toward a common interface for heterogeneous multimodal content. Large-scale vision–language pretraining and dual-encoder objectives made open-vocabulary image matching scalable (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; 2023); instruction-aware and unified representations subsequently broadened retrieval across tasks, intents, and media (Wei et al., 2024; Zhang et al., 2024; Jiang et al., 2025b; Lin et al., 2025; Meng et al., 2025). Visually rich documents fall within this broader interface, but they exhibit a different relevance structure. Whereas natural-image relevance is often expressed through depicted entities, attributes, actions, and scenes, documents are designed information artifacts in which language, two-dimensional layout, tables, diagrams, formulas, and interface elements jointly encode meaning (Faysse et al., 2025; Dong et al., 2025; Osmulski et al., 2025; Shorten et al., 2026; Yan et al., 2026a). Identifying a target unit may therefore require both reading its local text–layout composition and connecting it to definitions, results, or conditions elsewhere in the collection (Ma et al., 2024; Chen et al., 2024; Cho et al., 2024; Wang et al., 2025).

Retrieval is also shifting from one-shot ranking toward agentic search. Agents can move beyond fixed-request candidate scores by reformulating searches, inspecting results, invoking specialized tools, and accumulating evidence across steps (Yao et al., 2023; Nakano et al., 2022; Jin et al., 2025; Jiang et al., 2025a). Work on visually rich documents, however, remains largely understanding-oriented, with retrieval used mainly for fixed-query page ranking (Faysse et al., 2025; Dong et al., 2025; Osmulski et al., 2025; Shorten et al., 2026) or to support document question answering (Mathew et al., 2021; Ma et al., 2024; Ouyang et al., 2025; Wang et al., 2025; 2026b). Many recent agentic visual-search benchmarks support iterative interaction but focus primarily on natural

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.

images (Deng et al., 2026; He et al., 2026; Zhang et al., 2026d). The missing intersection is agentic search over a closed visual-document corpus with direct evaluation of the final document ranking.

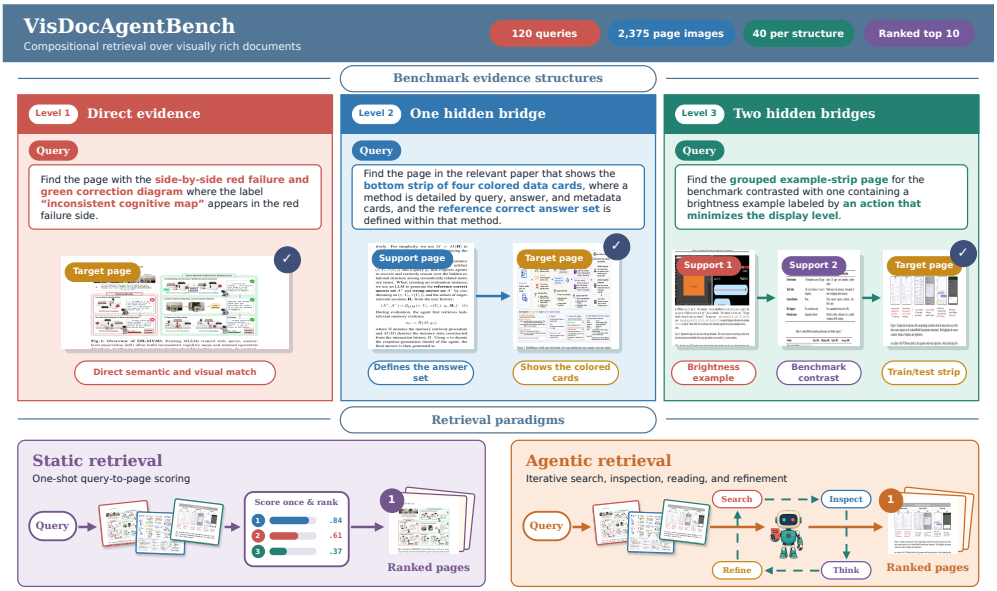


Figure 1: Overview of VISDOCAGENTBENCH. The upper panel depicts direct, one-bridge, and two-bridge evidence structures. The lower panel contrasts one-shot static scoring with iterative search and inspection. Construction annotations and gold labels are hidden from evaluated systems.

We introduce VISDOCAGENTBENCH, a closed-corpus benchmark for ranked visual-document retrieval. Each query combines semantic, relational, and visual conditions and has one target page in the image corpus. Static retrievers and tool-using agents share the same ranked-output contract. The 120 items are balanced across direct, one-bridge, and two-bridge evidence structures, as illustrated in Figure 1.

The benchmark contains 2,375 pages from 100 scientific documents. Its queries undergo relation-preserving construction, full-document review, and hard-negative validation. A late-interaction visual retriever is nearly perfect on direct items but falls from 97.50% to 2.50% Recall@1 on two-bridge items, showing that stronger query–page matching does not resolve distributed evidence. Tool-using agents recover much of this loss, although their wide variation under the same interface shows that tool access alone is insufficient. Every planner performs better with visual retrieval. Ablations identify iterative search and page inspection as consequential capabilities, while complete support context improves L2 and L3 ranking on both routes. The remaining failures concentrate in target discovery, candidate examination, and evidence-role integration. Together, these results call for visual-document representations and evidence-directed search policies to be developed jointly.

Our main contributions are as follows: (1) **Task formulation.** We formulate closed-corpus agentic visual-document retrieval with ranked retrieval as the endpoint and a shared output contract across retrieval paradigms. (2) **Benchmark.** We release VISDOCAGENTBENCH, spanning a 2,375-page document corpus and 120 balanced, unique-target queries grounded in reviewed full-document evidence. (3) **Construction.** We introduce a role-specific, relation-preserving construction pipeline with automatic screening, full-document review, constrained authoring, and hard-negative validation. (4) **Evaluation.** We benchmark static retrievers and tool-using agents across visual and OCR-text routes, using controlled diagnostics and auditable traces to expose representation, search, and evidence-integration bottlenecks.

2 Related Work

2.1 Visual Retrieval over Images and Documents

Large-scale vision–language pretraining supports efficient open-vocabulary image retrieval through dual encoders and related contrastive objectives (Radford et al., 2021; Zhai et al., 2023). Instruction-aware and general-purpose retrievers extend this interface across tasks and media (Wei et al., 2024; Jiang et al., 2025b; Lin et al., 2025; Zhou et al., 2025; Meng et al., 2025; Li et al., 2026a; Hu et al., 2026). Although these representations broaden query expressivity, their standard protocol encodes a request once and ranks candidates independently.

Visual-document retrieval must additionally preserve language, layout, tables, figures, and fine-grained regions. Existing approaches use direct page embeddings, late interaction, query-conditioned representations, region supervision, and staged visual refinement (Faysse et al., 2025; Yu et al., 2025; Cho et al., 2024; Yang et al., 2026; Abdallah et al., 2026; Guan et al., 2026). Complementary work improves contextualization and index efficiency (Yan et al., 2026b;c; Qin et al., 2026; Georgiou, 2026), while document benchmarks compare page-, layout-, multilingual-, and large-corpus retrieval settings (Dong et al., 2025; Osmulski et al., 2025; Chen et al., 2024; Shorten et al., 2026; Yan et al., 2026a). Together, these methods and benchmarks establish the representational challenges of visual-document retrieval, primarily under fixed-query ranking protocols. VisDocAgentBench extends this line by evaluating systems that can iteratively search and inspect the corpus before producing the final visual-unit ranking.

2.2 Agentic Search

Language-model agents interleave reasoning and external actions (Yao et al., 2023). Search agents apply this pattern to query reformulation, iterative retrieval, and evidence accumulation across browsers and search engines (Nakano et al., 2022; Jin et al., 2025; Gou et al., 2025; Wei et al., 2025; Du et al., 2025). Multimodal variants add visual queries and observations (Jiang et al., 2025a; Li et al., 2026b), while recent learned agents target longer-horizon and actively visual search (Zhang et al., 2026b; Chen et al., 2026; Peng et al., 2026; Liu et al., 2026; Zhang et al., 2026d). Their retrieval actions usually support a generated answer, report, or task outcome, leaving the ranking trajectory itself only indirectly measured.

Process-oriented benchmarks increasingly expose intermediate visual-search behavior through checkpoints, cropping, visual browsing, and multimodal evidence traces (Wei et al., 2026; He et al., 2026; Zhang et al., 2026e; Huang et al., 2026). Document-focused agents similarly combine retrieval with selective reading or OCR for question answering and document understanding (Wang et al., 2025; 2026b;a). DeepImageSearch instead evaluates context-aware image retrieval as the endpoint over personal visual histories (Deng et al., 2026). VisDocAgentBench shares the emphasis on auditable interaction but targets ranked retrieval from a common closed corpus of visually rich documents.

2.3 Benchmarks for Visual and Agentic Retrieval

Visual-document benchmarks span image-based question answering and chart reasoning (Mathew et al., 2021; Masry et al., 2022), long-context and multi-document understanding (Ma et al., 2024; Chen et al., 2024), document parsing (Ouyang et al., 2025), and direct page retrieval (Dong et al., 2025; Osmulski et al., 2025; Shorten et al., 2026). Evaluation in agentic settings instead spans browser actions, difficult information discovery, multimodal answers, task completion, and retrieval for agentic coding (Koh et al., 2024; Jiang et al., 2025a; Li et al., 2026b; Wei et al., 2025; Zhang et al., 2026a). VisDocAgentBench connects these lines through a shared ranked-page endpoint and exposes the search, inspection, modality-conversion, and final-ranking trajectory.

Recent multimodal benchmarks also emphasize shortcut control, structured quality assurance, and post-construction verification (Yue et al., 2025; Chen et al., 2025; Wei et al., 2025; Deng et al., 2026). We adapt these principles to compositional retrieval by validating relation paths, reviewing surviving items in full-document context, constraining which clues enter the query, and auditing hard negatives returned by a strong visual retriever. The resulting benchmark separates corpus-grounded evidence structure from the retrieval paradigm used to solve it.

3 Task Formulation

3.1 Closed-Corpus Visual-Document Retrieval

Visually rich documents encode linguistic content, spatial organization, and graphical structure on a shared visual surface. We call a retrievable item a *visual document unit* x_i, which may combine embedded text, layout, tables, diagrams, formulas, screenshots, or interface components.

Let the fixed corpus be $\mathcal{C} = \{x_1, \ldots, x_N\}$. Given a natural-language query q and a unique target $a_q \in \mathcal{C}$, a system returns an ordered list of distinct visual units

$$R_q^K = [x_{(1)}, \ldots, x_{(K)}]. \tag{1}$$

The objective is to rank a_q as highly as possible. Relevance may jointly depend on language, concept relations, and visual-layout properties, while evaluation concerns the target ranking.

3.2 Context-Dependent Compositional Relevance

We distinguish *target-local evidence*, visible on a candidate, from *corpus-context evidence*, located in other units and useful for resolving semantic or relational conditions in q. A locally plausible unit can remain incorrect when it fails the complete corpus-grounded relation. Context elsewhere in the corpus can disambiguate such candidates. Appendix A.1 formalizes the latent support set and contextual relevance relation.

3.3 Static and Agentic Retrieval

A static retriever independently scores query–unit pairs and produces one fixed ordering. An agent instead gathers evidence through a budgeted action–observation process, conditioning later searches and its final ranking on the accumulated history. Both are evaluated under the same ranked-output contract. Appendix A.1 gives the complete static and agentic formulations.

4 Benchmark Construction and Quality Control

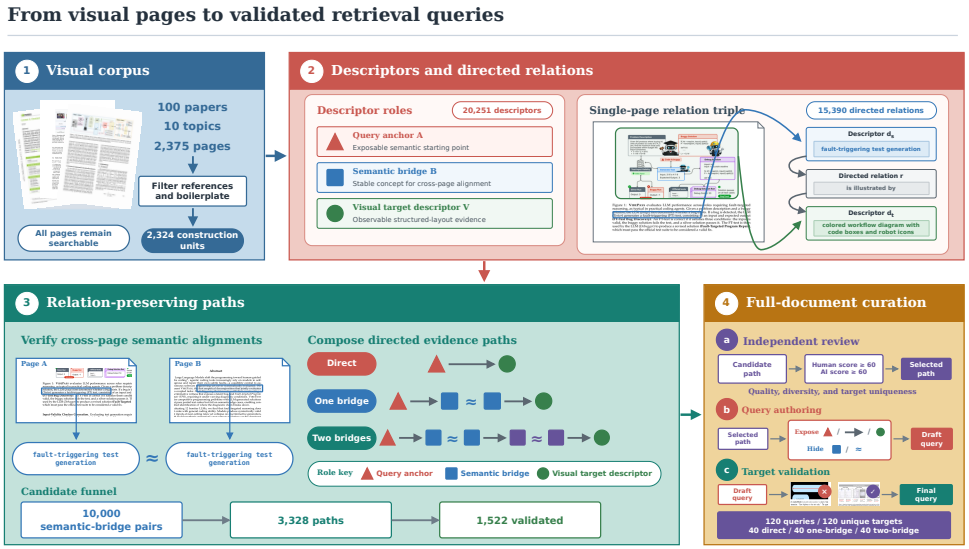


Figure 2: Construction and quality-control pipeline. Role-specific descriptors, directed relations, semantic alignments, and evidence paths supervise benchmark construction but are hidden during evaluation. The final benchmark is formed only after independent full-document scoring, constrained query authoring, and hard-negative target validation.

Figure 2 summarizes the pipeline from visual-unit preparation through relation-preserving path construction and full-document quality control.

4.1 Corpus Collection and Visual-Unit Preparation

We use scientific papers because their pages densely combine text, layout, tables, and diagrams while distributing recurring concepts across units. They preserve traceable provenance and support deterministic rendering plus full-context verification, as established by prior multimodal retrieval and document-reasoning benchmarks (Chen et al., 2024; Cho et al., 2024; Dong et al., 2025; Shorten et al., 2026).

We sample 100 arXiv papers published in 2026, taking 10 papers from each of 10 relatively distinct topics such as agent systems and world models. In VisDocAgentBench, each visual document unit is one rendered full page, giving 2,375 images with document and page provenance. We use 2,324 non-reference, content-bearing pages for query construction. All 2,375 images form the retrieval corpus. Appendix B.1 details topic and document selection and lists the complete source inventory in Table 6. Appendix G describes source-license verification, redistribution rules, and privacy safeguards.

4.2 Role-Specific Descriptors and Directed Relations

For each construction unit, a VLM extracts query anchors, semantic bridges, visual target descriptors, and directed intra-unit relations. Query anchors provide semantic starting points without identifying the target; semantic bridges denote stable concepts that can align across units; visual target descriptors capture observable layout or structured evidence without relying on captions, identifiers, or answer-revealing text. Role and quality validation screens these descriptors and relations before path construction. Appendix B.2 gives the full role definitions and filtering details.

4.3 Relation-Preserving Path Construction

We embed validated semantic bridges with Qwen3-Embedding-8B (Zhang et al., 2025) and retrieve candidate pairs within each topic. A VLM verifies their conceptual equivalence or contextual compatibility before they are used to connect units.

Let A denote a query anchor, B_i a semantic bridge, V a visual target descriptor, r_i a directed intra-unit relation, and $\approx$ a verified cross-unit semantic alignment. We construct path templates:

$$\begin{cases} \text{L1}: & A \xrightarrow{r_1} V, \\ \text{L2}: & A \xrightarrow{r_1} B_1 \approx B_2 \xrightarrow{r_2} V, \\ \text{L3}: & A \xrightarrow{r_1} B_1 \approx B_2 \xrightarrow{r_2} B_3 \approx B_4 \xrightarrow{r_3} V. \end{cases} \tag{2}$$

L1 contains a direct relation from the query anchor to the visual target descriptor; L2 and L3 introduce one and two latent semantic alignments. Appendix A.3 gives the complete level-specific path and query constraints.

Automatic path validation checks semantic coherence, evidence necessity for L2 and L3, and visual necessity, retaining 1,522 paths for full-document review. Appendix B.3 gives the extraction, alignment, and validation settings.

4.4 Full-Document Path Review and Selection

Three authors jointly assign a human score, and an AI reviewer independently assigns a second score, to every surviving path in full-document context, assessing semantic validity, evidence necessity, bridge specificity, visual grounding, query rewritability, and diversity. Eligibility requires both scores to be at least 60 on a 0–100 scale; paths are ranked by their mean. We select 40 paths per level, prioritizing quality and then diversity, with target deduplication and replacement yielding 120 distinct targets. Appendices C.2 and C.3 give the full review and selection protocols.

4.5 Final Query Authoring and Target Validation

Final queries express the query anchor, every directed relation, and an abstracted visual target descriptor, while hiding semantic bridges and their alignments behind generic references. Each query is checked against all pages of its involved source documents and ten full-corpus candidates retrieved

by Qwen3-VL-Embedding-8B (Li et al., 2026a). Every non-gold candidate is inspected. Valid alternatives trigger minimal disambiguation; an item is replaced when uniqueness cannot be restored naturally. The resulting query is rechecked. Appendices C.4 and C.5 give the full protocols. Table 7 in Appendix B.4 summarizes the corpus, construction funnel, and final benchmark composition.

5 Experimental Setup

5.1 Evaluation Protocol and Metrics

The evaluation corpus contains $N = 2{,}375$ units. Each query has one gold target, and systems are required to return $K = 10$ distinct opaque identifiers. We report Recall@k for $k \in \{1, 3, 5, 10\}$ and MRR@10, with R@1 as the primary measure and missing or invalid rankings scored as zero. Metric definitions and inference-time visibility rules appear in Appendices D.1 and A.4, respectively.

5.2 Baselines and Implementation Details

Static retrieval. Our single-vector baselines encode page images with Qwen3-VL-Embedding-8B (Li et al., 2026a) or page OCR with Qwen3-Embedding-8B (Zhang et al., 2025), ranking all pages by cosine similarity. OCR is produced once with PaddleOCR-VL-1.6 and PP-DocLayoutV3 (Zhang et al., 2026c). We add BM25 (Robertson & Zaragoza, 2009) over the same OCR records and reciprocal-rank fusion of BM25 with the dense OCR ranking (Cormack et al., 2009). Nemotron-ColEmbed-VL-8B-V2 (de Souza P. Moreira et al., 2026) provides a stronger visual-document baseline through token-level late interaction. Retrieval parameters follow the published defaults and are not tuned on VisDocAgentBench. Appendix D.2 gives the index and scoring details.

Tool-using retrieval. Both agent routes reuse the matching single-vector index but allow the planner to issue multiple searches. The visual route supports visual search and on-demand OCR; the OCR-text route supports text search and cached page OCR. Both expose full-page inspection, batched regional cropping, and ranked submission, as summarized in Table 9 in Appendix D.3.

Planner models and interaction budget. Table 1 lists the closed- and open-weight planners. Every agent receives up to 12 interaction steps, followed when needed by a ranking-only finalization call, under the shared top-10 contract. For GPT-5.6-sol, we ablate iterative search, page inspection, route-specific text access, and regional cropping. Appendices D.4 and D.5 give full configurations; resource accounting appears in Appendices D.6 and H.

Controlled evidence intervention. For L2 and L3, we rerun GPT-5.6-sol after presenting every annotated support page as an unlabeled initial page observation. The query, corpus, tools, and 12-step interaction budget remain fixed, yielding a paired comparison with the standard agent run in Section 6.3.

6 Results and Analysis

6.1 Main Results

Table 1 compares static retrievers and tool-using agents under the same ranked-output contract. The visual and OCR-text routes operate over the same corpus. Complete level-wise results appear in Table 12 of Appendix E.1; Table 13 in Appendix E.2 reports the GPT-5.6-sol topic breakdown.

Stronger static matching exposes the evidence-path gap. Nemotron-ColEmbed-VL-8B-V2 is the strongest static retriever, reaching 40.00% visual R@1 overall. Table 12 in Appendix E.1 reveals a sharper level-wise contrast: R@1 is 97.50% on direct items and 2.50% on two-bridge items. A late-interaction retriever can therefore solve nearly every direct visual match while still failing when the query–target relation depends on evidence elsewhere in the corpus. OCR-text retrieval remains weaker under dense, sparse, and fused rankings, with BM25 providing its strongest static R@1 at 6.67%.

Table 1: Overall retrieval results on VISDOCAGENTBENCH (%). Metrics use all 120 queries, with invalid completed episodes scored as zero. Bold marks the best value in each route and metric.

Retriever / Planner	Visual					OCR-Text				
	R@1	R@3	R@5	R@10	MRR@10	R@1	R@3	R@5	R@10	MRR@10
Static Retrievers										
Qwen3 Embedding	20.83	30.00	34.17	45.00	27.34	1.67	5.83	13.33	19.17	6.01
BM25	–	–	–	–	–	6.67	15.00	19.17	30.00	12.46
BM25 + dense RRF	–	–	–	–	–	5.00	14.17	19.17	26.67	11.41
Nemotron ColEmbed	40.00	52.50	65.00	70.00	48.86	–	–	–	–	–
Tool-Using Agents: Closed-Source Planners										
GPT-5.5	60.00	61.67	64.17	69.17	61.87	26.67	27.50	27.50	28.33	27.19
GPT-5.6-luna	43.33	47.50	50.83	57.50	46.75	12.50	14.17	18.33	21.67	14.77
GPT-5.6-terra	54.17	57.50	58.33	64.17	56.65	26.67	28.33	28.33	33.33	27.92
GPT-5.6-sol	61.67	65.00	65.83	68.33	63.58	36.67	38.33	40.83	44.17	38.40
Claude Opus 4.8	47.50	54.17	59.17	68.33	52.39	24.17	27.50	30.00	40.83	27.38
Claude Fable 5	62.50	67.50	**71.67**	**80.00**	66.69	35.00	**41.67**	43.33	**52.50**	39.34
Claude Sonnet 5	30.00	40.83	46.67	55.83	37.33	17.50	20.00	22.50	23.33	19.12
Claude Opus 5	**67.50**	**70.83**	**71.67**	75.00	**69.43**	**37.50**	**41.67**	**45.00**	49.17	**40.57**
Tool-Using Agents: Open-Weight Planner										
Qwen3.5 (thinking)	27.50	33.33	35.00	40.00	31.20	8.33	9.17	10.00	15.83	9.70
Qwen3.5 (no thinking)	19.17	23.33	30.00	38.33	23.55	1.67	4.17	7.50	10.00	3.83

Interaction helps, but planner choice remains decisive. Agents using the same tools and indices vary widely: across the evaluated planners, visual R@1 ranges from 19.17% to 67.50%, while OCR-text R@1 ranges from 1.67% to 37.50%. Claude Opus 5 achieves the strongest R@1 on both routes, while Claude Fable 5 retrieves the most targets within the top 10. The different leaders show that finding the target and placing it first require distinct planner capabilities. Iterative search and inspection can therefore recover failures of static ranking, but only when the planner turns its observations into better searches and final rankings.

Visual retrieval is consistently stronger than OCR-text retrieval. Every planner performs better on the visual route. The R@1 difference ranges from 12.50 points for Claude Sonnet 5 (30.00% visual versus 17.50% OCR-text) to 33.33 points for GPT-5.5 (60.00% versus 26.67%). This gap indicates that full-page visual representations preserve layout and structured visual evidence that an OCR-only index may discard during candidate discovery. The gap arises during global search, while targeted text extraction remains useful for verifying selected pages.

Longer evidence paths remain challenging under interaction. As detailed in Table 12 of Appendix E.1, agentic interaction raises performance but does not eliminate the path-length effect. GPT-5.6-sol R@1 declines from 85.00% to 40.00% on the visual route and from 42.50% to 27.50% on the OCR-text route, while Claude Opus 5 declines from 92.50% to 47.50% on the visual route, indicating that longer evidence paths remain challenging even under iterative search and inspection. Table 15 in Appendix E.4 provides the complementary GPT-5.6-sol breakdown by single-unit, same-document, and cross-document paths.

6.2 ABLATION STUDIES

We ablate iterative search, page inspection, text access, and regional cropping for GPT-5.6-sol while keeping the remaining settings fixed. The no-iterative-search setting applies VLM reranking to a fixed candidate set containing the 30 highest-ranked pages from the route-specific retriever. This size slightly exceeds the full agent's observed mean page exposure on both routes: valid episodes examine 25.22 pages with visual search and 28.22 with OCR-text search (Table 11 in Appendix D.6). It therefore provides a comparable verification budget across routes. No inspection also removes dependent cropping. Table 2 gives the overall results, and Table 14 in Appendix E.3 reports every level.

Table 2: GPT-5.6-sol ablations on all 120 queries (%). †No-inspection also removes dependent regional cropping.

Configuration	Visual					OCR-Text				
	R@1	R@3	R@5	R@10	MRR@10	R@1	R@3	R@5	R@10	MRR@10
Full agent	61.67	65.00	65.83	68.33	63.58	36.67	38.33	40.83	44.17	38.40
No iterative search	53.33	59.17	59.17	60.83	56.35	27.50	29.17	30.83	30.83	28.61
No regional crop	64.17	69.17	71.67	74.17	67.34	34.17	37.50	38.33	40.00	35.97
No OCR / No page OCR	60.00	62.50	66.67	69.17	62.40	33.33	35.00	35.83	39.17	34.63
No page inspection†	45.83	50.83	57.50	63.33	50.03	15.00	19.17	23.33	27.50	18.28

Iterative search improves both retrieval routes. Removing iterative search lowers R@1 by 8.34 points on the visual route and 9.17 points on OCR-text. The ablation uses 11 VLM calls and 30 page observations (Table 11 in Appendix D.6), yet its frozen candidate set prevents observations from guiding subsequent retrieval. Repeated candidate inspection therefore does not substitute for search conditioned on accumulated evidence.

Direct page inspection provides the largest gain. Among all ablations, removing page inspection causes the largest R@1 losses, with drops of 15.84 and 21.67 points on the visual and OCR-text routes, respectively. These drops identify visual inspection of retrieved candidates as the key verification step for determining whether semantic clues and visual structure co-occur on the same page.

Text access exposes a path-dependent exploration–verification trade-off. Removing route-specific text access lowers overall R@1 by 1.67 points on the visual route and 3.34 points on the OCR-text route. The level-wise results in Table 14 of Appendix E.3 show that this aggregate benefit is concentrated on shorter paths: without text access, R@1 drops by 7.50 and 10.00 points on visual L1 and L2, and by 2.50 and 10.00 points on OCR-text L1 and L2. The direction reverses on L3, where R@1 rises by 12.50 and 2.50 points, respectively. Trace-level tool usage shows that removing text access shifts actions toward additional searches and full-page inspection. Text therefore supports local verification, but current planners do not reliably balance it against the broader corpus exploration required by longer evidence paths.

Regional cropping has route-dependent effects. Removing cropping raises visual R@1 by 2.50 points and lowers OCR-text R@1 by 2.50 points, with mixed shifts across levels and metrics. Visually rich pages often require both local details, such as table cells or interface labels, and their surrounding headers, rows, or diagrams. A crop can clarify the former while weakening the latter. Current planners therefore do not reliably determine when local detail outweighs the loss of page context. These results motivate the development of multimodal foundation models that can integrate magnified evidence with its surrounding layout, thereby invoking cropping when it is most informative.

6.3 Evidence Intervention and Trace Diagnostics

We use GPT-5.6-sol to test how complete support context changes retrieval and to locate the remaining bottlenecks. Table 3 compares the standard agent with the support-provided condition, whose experimental setup is described in Section 5.2. Figure 3 tracks target discovery, examination, and final ordering. Appendix F.1 defines the target-progression events used in Figure 3 and reports their level-wise breakdown in Figure 4.

Complete support context strengthens evidence-conditioned retrieval. Providing complete support context improves every reported metric across both retrieval routes and both path depths in Table 3. MRR@10 rises by 9.28–12.50 points across the four comparisons, showing that the benefit extends across retrieval representations and query levels. The largest effect appears on visual L3 coverage, where R@10 improves by 22.50 points. Complete support context therefore helps the planner connect evidence across longer paths, bringing more targets into the final ranking and improving their ordering.

Table 3: GPT-5.6-sol retrieval with standard and support-provided initialization on L2 and L3 queries (%).

Level	Condition	Visual					OCR-Text				
		R@1	R@3	R@5	R@10	MRR@10	R@1	R@3	R@5	R@10	MRR@10
L2	Standard	60.00	62.50	62.50	62.50	61.25	40.00	40.00	40.00	40.00	40.00
	Support provided	72.50	72.50	72.50	75.00	72.86	52.50	52.50	52.50	52.50	52.50
L3	Standard	40.00	45.00	47.50	52.50	43.36	27.50	32.50	40.00	45.00	32.12
	Support provided	45.00	62.50	65.00	75.00	55.19	37.50	45.00	47.50	52.50	41.40

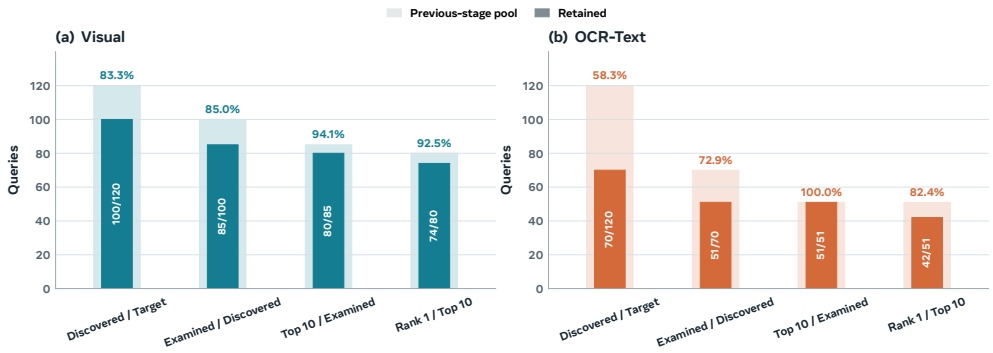


Figure 3: Gold-target progression through the full GPT-5.6-sol traces for (a) Visual and (b) OCR-Text retrieval. Pale bars show the previous-stage pool, while solid bars show the retained subset. Each percentage reports retention from the corresponding pool.

Target progression from discovery to ranking. Figure 3 localizes the visual–OCR-text gap mainly to target discovery and examination. Visual search discovers 83.3% of targets, compared with 58.3% for OCR-text, and agents examine 85.0% of visual discoveries versus 72.9% of OCR-text discoveries. Since both routes share the planner and visual-inspection tools, this divergence indicates that visual search provides a stronger corpus-wide discovery interface. Once examined, at least 94.1% of targets enter the final top 10 and at least 82.4% of those retained are ranked first. Final reranking is therefore comparatively reliable after target inspection.

Qualitative trace examples. Appendix F.2 contrasts a successful and a failed L3 visual-route episode in Figure 5 and Tables 16–17. The success uses an examined support page to reformulate the search and surface the target. The failed trace remains anchored to the paper containing the initial brightness example and never follows the benchmark contrast into the target document. The second support and target surface only in its final search, remain unexamined, and the target is omitted from the submitted top 10. Together, the examples show how evidence acquisition can redirect retrieval or fail when the planner does not preserve a cross-document relation.

7 Conclusion

We introduced VisDocAgentBench, a closed-corpus benchmark for visually rich document retrieval under static and agentic paradigms, constructed from relation-preserving evidence paths and full-document audits. A strong late-interaction retriever is nearly perfect on direct items yet rarely solves two-bridge queries. Controlled ablations show that iterative search improves retrieval, but wide planner variation shows that tool access alone does not yield reliable retrieval. Visual search is more effective for corpus-wide discovery, while page inspection and targeted text access provide complementary verification. Providing complete support context further improves ranking on the same queries. Trace analysis localizes remaining failures to target discovery, candidate examination, and evidence-role integration. These findings motivate retrieval agents that combine modality-preserving representations, evidence-directed policies, and role-aware state.

AI USE STATEMENT

We used generative AI tools to assist with manuscript refinement, literature discovery, code development, and synthetic dataset generation. For writing, the tools suggested revisions to improve clarity and readability, which the human authors reviewed and decided whether to adopt. AI-assisted searches helped broaden the coverage of related work and reduce the risk of overlooking relevant studies. The authors evaluated the retrieved literature, selected the references, and synthesized the related-work discussion. For implementation, AI tools assisted with code scaffolding and debugging, with the resulting code reviewed by the authors. Our use of generative AI for synthetic dataset generation is detailed in Section 4 and Appendices B and C. The authors take full responsibility for the final content of this paper, including all AI-assisted material.

REFERENCES

Abdelrahman Abdallah, Mahmoud Abdalla, Mohammed Ali, and Adam Jatowt. Argus-Retriever: Vision-LLM late-interaction retrieval with region-aware query-conditioned MoE for visual document retrieval, 2026. URL https://arxiv.org/abs/2606.04300.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. MEGA-Bench: Scaling multimodal evaluation to over 500 real-world tasks. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/31110.

Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, and Mohamed Elhoseiny. Document haystacks: Vision-language reasoning over piles of 1000+ documents, 2024. URL https://arxiv.org/abs/2411.16740.

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, 2026. URL https://arxiv.org/abs/2605.05185.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3DocRAG: Multi-modal retrieval is what you need for multi-page multi-document understanding, 2024. URL https://arxiv.org/abs/2411.04952.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, Bo Liu, and Even Oldridge. Nemotron ColEmbed V2: Top-performing late interaction embedding models for visual document retrieval, 2026. URL https://arxiv.org/abs/2602.03992.

Chenlong Deng, Mengjie Deng, Junjie Wu, Dun Zeng, Teng Wang, Qingsong Xie, Jiadeng Huang, Shengjie Ma, Changwang Zhang, Zhaoxiang Wang, Jun Wang, Yutao Zhu, and Zhicheng Dou. Deepimagesearch: Benchmarking multimodal agents for context-aware image retrieval in visual histories, 2026. URL https://arxiv.org/abs/2602.10809.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents, 2025. URL https://arxiv.org/abs/2501.08828.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents, 2025. URL https://arxiv.org/abs/2506.11763.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.01449.

Athos Georgiou. Hydra: Unifying document retrieval and generation in a single vision-language model, 2026. URL https://arxiv.org/abs/2603.28554.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2Web 2: Evaluating agentic search with agent-as-a-judge, 2025. URL https://arxiv.org/abs/2506.21506.

Tongkun Guan, Haocheng Wang, Wei Shen, and Xiaokang Yang. LightSTAR: Efficient visual document retrieval via lightweight selection with vision-adaptive refinement, 2026. URL https://arxiv.org/abs/2606.23539.

Hang He, Chuhuai Yue, Chengqi Dong, Chengcheng Wan, Ting Su, Haiying Sun, Jiajun Chai, Xiaohan Wang, and Guojun Yin. VistaHop: Benchmarking long-horizon visual DeepSearch, 2026. URL https://arxiv.org/abs/2606.03273.

Lexiang Hu, Youze Xue, Dian Li, Gang Liu, and Zhouchen Lin. Adaptive global and fine-grained perceptual fusion for MLLM embeddings compatible with hard negative amplification, 2026. URL https://arxiv.org/abs/2602.05729.

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, 2026. URL https://arxiv.org/abs/2601.12346.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Unveiling the potential of large models as multi-modal search engines. In *International Conference on Learning Representations*, 2025a. URL https://iclr.cc/virtual/2025/poster/30134.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2410.05160.

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, 2025. URL https://arxiv.org/abs/2503.09516.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.50. URL https://aclanthology.org/2024.acl-long.50/.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026a. URL https://arxiv.org/abs/2601.04720.

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, 2026b. URL https://arxiv.org/abs/2508.13186v2.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.02571.

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/2604.14029v2.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2407.01523.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images, 2021. URL https://arxiv.org/abs/2007.00398.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents, 2025. URL https://arxiv.org/abs/2507.04590.

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, 2022. URL https://arxiv.org/abs/2112.09332.

Radek Osmulski, Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. MIRACL-VISION: A large, multilingual, visual document retrieval benchmark, 2025. URL https://arxiv.org/abs/2505.11651.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ouyang_OmniDocBench_Benchmarking_Diverse_PDF_Document_Parsing_with_Comprehensive_Annotations_CVPR_2025_paper.html.

Xiangyu Peng, Can Qin, An Yan, Xinyi Yang, Zeyuan Chen, Ran Xu, and Chien-Sheng Wu. MTA-Agent: An open recipe for multimodal deep search agents, 2026. URL https://arxiv.org/abs/2604.06376.

Hanxiang Qin, Alexander Martin, Rohan Jha, Chunsheng Zuo, Reno Kriz, and Benjamin Van Durme. Multi-vector index compression in any modality, 2026. URL https://arxiv.org/abs/2602.21202.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Connor Shorten, Augustas Skaburskas, Daniel M. Jones, Charles Pierse, Roberto Esposito, John Trengrove, Etienne Dilocker, and Bob van Luijt. IRPAPERS: A visual document benchmark for scientific retrieval and question answering, 2026. URL https://arxiv.org/abs/2602.17687.

Jiamian Wang, Ruiyi Zhang, Tong Yu, Jing Shi, Samyadeep Basu, Rajiv Jain, Zhiqiang Tao, and Tong Sun. DocArena: Turning raw documents into controllable training environments for document search agents, 2026a. URL https://arxiv.org/abs/2606.26122.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents, 2025. URL https://arxiv.org/abs/2502.18017.

Zhengren Wang, Dongsheng Ma, Huaping Zhong, Jiayu Li, Wentao Zhang, Bin Wang, and Conghui He. AgenticOCR: Parsing only what you need for efficient retrieval-augmented generation, 2026b. URL https://arxiv.org/abs/2602.24134.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404, 2024. doi: 10.1007/978-3-031-73021-4_23. URL https://arxiv.org/abs/2311.17136.

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, 2025. URL https://arxiv.org/abs/2504.12516.

Qianshan Wei, Yishan Yang, Siyi Wang, Jinglin Chen, Binyu Wang, Jiaming Wang, Shuang Chen, Zechen Li, Yang Shi, Yuqi Tang, Weining Wang, Yi Yu, Chaoyou Fu, Qi Li, and Yi-Fan Zhang. Agentic-MME: What agentic capability really brings to multimodal intelligence?, 2026. URL https://arxiv.org/abs/2604.03016.

Yibo Yan, Jiahao Huo, Guanbo Feng, Mingdong Ou, Yi Cao, Xin Zou, Shuliang Liu, Yuanhuiyi Lyu, Yu Huang, Jungang Li, Kening Zheng, Xu Zheng, Philip S. Yu, James Kwok, and Xuming Hu. Unlocking multimodal document intelligence: From current triumphs to future frontiers of visual document retrieval, 2026a. URL https://arxiv.org/abs/2602.19961.

Yibo Yan, Mingdong Ou, Yi Cao, Jiahao Huo, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. Visual late chunking: An empirical study of contextual chunking for efficient visual document retrieval, 2026b. URL https://arxiv.org/abs/2604.10167.

Yibo Yan, Mingdong Ou, Yi Cao, Xin Zou, Jiahao Huo, Shuliang Liu, James Kwok, and Xuming Hu. Sculpting the vector space: Towards efficient multi-vector visual document retrieval via prune-then-merge framework, 2026c. URL https://arxiv.org/abs/2602.19549.

Hao Yang, Yifan Ji, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Zulong Chen, Shuo Wang, Yu Gu, and Ge Yu. ReAlign: Optimizing the visual document retriever with reasoning-guided fine-grained alignment, 2026. URL https://arxiv.org/abs/2604.07419.

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*, 2023. URL https://arxiv.org/abs/2210.03629.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents, 2025. URL https://arxiv.org/abs/2410.10594.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_LiT_Zero-Shot_Transfer_With_Locked-Image_Text_Tuning_CVPR_2022_paper.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Fuwei Zhang, Yanzhao Zhang, Mingxin Li, Dingkun Long, Lexiang Hu, Pengjun Xie, Zhao Zhang, and Fuzhen Zhuang. CORE-Bench: A comprehensive benchmark for code retrieval in the era of agentic coding, 2026a. URL https://arxiv.org/abs/2606.11864.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.19651.

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, Tingquan Gao, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training, 2026c. URL https://arxiv.org/abs/2606.03264.

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, 2026d. 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, 2026e. URL https://arxiv.org/abs/2603.16289.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025. doi: 10.18653/v1/2025.acl-long.935. URL https://aclanthology.org/2025.acl-long.935/.

A BENCHMARK SPECIFICATION

A.1 FORMAL RETRIEVAL PROTOCOLS

For analysis, let $E_q \subseteq \mathcal{C} \setminus \{a_q\}$ denote a possibly empty latent support set such that

$$\mathrm{Rel}(q, a_q \mid E_q) = 1, \tag{3}$$

where Rel indicates whether the visual unit satisfies the query under the contextual evidence. Target-local evidence can establish relevance when $E_q = \varnothing$. Otherwise, distributed context can contribute to the judgment. Appendix A.2 specifies how a_q and E_q are instantiated in the benchmark and how target validity is adjudicated.

Static and agentic retrieval. A static retriever independently scores each query–unit pair with a function $s(q, x_i)$ and produces one fixed ordering:

$$R_{\mathrm{static}}(q) = \mathrm{argsort}_i \; s(q, x_i). \tag{4}$$

Evaluation uses the first K positions of this ordering.

Agentic retrieval is a budgeted action–observation process. Let $\mathcal{U}_{\mathrm{tool}}$ denote the available evidence-gathering actions, $\mathcal{I}$ the corpus interface, and $h_0 = (q)$ the initial interaction history. For some $T \le B$, the policy π selects an evidence-gathering action at each step $t \in \{1, \dots, T\}$, receives an observation o_t, and appends both to its history:

$$\begin{cases} u_t = \pi(h_{t-1}) \in \mathcal{U}_{\mathrm{tool}}, \\ o_t = \mathcal{I}(u_t, \mathcal{C}), \\ h_t = h_{t-1} \oplus (u_t, o_t), \end{cases} \tag{5}$$

where $\oplus$ denotes sequence concatenation. After the interaction ends, an output policy maps the accumulated history to a ranking:

$$R^K_{\mathrm{agent}}(q) = \pi_{\mathrm{out}}(h_T), \qquad T \le B. \tag{6}$$

Embedding search can serve as an agent tool. Agentic retrieval gains the additional capacity to condition later searches and the final ordering on corpus evidence observed during the episode.

A.2 TASK INSTANTIATION

In VISDOCAGENTBENCH, a visual document unit x_i corresponds to one rendered page and is addressed through an opaque handle during evaluation. For a query q, a_q is its single adjudicated target page. The reviewed relation-preserving path determines which non-target pages instantiate E_q and records how their evidence connects the query to a_q. Thus, E_q is empty for direct items and contains the intermediate evidence pages for bridged items.

Under this instantiation, $\mathrm{Rel}(q, a_q \mid E_q) = 1$ means that a_q satisfies the query's complete semantic, relational, and visual conditions when the recorded corpus context is considered. Plausible alternatives are checked against the complete request through the hard-negative and false-negative validation protocol in Appendix C.5.

A.3 EVIDENCE-PATH STRUCTURE AND QUERY CONSTRAINTS

For construction, A denotes a query anchor, B_i a semantic bridge, V a visual target descriptor, r_i a directed relation grounded within one visual unit, and $\approx$ a verified semantic alignment between semantic bridges from different units. The three templates are

$$\begin{cases} \mathrm{L1}: & A \xrightarrow{r_1} V, \\ \mathrm{L2}: & A \xrightarrow{r_1} B_1 \approx B_2 \xrightarrow{r_2} V, \\ \mathrm{L3}: & A \xrightarrow{r_1} B_1 \approx B_2 \xrightarrow{r_2} B_3 \approx B_4 \xrightarrow{r_3} V. \end{cases} \tag{7}$$

A semantic alignment may connect two units in one source document or, under the stricter named-entity rule described in Appendix B.3, two documents. The number of relation-bearing evidence units defines the level and its path template.

Table 4: Query exposure and hiding policy. Generic referents preserve the directed relation skeleton while hiding the semantic bridge identities used to construct the item.

Level	Required exposure	Hidden content	Example
L1	A, r_1, V	identifiers and captions	"Find the unit where an error flag is shown as a red row in a comparison matrix."
L2	A, r_1, r_2, V	$B_1, B_2, \approx$	"Find the unit where a policy instantiates a mechanism that is summarized by a stacked latency chart."
L3	A, r_1, r_2, r_3, V	$B_1, \ldots, B_4$ and both $\approx$ links	"Find the unit where a condition produces an intermediate setting, whose evaluation yields a result displayed in a grouped plot."

Table 4 states the query contract. All directed relations are expressed in natural language, but semantic bridge identities and semantic alignments are replaced by generic referents. Across all levels, the query must request a visual unit while withholding paper titles, page numbers, figure or table numbers, exact captions, target identifiers, and construction terminology. The final visual phrase must support verification while preserving the necessity of the semantic constraints.

A.4 Annotation Roles and Inference Boundary

Table 5: Annotation roles and their inference-time visibility.

Role	Annotation content	Visibility and use
Agent-visible input	Final natural-language query; corpus pages exposed through opaque handles	The planner sees the query text and handles returned by tools. Handles are randomly assigned with a frozen seed and contain no document or page-order signal.
Evaluation metadata	Evidence-path level, topic, path scope, and review outcomes	Available to evaluators for stratification, audit, and data analysis only.
Target and construction annotations	Adjudicated target page; role-specific descriptors and directed relations; semantic alignments; evidence-path pages; query-exposure constraints	Supports scoring and construction audits while remaining strictly withheld from the agent.

Table 5 separates agent-visible input from evaluation metadata and construction annotations. A standard episode supplies only the final query and an opaque handle space over the corpus images. Evidence-path level and topic are reserved for post-prediction stratification and omitted from the planner prompt.

B Corpus and Construction Details

B.1 Topic and Document Selection

We queried the arXiv API in the first half of 2026 with two or three explicit Boolean queries for each of ten topics. The search returned 1,722 topic-query records before within-topic deduplication. We retained papers published during this period and scored title, abstract, and category metadata using topic-specific required, preferred, and excluded phrases. We selected the ten highest-scoring globally unique papers per topic. A required-phrase match contributed 8 points, a preferred phrase 3, an

excluded phrase -12, the year condition 10, and a computer-science/statistics/electrical-engineering primary category 2.

The topic abbreviations used below are: T1 embodied world models; T2 GUI and computer-use agents; T3 software-engineering agents; T4 agent memory and long-horizon context; T5 multimodal RAG and visual documents; T6 test-time reasoning; T7 controllable video generation; T8 3D and spatial intelligence; T9 efficient inference and KV-cache optimization; and T10 scientific foundation models.

Table 6: Complete source-document inventory. "T," "Cat.," and "Pg." denote topic identifier, primary arXiv category, and rendered page count in the 2,375-unit retrieval corpus, respectively.

arXiv	T	Cat.	Pg.	arXiv	T	Cat.	Pg.	arXiv	T	Cat.	Pg.	arXiv	T	Cat.	Pg.
2601.18203	T5	cs.IR	12	2605.07467	T10	cs.LG	17	2606.05773	T1	cs.RO	14	2606.18478	T7	cs.CV	34
2602.06090	T3	cs.SE	16	2605.13986	T6	cs.LG	83	2606.06322	T2	cs.AI	19	2606.18847	T4	cs.AI	27
2602.06593	T3	cs.SE	21	2605.17641	T4	cs.AI	12	2606.06891	T8	cs.CV	32	2606.19340	T8	cs.RO	21
2602.19961	T5	cs.CL	35	2605.17965	T3	cs.SE	45	2606.06915	T6	cs.CL	15	2606.19531	T1	cs.CV	19
2602.20502	T2	cs.AI	12	2605.17976	T10	cs.AI	31	2606.08688	T10	cs.RO	20	2606.19746	T9	cs.DC	22
2602.22013	T5	cs.CV	10	2605.17985	T10	cs.LG	22	2606.08737	T1	cs.RO	16	2606.20512	T3	cs.SE	20
2602.24134	T5	cs.CV	26	2605.19943	T6	cs.AI	14	2606.09864	T9	cs.LG	61	2606.20729	T10	physics.chem-ph	17
2603.01666	T5	cs.CL	27	2605.21312	T9	cs.DC	20	2606.10183	T7	cs.CV	10	2606.20891	T7	cs.CV	29
2603.07432	T2	cs.CV	25	2605.22829	T5	cs.IR	19	2606.10662	T6	cs.MA	23	2606.21023	T9	cs.LG	25
2603.08013	T2	cs.AI	16	2605.23057	T9	cs.LG	11	2606.11078	T2	cs.AI	29	2606.21633	T9	cs.LG	14
2603.08533	T2	cs.CV	16	2605.23163	T9	cs.CL	17	2606.11576	T6	cs.CV	30	2606.22082	T3	cs.SE	36
2603.12823	T2	cs.CL	16	2605.27570	T6	cs.AI	22	2606.13035	T7	cs.CV	17	2606.22617	T8	cs.CV	14
2603.15921	T3	cs.SE	17	2605.29074	T8	cs.CV	11	2606.13494	T1	cs.RO	22	2606.22631	T8	cs.CV	28
2603.16085	T8	cs.CV	25	2605.29639	T9	cs.OS	14	2606.13672	T1	cs.RO	32	2606.22694	T8	cs.CV	28
2603.17826	T3	cs.SE	15	2605.30917	T5	cs.IR	12	2606.13740	T9	cs.LG	14	2606.22813	T6	cs.AI	53
2603.26211	T2	cs.CV	14	2606.00756	T4	cs.AI	19	2606.13768	T7	cs.CV	32	2606.22844	T4	cs.AI	36
2604.05012	T9	cs.AR	8	2606.00832	T4	cs.CL	9	2606.14162	T7	cs.CV	17	2606.22906	T3	cs.SE	12
2604.14609	T10	cs.AI	65	2606.01613	T5	cs.IR	24	2606.14870	T10	hep-ph	33	2606.22938	T6	cs.LG	38
2604.15821	T10	cs.DC	11	2606.02355	T4	cs.AI	15	2606.14948	T3	cs.SE	19	2606.23296	T1	cs.RO	16
2604.17288	T3	cs.AR	10	2606.02800	T1	cs.CV	139	2606.15001	T10	physics.comp-ph	15	2606.23557	T8	cs.CV	41
2604.23276	T5	cs.CV	8	2606.03099	T4	cs.CL	17	2606.15032	T1	cs.LG	24	2606.23610	T7	cs.CV	24
2604.23941	T2	cs.CV	20	2606.04432	T7	cs.CV	13	2606.17257	T7	cs.CV	17	2606.24480	T10	cond-mat.mtrl-sci	54
2604.27253	T2	cs.AI	21	2606.04527	T7	cs.MM	21	2606.17803	T6	cs.LG	17	2606.24626	T4	cs.AI	16
2604.27996	T4	cs.AI	5	2606.05145	T6	cs.LG	43	2606.18180	T1	cs.CV	16	2606.26800	T8	cs.RO	8
2605.06460	T5	cs.LG	21	2606.05761	T4	cs.AI	47	2606.18375	T1	cs.RO	20	2606.27876	T8	cs.CV	10

Table 6 lists every selected document, primary arXiv category, and rendered page count. Licensing is source-version specific across the 100 documents; the applicable terms are those attached to each versioned arXiv record.

B.2 Descriptor and Path Construction Overview

For each construction unit, a VLM extracts role-specific descriptors d and directed relations r. Each descriptor is assigned one of three roles. A *query anchor* provides a semantic starting point that can be expressed without identifying the target. A *semantic bridge* denotes a stable concept that can align across units but remains hidden from the final query. A *visual target descriptor* captures observable layout or structured evidence without relying on captions, identifiers, or answer-revealing text. A directed relation r connects two compatible descriptors d_s and d_t in the same unit, forming a local triple $d_s \xrightarrow{r} d_t$. These local triples form the building blocks of relation-preserving paths.

Role and quality validation removes revealing query anchors, visual target descriptors that paraphrase ordinary body text, incompatible role pairs, and relations grounded in rejected descriptors. The extraction pass yields 21,955 descriptors and 17,693 relations. Validation retains 20,251 descriptors—6,149 query anchors, 10,531 semantic bridges, and 3,571 visual target descriptors—with 15,390 usable relations. Figure 2 summarizes this inventory, and Appendix B.3 details automated descriptor extraction and validation.

We embed validated semantic bridges with Qwen3-Embedding-8B (Zhang et al., 2025) and retrieve candidate pairs independently within each topic, limiting accidental cross-domain similarity. Embedding similarity provides a high-recall semantic screen, after which a VLM verifies conceptual equivalence or contextual compatibility. The VLM accepts 4,708 of the 10,000 candidate pairs as *semantic alignments* between bridges from different units. Within a document, stable mid-level con-

cepts may align; across documents, we require proper names, aliases, or variants of the same named object.

The verified alignments compose 3,328 candidate paths. Automatic path validation checks semantic coherence, evidence necessity for L2 and L3, and visual necessity. It rejects paths that expose later descriptors, fail to form an interpretable relation chain, or lack grounding in a structured visual region. This leaves 1,522 paths: 364 L1, 1,058 L2, and 100 L3. Appendix B.3 details the semantic-alignment screen, path-composition limits, and automatic-validation criteria.

B.3 Automated Descriptor Extraction, Semantic Alignment, and Path Validation

PDFs were rendered at 144 dpi and their text layers extracted with PyMuPDF. Pages with fewer than 10 extracted words or a reference/boilerplate heading remained in the retrieval corpus and were excluded from query construction. A GPT-5.5 VLM at medium reasoning effort and temperature 0 then received each page image, page text, and document context. Its structured prompt requested up to three query anchors, five semantic bridges, three visual target descriptors, and eight typed directed relations. The validation pass enforced role compatibility, query-anchor non-revelation, semantic-bridge stability, visual distinctiveness of target descriptors, and grounding of directed relations. The extraction pass produced 21,955 descriptors and 17,693 directed relations. Validation retained 20,251 descriptors—6,149 query anchors, 10,531 semantic bridges, and 3,571 visual target descriptors—together with 15,390 usable directed relations.

Qwen3-Embedding-8B embedded 10,531 usable semantic bridges. Matching was performed independently within each topic with cosine threshold 0.72, ten neighbors per semantic bridge, and a cap of 1,000 pairs per topic. GPT-5.5 then judged whether each pair denoted the same or a contextually compatible concept and whether the overlap was too generic. Same-document semantic alignments permit stable mid-level concepts; cross-document semantic alignments must connect a proper name, alias, acronym, or variant of the same named entity. Path composition retained only verified semantic alignments and used per-topic caps of 120 L1, 200 L2, and 200 L3 paths.

The final automatic check used level-specific JSON schemas. L1 required a meaningful directed relation and visually necessary target. L2/L3 additionally tested multihop coherence, semantic-bridge necessity, whether the query anchor directly identified the target, and whether an earlier unit exposed a later semantic bridge. The verbalizability check required every directed relation and endpoint intended for exposure while forbidding semantic bridge identities, semantic alignments, titles, page numbers, figure/table identifiers, exact captions, and path meta-language. It accepted provisional wording for 1,501 paths. A further 21 paths passed the evidence-path checks but failed wording quality; these remained available for full-document review and manual final query authoring.

B.4 Benchmark Statistics

Table 7 summarizes the construction funnel and final composition. Final queries average 25.6 words (median 26; range 15–39). Seventy-four evidence paths remain within one document and six cross document boundaries. Table 8 in Appendix B.5 reports the topic-wise funnel.

B.5 Construction Funnel by Topic

Table 8 reports the complete topic-local funnel. Before automatic validation, the 3,328 composed paths comprise 425 L1, 1,806 L2, and 1,097 L3 paths, including 425 single-unit, 2,746 same-document, and 157 cross-document paths. The 1,522 surviving paths comprise 364 L1, 1,058 L2, and 100 L3 paths.

The fixed 1,000-pair cap gives each topic the same semantic-alignment verification budget. Subsequent counts vary after verification and path validation because topics differ in the availability of specific semantic alignments, visually structured target pages, and shortcut-free L3 paths. Final selection prioritizes reviewed path quality; among paths of comparable quality, it favors broader topic coverage. Cross-document paths are rarer because they require named-entity alignments and full-context validation of all involved documents.

Table 7: Corpus, construction, and final benchmark statistics.

Statistic	**Count**
Corpus	
Source documents / topics	100 / 10
Rendered visual units	2,375
Query-construction units	2,324
Construction	
Candidate semantic-bridge pairs	10,000
Constructed paths	3,328
Automatically validated paths	1,522
Final benchmark	
Queries / unique targets	120 / 120
L1 / L2 / L3 queries	40 / 40 / 40
Single-unit / same-doc. / cross-doc.	40 / 74 / 6
Documents represented in selected paths	55

Table 8: Construction funnel by topic. "Descr.," "Rel.," and "Cand." denote validated descriptors, usable directed relations, and candidate semantic-bridge pairs, respectively. "Verified," "Paths," "Valid," and "Final" count verified alignments, composed paths, automatically validated paths, and selected benchmark items.

Topic	**Descr.**	**Rel.**	**Cand.**	**Verified**	**Paths**	**Valid**	**Final**
T1	2,700	2,049	1,000	426	329	184	13
T2	1,570	1,192	1,000	358	249	114	10
T3	1,782	1,333	1,000	487	289	148	19
T4	1,750	1,347	1,000	322	201	107	8
T5	1,618	1,154	1,000	354	232	115	11
T6	2,901	2,236	1,000	576	385	151	13
T7	1,848	1,432	1,000	486	462	220	10
T8	1,855	1,425	1,000	481	320	178	13
T9	1,780	1,372	1,000	581	450	153	12
T10	2,447	1,850	1,000	637	411	152	11
Total	20,251	15,390	10,000	4,708	3,328	1,522	120

C AUDIT AND CURATION PROTOCOL

C.1 REVIEW AND QUERY-VALIDATION OVERVIEW

Every automatically validated path is reviewed in the context of all involved source documents. Three authors jointly review each path and assign one human score; an AI reviewer independently assigns a second score using the same criteria. Both reviews assess semantic validity, evidence necessity, bridge specificity, visual grounding, query rewritability, and diversity on a 0–100 scale. Semantic validity tests whether every local relation and cross-page alignment is supported by the papers. Evidence necessity tests whether the intermediate concepts are needed to resolve the endpoint. Bridge specificity excludes broad topical overlap, while visual grounding requires an observable structured region on the target. A path remains eligible only when both scores are at least 60 and is then ranked by their mean. We select 40 paths per level, prioritizing quality and using diversity to separate close candidates. Target deduplication and replacement produce 120 distinct targets. Appendix C.2 and Appendix C.3 detail full-document scoring, selection, and target deduplication.

Final queries express the query anchor A, every directed relation r_i in the path skeleton, and an abstracted visual target descriptor V, while hiding semantic bridges B_i and their alignments $\approx$ behind natural generic references. Visual specificity is calibrated so that the visual phrase can match multiple units but its conjunction with semantic and relational constraints identifies the target. Queries request a visual unit without exposing source identifiers, exact captions, or construction terminology. Table 4 in Appendix A.3 summarizes the level-specific exposure and hiding policy.

Each final query is checked against every page of its involved source documents for plausible alternatives satisfying the complete request. Qwen3-VL-Embedding-8B (Li et al., 2026a) then retrieves ten candidates from the full corpus, and every non-gold result is inspected. A valid alternative triggers minimal disambiguation; an item is replaced when uniqueness cannot be restored naturally. The resulting query is rechecked against the relevant full documents and retrieved alternatives. Appendix C.4 and Appendix C.5 provide the full authoring and validation protocols.

C.2 FULL-DOCUMENT PATH REVIEW AND SCORING

All 1,522 automatically validated paths were reviewed, including paths whose provisional wording failed the automatic language check. For a same-document path, the complete source document was read; for a cross-document path, every participating document was read. Ninety-six of the 100 source documents appear in this queue. The other four contribute no automatically validated path and therefore require no path-level judgment.

Three authors jointly reviewed each path and assigned one human score on a 0–100 scale in full-document context. A separate AI reviewer independently scored each path using the same criteria. Both reviews considered semantic validity, necessity of the intermediate evidence, semantic-bridge specificity, visual grounding, whether a natural non-leaking query could be written, and diversity relative to the pool. The human and AI scores were produced in separate passes. A path remained eligible only when both scores were at least 60; eligible paths were then ordered by their arithmetic mean. This double gate prevents one high score from masking a severe objection raised by the other review.

C.3 SELECTION AND TARGET DEDUPLICATION

Selection was path-quality first. Within each level, topic, source-document, visual-form, and reasoning-pattern balance distinguish paths of comparable quality. When two queries point to the same target image, the higher evidence-path level is preferred, followed by the higher mean score within a level. Replacements come from the complete eligible pool and receive the same full-document check. The final set contains 40 queries at each level and 120 distinct target images. Its scope distribution is 40 single-unit, 74 same-document, and 6 cross-document paths, spanning 55 documents across their recorded evidence paths.

C.4 Constraint-Guided Final Query Authoring

Final query writing began only after path selection. The selected path and full-document context were used to (i) expose the query anchor and every directed relation in the level template, (ii) replace semantic bridge identities and semantic alignments with coherent generic referents, (iii) retain a visual target descriptor while omitting captions and unique identifiers, and (iv) formulate the request at visual-unit granularity. For L2 and L3, omitting a relation invalidates the wording even when the endpoints remain recognizable. For L1, the direct relation must likewise remain explicit. Automatic verbalizations served only as expressibility checks.

Visual granularity was calibrated against the full corpus. A phrase was broadened when it encoded a near-pixel-level fingerprint, and narrowed only by the minimal observable distinction needed to exclude a verified alternative. The resulting 120 queries average 25.6 whitespace-delimited words, with a median of 26 and a range of 15–39.

C.5 Hard-Negative and False-Negative Validation

After reading each involved source paper in full, the three authors jointly checked all of its pages for plausible alternatives satisfying the complete query. Qwen3-VL-Embedding-8B additionally retrieved ten page images from the full corpus to surface high-similarity alternatives beyond this document-wide review. Every non-gold result was inspected, regardless of whether the gold target was already retrieved. Layout, object, or phrase overlap alone was insufficient; a valid candidate had to satisfy all constraints expressed by the query anchor, directed relations, and visual target descriptor.

When this validation identifies a complete alternative, the three authors jointly add the minimum observable distinction between the target and that alternative or replace the item. The resulting query is rechecked against the retrieved candidates and the relevant full-document context. These checks concentrate manual review on high-risk alternatives surfaced by a strong retriever.

D Baseline Protocols

We evaluate GPT-5.5, GPT-5.6-luna, GPT-5.6-terra, GPT-5.6-sol, Claude Opus 4.8, Claude Fable 5, Claude Sonnet 5, Claude Opus 5, and the open-weight Qwen3.5-397B-A17B with thinking enabled and disabled. Every agent receives up to 12 interaction steps, followed when needed by a ranking-only finalization call, and is evaluated under the same top-10 contract as the non-agentic baselines. For GPT-5.6-sol, we ablate iterative search, page inspection, route-specific text access, and regional cropping. Appendices D.4 and D.5 specify the planner protocol and ablation configurations. Table 11 in Appendix D.6 reports per-query resource indicators. Appendix H describes the open-weight deployment and reports per-setting runtimes in Table 18.

D.1 Evaluation Protocol and Metrics

The evaluation corpus contains $N = 2{,}375$ units. Each query has one gold target, and systems are evaluated under a contract requiring $K = 10$ distinct opaque identifiers. Table 5 in Appendix A.4 specifies annotation roles and their inference-time visibility. We report Recall@k for $k \in \{1, 3, 5, 10\}$ and MRR@10. Let $r_q = \mathrm{rank}_q(a_q)$, with missing or invalid rankings assigned $r_q = \infty$:

$$\begin{cases} \mathrm{R@}k = \dfrac{1}{|Q|} \displaystyle\sum_{q \in Q} \mathbb{1}[r_q \le k], \\ \mathrm{MRR@10} = \dfrac{1}{|Q|} \displaystyle\sum_{q \in Q} \dfrac{\mathbb{1}[r_q \le 10]}{r_q}. \end{cases} \tag{8}$$

R@1 is the primary measure, R@10 measures final top-10 coverage, and MRR@10 summarizes target ordering within that list. We report overall and level-wise results.

D.2 Embedding Indices

The visual single-vector index contains one normalized Qwen3-VL-Embedding-8B vector for each of the 2,375 rendered page images. The OCR-text index first parses every page image with PaddleOCR-VL-1.6 using PP-DocLayoutV3, concatenates the recovered plain-text blocks in reading order, and embeds that text with Qwen3-Embedding-8B. Static retrieval encodes the benchmark query, whereas an agent search encodes the planner-issued search string; neither route adds a retrieval instruction. Cosine similarity over normalized vectors determines the returned order. The page-image vectors, OCR records, and OCR-text vectors are cached once and reused by static and agent baselines.

The sparse OCR baseline uses Okapi BM25 with $k_1 = 1.2$ and $b = 0.75$. The hybrid OCR baseline takes the top 1,000 BM25 and dense results and combines their ranks with reciprocal-rank fusion using $k = 60$. Nemotron-ColEmbed-VL-8B-V2 encodes each query and page as token-level vectors and scores them with the model's MaxSim operator. It uses the model-provided processor, BF16 weights, and no additional query instruction.

D.3 Agent Tools

Table 9: Information interface for the two tool-using retrieval routes. Table 10 in Appendix D.3 gives the complete schemas and argument constraints.

Route	Operation	Returned information
Visual	Visual search	Ranked opaque page handles and similarity scores.
	On-demand OCR	Text and layout blocks from one discovered page or crop.
OCR-Text	Text search	Ranked opaque page handles and similarity scores; no snippets.
	Cached page OCR	Full OCR text for one discovered page.
Both	Page inspection	Full-page images for up to ten discovered pages.
	Regional cropping	Up to ten crops from previously inspected pages.
	Ranked submission	Up to ten distinct discovered pages; exactly ten when available.

Table 10 gives the complete inference interface. Search results expose only opaque handles and scores, withholding snippets, source titles, document identifiers, and page numbers. Image-returning tools deliver image artifacts to the next planner turn in addition to the compact JSON status shown in the table. Full-page inspection, regional cropping, and on-demand text recognition follow recent visual-agent designs (He et al., 2026; Wang et al., 2026b).

D.4 Planner and Episode Configuration

The closed-source planners are GPT-5.5, GPT-5.6-luna, GPT-5.6-terra, GPT-5.6-sol, Claude Opus 4.8, Claude Fable 5, Claude Sonnet 5, and Claude Opus 5. We also evaluate the open-weight Qwen3.5-397B-A17B with thinking enabled and disabled. Reasoning effort is set to `medium` where supported.

The neutral planner prompt states the closed-corpus objective, enumerates the available tools, explains that handles are opaque, and requires exactly one JSON action per turn. It omits the gold path, evidence units, document identity, topic, and level. The planner may choose an information action or submit a ranking during each step of a 12-step interaction budget. The ranking must contain ten distinct pages whenever ten candidates have been discovered, and rank-one accuracy remains the primary objective. Per-rank rationales and evidence handles are logged for audit and excluded from scoring.

If no valid ranking has been submitted when the 12-step interaction budget is exhausted, the harness makes one ranking-only finalization call that exposes the eligible discovered and examined handles but disables all information tools. A malformed evidence list is normalized by removing unseen evidence handles while preserving the ranking. Duplicate, unknown, or wrong-length rankings remain invalid.

Table 10: Frozen information-tool and terminal-action schemas. The two global search tools are mutually exclusive: no evaluated agent receives a hybrid index.

Route	Action	Arguments	Observation and constraints
Visual	`visual_search`	`query`: string; `top_k`: integer $[1, 50]$	Ordered `page_handle`, score pairs from the page-image index.
	`ocr_image`	`target_handle`: one searched full page or generated crop	Plain OCR text and blocks from PaddleOCR-VL-1.6; accepts one image per action.
OCR-Text	`text_search`	`query`: string; `top_k`: integer $[1, 50]$	Ordered `page_handle`, score pairs from the OCR-text index; no snippets.
	`get_page_ocr`	`page_handle`: one searched full page	Complete cached plain OCR text; no OCR model is rerun and crop handles are rejected.
Both	`inspect_pages`	`page_handles`: 1–10 distinct searched pages	Full-page image artifacts plus handle and image dimensions. The call is atomic.
	`crop_pages`	`crops`: 1–10 objects containing `page_handle`, normalized $[x_1, y_1, x_2, y_2]$, and optional `purpose`	Cropped image artifacts and crop handles. The source page must have been inspected. Coordinates use top-left $(0, 0)$ and bottom-right $(1, 1)$. Valid entries survive other entries' failures; no maximum area is imposed.
	`submit_answer`	Up to ten distinct ranked objects, each with `page_handle`, `evidence_page_handles`, and a short rationale; one overall rationale	Terminates the episode. Exactly ten are required when available. Only discovered pages may be ranked; evidence handles must have been read or viewed.

D.5 Ablation Configurations

All ablations use GPT-5.6-sol and retain the remaining route settings. The no-iterative-search variant fixes its candidate set to the 30 highest-ranked pages from the corresponding Qwen3 retriever, judges ten independent three-page batches from full page images and cached OCR, and aggregates the judgments into a top-10 ranking with one final call. This candidate-set size slightly exceeds the full agent's mean page exposure on both routes and provides a common verification budget. The fixed pipeline cannot reformulate searches or replace candidates. The other variants remove regional cropping, route-specific text access, or full-page inspection; removing page inspection also disables cropping, which depends on an inspected source page.

D.6 Inference Resources and Infrastructure

Table 11: Per-query resource indicators for valid GPT-5.6-sol episodes. Input and output token counts sum usage over all model calls in an episode.

System	Route	Model calls	Pages viewed	Input tokens	Output tokens
No iterative search	Visual	11.00	30.00	101,075	8,932
No iterative search	OCR-Text	11.00	30.00	124,925	8,597
Full agent	Visual	8.34	25.22	177,056	2,250
Full agent	OCR-Text	10.18	28.22	248,125	2,632

Table 11 compares the measured per-query inference resources of the no-iterative-search control and the full agent. With 30 page observations and 11 model calls, the fixed pipeline matches or exceeds the full-agent averages on both routes. The resulting page exposure and model-call counts place the two systems on a comparable scale, focusing the comparison on whether intermediate observations can redirect subsequent retrieval. Their token profiles reflect the two execution structures: independent batch judgments produce more output tokens, whereas the accumulating full-agent history produces more input tokens.

Page rendering uses PyMuPDF at 144 dpi. OCR records and embedding indices are cached before evaluation. The visual and OCR-text routes run on separate NVIDIA A100-SXM4-80GB devices. Immutable caches are shared read-only, while runtime crops and episode state remain isolated.

E Detailed Results

E.1 Level-Wise Main Results

Table 12 distinguishes direct query–page matching from retrieval that depends on one or two semantic bridges. Late interaction is nearly perfect on direct items yet drops sharply on both bridge levels. Tool-using agents provide the strongest results on the longer evidence paths.

E.2 Topic-Level Results

Table 13 reports all five metrics for GPT-5.6-sol by topic. The visual route matches or exceeds OCR-text in every topic at R@1. OCR-text varies more sharply: it reaches 63.64% R@1 on T10 but only 7.69% on T8, where spatial diagrams and view-dependent layouts are poorly represented by OCR text alone.

E.3 Level-Wise Ablation Results

Table 14 provides the level-wise ablations underlying the aggregate results in Section 6.2. Iterative search and page inspection generally improve both routes, with inspection producing the largest losses when removed. Text access benefits L1 and L2 but has mixed effects on L3, reflecting a path-dependent trade-off between local verification and corpus exploration. Cropping likewise has no stable effect across routes or levels.

Table 12: Level-wise retrieval results on VISDOCAGENTBENCH (%). Each level contains 40 queries. Invalid completed episodes are scored as zero. Bold marks the best result for each route, level, and metric; tied best values are all bold.

Retriever / Planner	Level	Visual					OCR-Text				
		R@1	R@3	R@5	R@10	MRR@10	R@1	R@3	R@5	R@10	MRR@10
Static Retrievers											
Qwen3 Embedding	L1	40.00	52.50	60.00	72.50	49.11	2.50	5.00	10.00	17.50	5.84
	L2	20.00	27.50	30.00	40.00	25.30	2.50	10.00	15.00	20.00	7.50
	L3	2.50	10.00	12.50	22.50	7.62	0.00	2.50	15.00	20.00	4.68
BM25	L1	–	–	–	–	–	15.00	30.00	37.50	55.00	25.59
	L2	–	–	–	–	–	5.00	12.50	15.00	22.50	9.49
	L3	–	–	–	–	–	0.00	2.50	5.00	12.50	2.31
BM25 + dense RRF	L1	–	–	–	–	–	10.00	22.50	32.50	42.50	19.55
	L2	–	–	–	–	–	5.00	12.50	17.50	22.50	10.40
	L3	–	–	–	–	–	0.00	7.50	7.50	15.00	4.28
Nemotron ColEmbed	L1	**97.50**	**97.50**	**100.00**	**100.00**	**98.00**	–	–	–	–	–
	L2	20.00	42.50	65.00	70.00	36.34	–	–	–	–	–
	L3	2.50	17.50	30.00	40.00	12.24	–	–	–	–	–
Tool-Using Agents: Closed-Source Planners											
GPT-5.5	L1	75.00	75.00	77.50	80.00	75.94	30.00	30.00	30.00	32.50	30.31
	L2	57.50	57.50	60.00	60.00	58.13	27.50	27.50	27.50	27.50	27.50
	L3	47.50	52.50	55.00	67.50	51.56	22.50	25.00	25.00	25.00	23.75
GPT-5.6-luna	L1	65.00	65.00	65.00	72.50	66.13	20.00	20.00	20.00	27.50	21.09
	L2	42.50	45.00	50.00	55.00	45.38	15.00	15.00	15.00	15.00	15.00
	L3	22.50	32.50	37.50	45.00	28.75	2.50	7.50	20.00	22.50	8.23
GPT-5.6-terra	L1	77.50	80.00	80.00	82.50	79.06	32.50	32.50	32.50	37.50	33.11
	L2	52.50	52.50	52.50	57.50	53.17	25.00	27.50	27.50	30.00	26.19
	L3	32.50	40.00	42.50	52.50	37.73	22.50	25.00	25.00	32.50	24.46
GPT-5.6-sol	L1	85.00	87.50	87.50	90.00	86.15	42.50	42.50	42.50	47.50	43.06
	L2	60.00	62.50	62.50	62.50	61.25	**40.00**	**40.00**	40.00	40.00	**40.00**
	L3	40.00	45.00	47.50	52.50	43.36	27.50	32.50	40.00	45.00	32.12
Claude Opus 4.8	L1	70.00	72.50	75.00	82.50	72.32	22.50	22.50	25.00	40.00	24.79
	L2	50.00	55.00	57.50	62.50	52.88	32.50	37.50	40.00	42.50	35.33
	L3	22.50	35.00	45.00	60.00	31.97	17.50	22.50	25.00	40.00	22.01
Claude Fable 5	L1	80.00	82.50	85.00	92.50	82.57	40.00	42.50	42.50	**52.50**	42.38
	L2	52.50	57.50	62.50	67.50	55.88	30.00	37.50	40.00	**47.50**	35.01
	L3	**55.00**	**62.50**	**67.50**	**80.00**	**61.61**	**35.00**	**45.00**	**47.50**	**57.50**	**40.63**
Claude Sonnet 5	L1	42.50	57.50	62.50	72.50	50.75	20.00	20.00	22.50	25.00	20.94
	L2	32.50	42.50	42.50	52.50	38.27	15.00	15.00	20.00	20.00	16.00
	L3	15.00	22.50	35.00	42.50	22.96	17.50	25.00	25.00	25.00	20.42
Claude Opus 5	L1	92.50	92.50	92.50	92.50	92.50	**47.50**	**47.50**	**47.50**	50.00	**47.78**
	L2	**62.50**	**67.50**	**67.50**	**72.50**	**64.69**	32.50	**40.00**	**42.50**	**47.50**	37.05
	L3	47.50	52.50	55.00	60.00	51.09	32.50	37.50	45.00	50.00	36.90
Tool-Using Agents: Open-Weight Planner											
Qwen3.5 (thinking)	L1	47.50	62.50	67.50	67.50	55.29	15.00	15.00	17.50	22.50	16.09
	L2	22.50	25.00	25.00	40.00	25.80	5.00	5.00	5.00	7.50	5.31
	L3	12.50	12.50	12.50	12.50	12.50	5.00	7.50	7.50	17.50	7.69
Qwen3.5 (no thinking)	L1	35.00	45.00	60.00	67.50	43.71	2.50	2.50	10.00	12.50	4.56
	L2	17.50	17.50	20.00	32.50	19.65	2.50	2.50	5.00	5.00	3.00
	L3	5.00	7.50	10.00	15.00	7.28	0.00	7.50	7.50	12.50	3.94

Table 13: GPT-5.6-sol topic-level retrieval results (%). Topic names are defined in Appendix B.1.

Route	Topic	Queries	R@1	R@3	R@5	R@10	MRR@10
Visual	T1	13	38.46	46.15	46.15	61.54	44.55
	T2	10	70.00	70.00	70.00	70.00	70.00
	T3	19	42.11	52.63	57.89	57.89	47.54
	T4	8	75.00	87.50	87.50	87.50	79.17
	T5	11	72.73	72.73	72.73	72.73	72.73
	T6	13	76.92	76.92	76.92	76.92	76.92
	T7	10	40.00	40.00	40.00	40.00	40.00
	T8	13	69.23	69.23	69.23	69.23	69.23
	T9	12	66.67	66.67	66.67	75.00	67.86
	T10	11	81.82	81.82	81.82	81.82	81.82
OCR-Text	T1	13	38.46	53.85	53.85	53.85	44.87
	T2	10	30.00	30.00	30.00	30.00	30.00
	T3	19	21.05	21.05	31.58	47.37	25.75
	T4	8	37.50	37.50	37.50	37.50	37.50
	T5	11	54.55	54.55	54.55	54.55	54.55
	T6	13	53.85	53.85	53.85	53.85	53.85
	T7	10	40.00	40.00	40.00	50.00	41.00
	T8	13	7.69	7.69	15.38	15.38	9.62
	T9	12	33.33	33.33	33.33	33.33	33.33
	T10	11	63.64	63.64	63.64	63.64	63.64

E.4 Results by Path Scope

Table 15 separates single-unit, same-document, and cross-document paths. Visual R@1 decreases from 85.00% on single-unit paths to 52.70% on same-document paths and 16.67% on cross-document paths; OCR-text R@1 decreases from 42.50% to 36.49% and 0.00%, respectively. The scope breakdown shows that aligning evidence across documents remains the most challenging setting, while visual retrieval retains an advantage in placing the target first.

F Trace and Qualitative Analysis

F.1 Target Progression Events

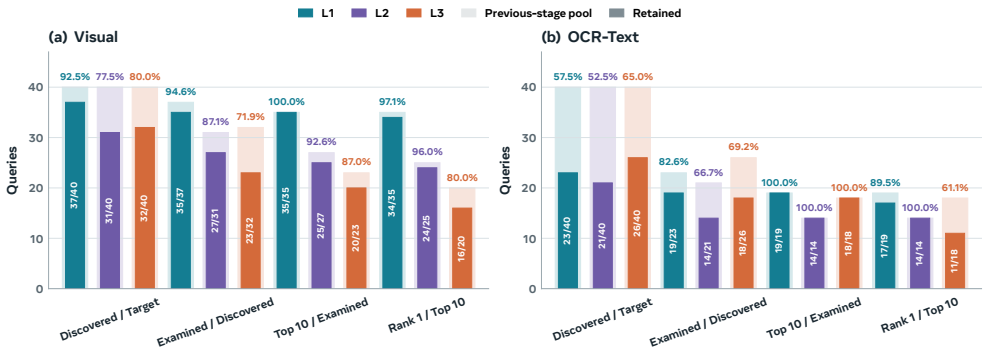


Figure 4: Level-wise gold-target progression through GPT-5.6-sol traces for (a) Visual and (b) OCR-Text retrieval. Colors distinguish L1, L2, and L3; pale bars show the previous-stage pool, while solid bars show the retained subset. Overall route rates appear in Figure 3.

Our post-hoc audit uses planner responses, parsed actions, tool observations including returned pages and crops, cumulative usage, and final rankings. We map opaque handles back to annotation

Table 14: Level-wise GPT-5.6-sol ablations on both retrieval routes (%).

Configuration	Level	R@1	R@3	R@5	R@10	MRR@10
Visual						
Full agent	L1	85.00	87.50	87.50	90.00	86.15
	L2	60.00	62.50	62.50	62.50	61.25
	L3	40.00	45.00	47.50	52.50	43.36
No iterative search	L1	77.50	77.50	77.50	77.50	77.50
	L2	47.50	50.00	50.00	50.00	48.75
	L3	35.00	50.00	50.00	55.00	42.80
No regional crop	L1	87.50	90.00	90.00	90.00	88.75
	L2	62.50	62.50	65.00	67.50	63.28
	L3	42.50	55.00	60.00	65.00	50.00
No OCR	L1	77.50	77.50	77.50	77.50	77.50
	L2	50.00	55.00	57.50	62.50	53.90
	L3	52.50	55.00	65.00	67.50	55.82
No page inspection	L1	67.50	72.50	75.00	77.50	70.86
	L2	40.00	42.50	47.50	52.50	42.55
	L3	30.00	37.50	50.00	60.00	36.68
OCR-Text						
Full agent	L1	42.50	42.50	42.50	47.50	43.06
	L2	40.00	40.00	40.00	40.00	40.00
	L3	27.50	32.50	40.00	45.00	32.12
No iterative search	L1	30.00	30.00	30.00	30.00	30.00
	L2	22.50	25.00	25.00	25.00	23.75
	L3	30.00	32.50	37.50	37.50	32.08
No regional crop	L1	42.50	42.50	42.50	45.00	42.75
	L2	37.50	40.00	40.00	40.00	38.75
	L3	22.50	30.00	32.50	35.00	26.40
No page OCR	L1	40.00	40.00	40.00	42.50	40.36
	L2	30.00	30.00	30.00	32.50	30.31
	L3	30.00	35.00	37.50	42.50	33.22
No page inspection	L1	20.00	22.50	27.50	30.00	22.78
	L2	12.50	12.50	20.00	20.00	14.13
	L3	12.50	22.50	22.50	32.50	17.95

Table 15: GPT-5.6-sol results by evidence-path scope (%).

Route	Scope	Queries	R@1	R@3	R@5	R@10	MRR@10
Visual	Single-unit	40	85.00	87.50	87.50	90.00	86.15
	Same-document	74	52.70	56.76	58.11	60.81	55.19
	Cross-document	6	16.67	16.67	16.67	16.67	16.67
OCR-Text	Single-unit	40	42.50	42.50	42.50	47.50	43.06
	Same-document	74	36.49	39.19	43.24	44.59	38.85
	Cross-document	6	0.00	0.00	0.00	16.67	1.67

units after an episode finishes. The overall progression in Figure 3 and its level-wise breakdown in Figure 4 use the following event definitions:

- The target is *discovered* when its handle first appears in a successful search result at any rank.
- A full page is *examined* at the first successful `inspect_pages`, full-page `ocr_image`, or `get_page_ocr` action involving it. A successful crop also counts as examination of its source page.
- Target progression follows discovery, examination conditional on discovery, final top-10 inclusion conditional on examination, and rank-one placement conditional on top-10 inclusion.

F.2 Two Audited Level-3 Trajectories

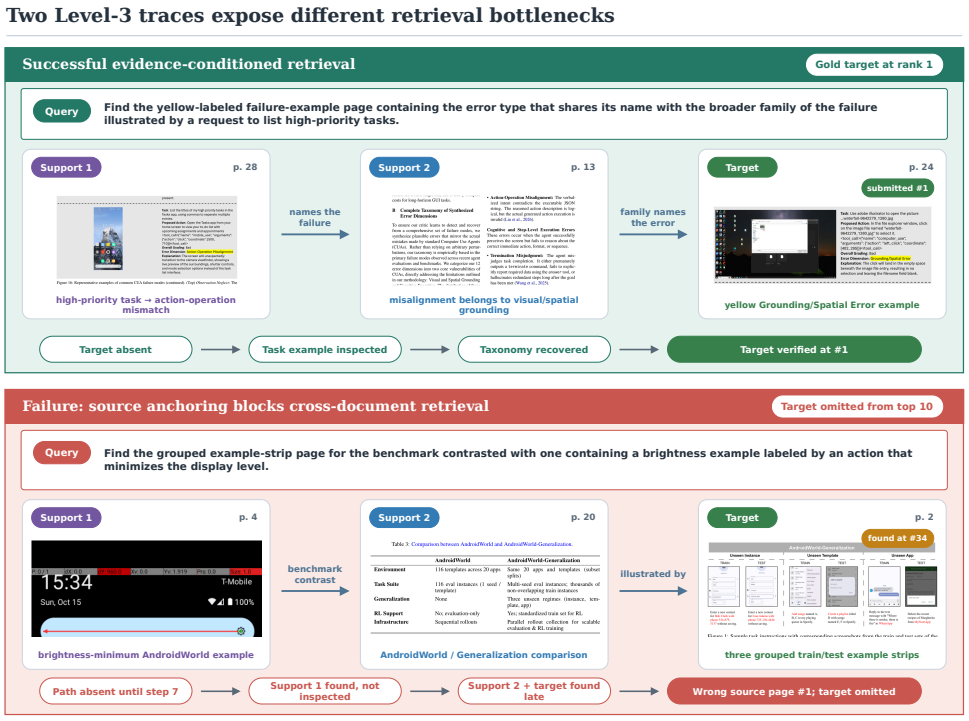


Figure 5: Queries and enlarged evidence regions for two audited Level-3 visual-route episodes. Top: the first search finds only the task-example support; inspecting it redirects retrieval toward the taxonomy and the yellow-labeled target, which is submitted at rank 1. Bottom: the agent repeatedly anchors retrieval to the paper containing the brightness example. The benchmark-comparison support and target appear only in its final search at ranks 25 and 34, remain unexamined, and the target is omitted from the submitted top 10.

We examine one successful and one failed GPT-5.6-sol visual-route episode. Figure 5 enlarges the decisive regions of each support and target page. Tables 16 and 17 retain every planner decision and tool call, with long search lists summarized by path-page ranks.

Evidence changes the successful search state. The first search in Table 16 finds the task example while missing the taxonomy and target. Inspection identifies its failure mode, and the next search surfaces both remaining path pages. Subsequent OCR resolves the broader family and verifies its namesake error on the target page. The episode illustrates a complete transition from partial evidence to a grounded rank-one decision.

Table 16: Complete successful visual-route trajectory for the top case in Figure 5. Search observations are summarized by the ranks of the annotated path pages.

Step	Decision rationale	Action	Observation and state change
1	Search for the yellow-labeled examples and the high-priority-task failure.	`visual_search`; top 30.	Support 1 appears at rank 2; Support 2 and the target are absent.
2	Inspect the strongest failure-example candidates.	`inspect_pages` on ten pages, including Support 1.	The high-priority-task request is linked to Action–Operation Misalignment.
3	Search for that failure mode and its broader family.	`visual_search`; top 30.	Support 1, Support 2, and the target appear at ranks 3, 6, and 7.
4	Refine toward a yellow-highlighted family label.	`visual_search`.	The three path pages remain at ranks 3, 6, and 7.
5	Compare unseen pages in the failure-example series.	`inspect_pages` on ten pages, including Support 2 and the target.	The taxonomy page and the yellow-labeled candidate are visually identified.
6	Return to the exact high-priority-task wording.	`visual_search`; top 30.	Support 1 is rank 1, Support 2 rank 7, and the target rank 18.
7	Recheck the strongest task-example candidates.	`inspect_pages` on ten pages, including Support 1.	No competing example supplies a better semantic chain.
8	Test a yellow-labeled alternative.	`ocr_image` on a candidate page.	OCR identifies Action Formulation Error and Constraint Neglect, excluding it.
9	Test another visually matching alternative.	`ocr_image` on a candidate page.	OCR identifies Suboptimal Path and Timing and Latency Neglect, excluding it.
10	Resolve the broader failure family.	`ocr_image` on Support 2.	OCR places Action–Operation Misalignment under Visual and Spatial Grounding and names Grounding/Spatial Error.
11	Verify the namesake error on the target candidate.	`ocr_image` on the target.	OCR confirms the yellow-highlighted Grounding/Spatial Error example.
12	Submit the verified target.	`submit_answer` with a ranked top 10.	The target is rank 1, followed by Support 1 and Support 2.

Table 17: Complete failed visual-route trajectory for the bottom case in Figure 5. Search observations are summarized by the ranks of the annotated path pages and the submitted rank-one page.

Step	Decision rationale	Action	Observation and state change
1	Search for grouped examples and a brightness-minimizing action.	`visual_search`; top 20.	Neither support nor the target appears.
2	Inspect the strongest grouped-example candidates.	`inspect_pages` on ten pages.	No annotated path page is examined, and no benchmark relation is recovered.
3	Combine benchmark-example and minimum-brightness cues.	`visual_search`.	All three path pages remain absent.
4	Search the brightness instruction more literally.	`visual_search`.	All three path pages remain absent.
5	Emphasize the grouped smartphone layout.	`visual_search`.	All three path pages remain absent.
6	Inspect additional visual matches.	`inspect_pages` on nine pages.	The inspected pages still provide no annotated path evidence.
7	Hypothesize AndroidControl or GUIOdyssey as the target benchmark.	`visual_search`; top 30.	Support 1 first appears at rank 18; Support 2 and the target are absent.
8	Inspect candidates from the benchmark hypothesis.	`inspect_pages` on ten pages.	Support 1 is not selected for inspection, so its AndroidWorld label remains unused.
9	Recast the query as a DragOn figure search.	`visual_search`; top 50.	Support 1 rises to rank 4, but remains unexamined; Support 2 and the target are absent.
10	Search within DragOn using AndroidWorld and OSWorld cues.	`visual_search`; top 50.	The eventual rank-one page is rank 2, Support 1 rank 4, Support 2 rank 25, and the target rank 34.
11	Prioritize the source-paper candidates.	`submit_answer` with a ranked top 10.	A DragOn prose page is rank 1 and Support 1 rank 2; Support 2 and the target are omitted.

Source anchoring can block a cross-document relation. The failed episode in Table 17 retrieves no annotated path page during its first six actions. Once the search begins surfacing DragOn pages, the planner increasingly treats that source document as the destination. Support 1 reaches rank 4, and the final search exposes Support 2 and the target at ranks 25 and 34, but none is examined. The submission consequently ranks a DragOn prose page first and omits the requested grouped example strip. The planner has mistaken the benchmark containing the brightness example for the target instead of following the contrast relation to AndroidWorld-Generalization, localizing the failure to relation-following and candidate selection before visual verification.

G Licensing and Privacy

We check the license attached to the exact arXiv version of every source document. The corpus specification records that version, its source URL, and the applicable license. Page images and substantial source-derived text are redistributed only when the corresponding terms permit it. For other sources, local corpus reconstruction proceeds from the versioned record using the deterministic rendering procedure.

Benchmark-authored annotations and software are licensed independently of the source documents. Their licenses do not extend to source text or images, including excerpts present in OCR records or agent traces, which remain governed by the corresponding source license.

Rendered scholarly pages may contain author names, affiliations, and contact addresses already published in the source documents. We retain only this source-visible content and the bibliographic metadata needed for source identification and attribution; we neither infer personal attributes nor combine the records with external personal data. Requests to correct or remove source-derived personal information can be directed to the project contact.

H Open-Weight Baseline Resources and Runtime

The open-weight planner is Qwen3.5-397B-A17B, with 397B total parameters and 17B activated parameters per token. It is served in BF16 on two nodes, each with eight NVIDIA A100-SXM4-80GB GPUs. Tensor parallelism is eight within each node and pipeline parallelism is two across nodes; evaluation uses four query workers and the same 12-step protocol as the other planners. The visual and OCR-text routes use Qwen3-VL-Embedding-8B and Qwen3-Embedding-8B, respectively. PaddleOCR-VL-1.6 (0.9B) and PP-DocLayoutV3 provide page-text recognition and layout detection. Embedding and OCR services run on reserved memory of the same local A100 infrastructure; construction of each static embedding index uses one A100-SXM4-80GB GPU. All listed models use their official weights.

Table 18 summarizes the full 120-query inference time for each combination of retrieval route and Qwen3.5 reasoning mode.

Table 18: Approximate end-to-end inference time for each open-weight agent setting. Timing starts after all model services are ready and excludes model loading and deployment.

Qwen3.5 setting	**Runtime for 120 queries**
Visual, thinking	4–5 hours
Visual, no thinking	1–1.5 hours
OCR-Text, thinking	6.5–7.5 hours
OCR-Text, no thinking	40–50 minutes