ATTRISWARM: ORCHESTRATING AGENT SWARMS FOR CROSS-PAGE VISUAL ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) increasingly address knowledge-intensive tasks by integrating evidence scattered across multiple document pages into a final answer. Because such answers are synthesized from many pages, the link between the answer and its supporting visual evidence can become obscured, hindering verification and trust. Cross-page visual attribution is therefore essential: it requires identifying the supporting pages and localizing fine-grained evidence regions within them, thereby making multi-page reasoning transparent and auditable. Existing approaches either process multiple pages jointly or iteratively refine evidence within a single trajectory, risking cross-page context pollution and, for interactive methods, substantial sequential overhead. We introduce **AttriSwarm**, an agent-swarm framework that assigns these concerns to separate roles: a leader handles global cross-page reasoning, while subagents perform page-wise attribution in isolation and in parallel. To support training and evaluation, we construct **AttriDoc**, a cross-page QA dataset linking supporting pages to their annotated evidence regions, and use it for role-specific training of AttriSwarm, improving answer accuracy, page-level F_1, and region-level IoU over its training-free counterpart by an average of 18.94, 29.34, and 23.04 percentage points, respectively, across three open-source backbones and three datasets.

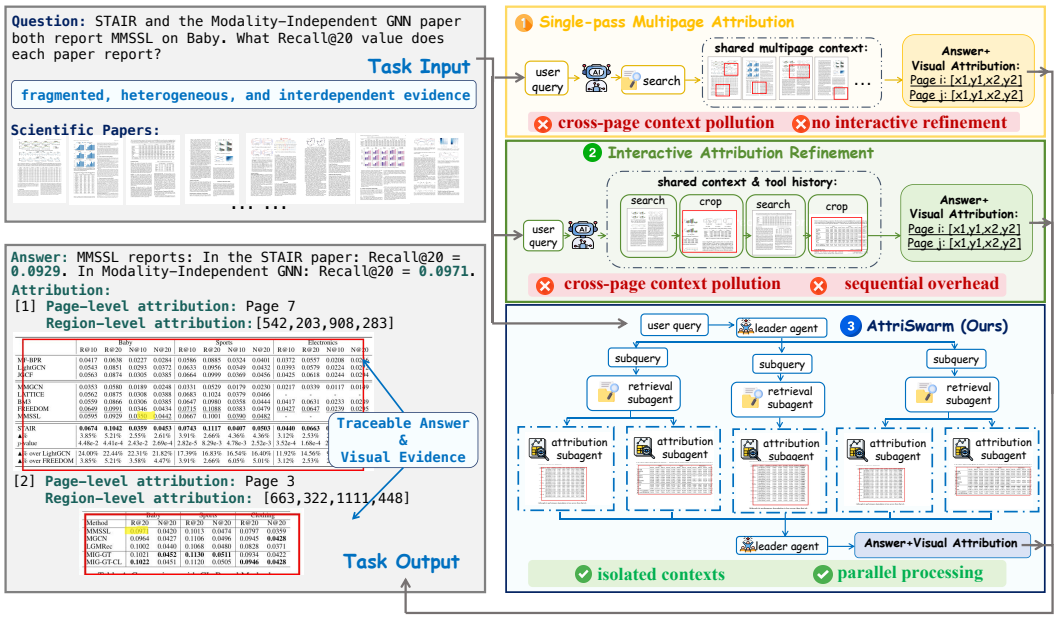

Figure 1: **Example and paradigms for cross-page visual attribution.** *Single-pass multipage attribution* jointly processes candidate pages without interactive refinement. *Interactive attribution refinement* supports iterative evidence localization but introduces sequential overhead. Both paradigms operate in shared contexts and are susceptible to cross-page context pollution. **AttriSwarm** isolates and parallelizes page-wise attribution while preserving global evidence integration.

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.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have enabled their application to complex knowledge-intensive tasks, such as multimodal deep research (Huang et al., 2026b;a; Yu et al., 2026; Yao et al., 2026; Zeng et al., 2026; Huang et al., 2026c; Yang et al., 2025) and scientific literature analysis (Xiong et al., 2026; Sun et al., 2025a; Li et al., 2026). When the evidence supporting an answer is contained within a single page, visual attribution can localize the relevant regions within that page, making the answer traceable to its source. However, many knowledge-intensive tasks require models to search, select, and integrate evidence distributed across multiple pages. In such scenarios, **cross-page visual attribution** requires identifying both the supporting pages and the corresponding evidence regions within each page, allowing generated answers to be verified against their source material, as illustrated in Figure 1. Compared with the single-page setting, cross-page attribution is more challenging because the evidence required for an answer is often fragmented across pages, heterogeneous in form, and is highly interdependent.

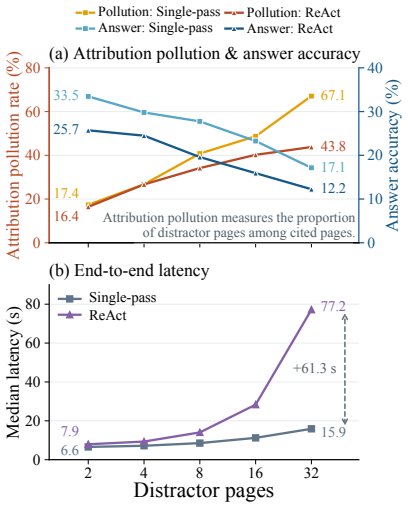

Figure 2: **Cross-page context pollution and sequential overhead.** As distractor pages increase, (a) both methods exhibit higher attribution pollution and lower answer accuracy; (b) ReAct exhibits steeper latency growth.

Existing approaches to cross-page visual attribution broadly follow two paradigms. *Single-pass multipage attribution* processes all candidate pages jointly (Ma et al., 2025; Zhao et al., 2025), but cannot revisit ambiguous evidence or inspect finer-grained details through subsequent interactions. To address this limitation, *interactive attribution refinement* supports iterative searching, cropping, or zooming (Wang et al., 2025b; 2026c;b; Zhao et al., 2026a). However, despite this additional flexibility, both paradigms can suffer from **cross-page context pollution** (Phillips, 2026), where irrelevant pages in a shared context interfere with evidence selection and localization. To assess their susceptibility to this problem, we inject 2–32 distractor pages into a subset of CiteVQA (Ma et al., 2026) and compare single-pass attribution with a representative interactive refinement method, ReAct (Yao et al., 2023), under the same Qwen3-VL-4B-Instruct backbone (Team, 2025). As shown in Figure 2 (a), both methods exhibit higher attribution pollution (the proportion of distractor pages among cited pages) and lower answer accuracy as distractors increase. Meanwhile, ReAct's latency increases sharply with page count (Figure 2 (b)), highlighting the **sequential overhead** of serial page-wise attribution.

Architecturally, these issues arise from coupling heterogeneous reasoning states and page-wise operations within a single trajectory. While recent studies on agent swarms and multi-agent delegation leverage context isolation, task decomposition, and parallel execution to alleviate this burden (Xu et al., 2026; Song et al., 2026; Team, 2026; Zhang et al., 2026a; Virk et al., 2026), cross-page visual attribution faces a unique challenge. Its primary bottleneck is the intrinsic conflict between isolating local relevance assessment and region localization, and preserving the global evidence relationships essential for reasoning.

Building on these capabilities, we introduce **AttriSwarm**, an agent-swarm framework that preserves global evidence integration while isolating local attribution. The leader decomposes the query into subqueries that guide parallel retrieval and specify the evidence sought in local attribution. Each attribution subagent receives a single candidate page and its corresponding subquery to assess relevance and localize supporting regions within an isolated context. Attribution runs in parallel to reduce sequential overhead, while the leader integrates the returned evidence for cross-page reasoning and generates the final answer with visual attributions.

To support training and evaluation, we construct **AttriDoc** with cross-page QA and page- and region-level annotations: **AttriDoc-Paper** is synthesized from tables linked by shared entities across scientific papers, while **AttriDoc-Slide** augments multipage QA pairs from SlideVQA (Tanaka et al.,

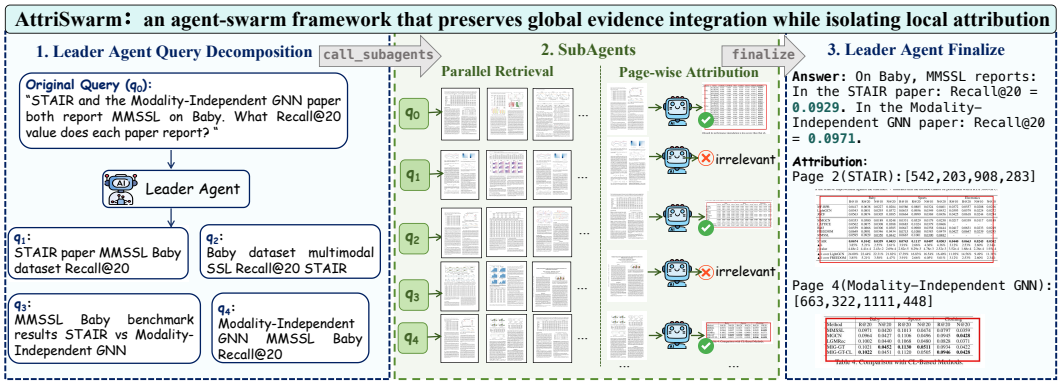

Figure 3: **Overview of AttriSwarm.** AttriSwarm decomposes cross-page reasoning into leader-based query decomposition, parallel retrieval and page-wise attribution of subagents, and final evidence aggregation by the leader agent.

2023) with region-level attribution annotations. To make this decomposition effective with general-purpose MLLMs, we further develop a role-specific training strategy for the leader and attribution subagents, improving answer accuracy, page-level F_1, and region-level IoU over its training-free counterpart by an average of 18.94, 29.34, and 23.04 percentage points, respectively, across three open-source backbones and three datasets.

Our contributions are threefold for complex cross-page visual attribution: (1) Architecturally, we propose **AttriSwarm**, an agent-swarm framework that isolates and parallelizes page-wise attribution while preserving global cross-page reasoning. (2) We construct **AttriDoc**, a multimodal dataset comprising scientific-paper and slides, with cross-page QA pairs and page- and region-level attributions. (3) We develop a role-specific training strategy for leader and attribution subagents, and extensive experiments demonstrate improvements in both answer accuracy and visual attribution.

2 Architecture of AttriSwarm

We present AttriSwarm (Figure 3), a swarm-based framework that mitigates cross-page context pollution and sequential overhead. The leader decomposes the query and integrates evidence, while parallel, context-isolated subagents assess page relevance and localize supporting regions.

2.1 Task Formulation

We formulate the cross-page visual attribution task in a retrieval-augmented multimodal setting. Given a user query q and a document page corpus $\mathcal{D} = \{p_1, p_2, \ldots, p_{|\mathcal{D}|}\}$, the goal is to generate a grounded answer a along with explicit visual attributions at both the page and region levels, denoted as $\mathcal{S}_q$ and $\mathcal{B}_q$, respectively:

$$(q, \mathcal{D}) \rightarrow (a, \mathcal{S}_q, \mathcal{B}_q) \tag{1}$$

Given a query q, a retrieval module retrieves the top-K candidate pages $\mathcal{C}_q = \{p_k\}_{k=1}^{K} \subset \mathcal{D}$. The model then identifies the supporting pages $\mathcal{S}_q \subseteq \mathcal{C}_q$, defining the **page-level attribution**, and localizes evidence regions on each $p \in \mathcal{S}_q$ as $B_p = \{b_{p,j}\}_{j=1}^{m_p}$, where $b_{p,j} = (x_1, y_1, x_2, y_2)$. The **region-level attribution** is therefore $\mathcal{B}_q = \{(p, B_p) \mid p \in \mathcal{S}_q\}$.

2.2 Leader Agent for Global Cross-Page Reasoning

The leader focuses on what evidence is needed and whether it jointly supports an answer, leaving retrieval and page-wise attribution execution to subagents. It uses two actions: `call_subagents`, which triggers retrieval and subsequent page-wise attribution based on subqueries, and `finalize`, which terminates evidence acquisition and produces the final answer with visual attributions.

Query Decomposition and Subagent Invocation. Given a user query q, the leader first analyzes the required evidence and reformulates it into a set of retrieval-oriented subqueries, $Q_q =$

$\{q_0, q_1, \ldots, q_M\}$, where $q_0 = q$ preserves the original query and the remaining subqueries provide complementary retrieval views of the required evidence. As illustrated in Figure 3, the leader decomposes the original query into multiple subqueries targeting complementary aspects and executes `call_subagents`, which triggers parallel retrieval followed by page-wise attribution. In this way, the leader specifies what evidence should be acquired, while retrieval and visual attribution are delegated to specialized subagents and executed outside the leader's reasoning context. We describe these two actions in Section 2.3.

Cross-Page Attribution Selection and Answer Synthesis. After the subagent calls are completed, the leader receives the relevant pages together with their localized supporting regions and assesses whether the collected evidence is sufficient to answer the query. If so, it executes `finalize`, which aggregates and verifies the page-specific evidence to produce the final answer a, along with the selected supporting pages $\mathcal{S}_q$ and their corresponding region-level attributions $\mathcal{B}_q$.

2.3 Parallel Subagents for Retrieval and Page-Wise Attribution

The subagent layer organizes retrieval by subquery and attribution by page, allowing evidence acquisition to proceed in parallel while keeping local attribution contexts isolated.

Parallel Retrieval. Given the retrieval-oriented subqueries generated by the leader, each retrieval subagent searches the document corpus for candidate pages relevant to its assigned subquery. As these retrieval tasks target complementary evidence and can be executed independently, retrieval proceeds in parallel rather than through a sequential single-agent trajectory. The retrieved pages are then collected and deduplicated by the externalized harness, with their associated subqueries preserved for subsequent attribution.

Isolated Page-wise Attribution. Each attribution subagent receives a single candidate page p_i and its corresponding subquery $\tilde{q}_i \in Q_q$, assesses the page's relevance to $\tilde{q}_i$, and localizes supporting regions when the page is relevant. Formally, $(\tilde{q}_i, p_i) \rightarrow (r_i, B_i)$, where $r_i \in \{0, 1\}$ indicates page relevance to $\tilde{q}_i$, and $B_i = \{b_{i,1}, \ldots, b_{i,m_i}\}$ contains the localized evidence boxes when $r_i = 1$; otherwise, $B_i = \emptyset$. The externalized harness collects these outputs and forwards only pages with $r_i = 1$, along with their localized regions, to the leader for cross-page reasoning and synthesis. By operating on a single candidate page within an isolated context, each subagent limits cross-page context pollution while enabling parallel attribution across multiple pages.

2.4 Externalized Harness for State and Evidence Management

To support the coordination between the leader and parallel subagents, AttriSwarm employs an externalized harness as a programmatic state-management layer. It maintains four types of externalized state: (1) a `Query Registry` that stores the original query and its decomposed subqueries, (2) a `Retrieved Page Pool` that collects and deduplicates retrieved pages while preserving their retrieval scores and provenance information, (3) a `Dispatch Table` that tracks the assignment of candidate pages to attribution subagents, and (4) an `Evidence Ledger` that records page relevance assessments and localized supporting regions returned by the subagents. By externalizing these workflow states, the harness keeps intermediate retrieval and attribution information outside the agents' reasoning contexts, ensuring both the leader and each subagent access only the information required for its current stage, while preserving explicit evidence throughout the workflow.

3 AttriDoc Construction

Cross-page visual attribution requires page- and region-level evidence annotations. **AttriDoc** comprises 5,471 training and 1,236 test QA pairs across **AttriDoc-Paper** and **AttriDoc-Slide**, annotated with supporting pages and evidence boxes.

3.1 Data Sources and Annotation

AttriDoc-Paper. This subset is constructed from computer science papers for cross-page reasoning over experimental results. We collect seed papers from arXiv and their cited papers, parsing

PDF and HTML representations to extract experimental tables. GPT-5.1 (Singh et al., 2026) groups evidence across papers by shared entities (methods, datasets, and metrics) and synthesizes QA pairs that require cross-page evidence integration from related tables and page images. We extract table boxes using MinerU (Wang et al., 2026a), where source pages and boxes serve as page- and region-level attributions, respectively.

AttriDoc-Slide. This subset is built from the multipage QA pairs in SlideVQA (Tanaka et al., 2023). We retain the original questions, answers, and supporting-pages and augment them with region-level bounding boxes. Following AttriSwarm, we use Gemini-3-Flash-Preview (Comanici et al., 2025) as the attribution subagent to localize supporting regions in trajectory generation.

3.2 Quality Control and Statistics

Table 1: AttriDoc dataset statistics.

Subset	**Train QA**	**Test QA**	**Pages**
Paper	4,079	669	42,306
Slide	1,392	567	19,749
Total	**5,471**	**1,236**	**62,055**

Evidence pages	**1**	**2**	**3**	**4**	**5**	**6**
Count	3	830	193	137	67	6
Share (%)	0.2	67.2	15.6	11.1	5.4	0.5

Quality Control. For AttriDoc-Paper, we retain QA pairs only when GPT-5.1 (Singh et al., 2026), Gemini-3-Flash-Preview (Comanici et al., 2025), and Claude Sonnet 4.6 (Anthropic, 2024) unanimously agree that the answers are correct based on the questions and cropped evidence images. We further manually filter the test set for: (1) clearly specified information requests; (2) unambiguous references to methods, datasets, and metrics; (3) necessary and sufficient evidence; and (4) answer values that are consistent with the corresponding tables. For AttriDoc-Slide, GPT-5.1 evaluates whether the cropped evidence supports the answer, covers the required information, and avoids excessive irrelevant content. Failed annotations undergo up to three rounds of revision, and cases that still fail are discarded.

Dataset Statistics. Table 1 summarizes the dataset scale and the distribution of distinct annotated evidence pages per test question. After quality control, AttriDoc-Paper contains 4,079 training and 669 test QA pairs, with a retrieval corpus of 42,306 pages from 2,359 papers. AttriDoc-Slide contains 1,392 training and 567 test QA pairs, with a retrieval corpus of 19,749 slide images.

4 Role-Specific Training

While AttriSwarm provides the architectural decomposition needed to isolate and parallelize page-wise attribution, general-purpose MLLMs are not inherently optimized for the specialized roles introduced by this decomposition. In particular, false-negative relevance decisions may discard useful pages, while inaccurate localization degrades the evidence returned to the leader. We therefore develop role-specific training for both the leader and attribution subagents.

4.1 Trajectory-Based Supervised Fine-Tuning

We perform SFT separately for the leader and attribution subagents using role-specific interaction trajectories. The trajectories are constructed as follows. For both AttriDoc subsets, GPT-5.1 serves as the leader. It generates retrieval-oriented subqueries and invokes subagents to retrieve candidate pages. To avoid redundant retrieval, we discard any subquery whose retrieved pages are fully covered by other retained queries. Each retrieved page is then paired with the first retained query that retrieves it, with the original user query ordered first. Page-wise attribution responses are derived from AttriDoc-Paper annotations, or generated by Gemini-3-Flash-Preview for AttriDoc-Slide. The leader then integrates the returned evidence and invokes finalize to produce an answer with selected visual attributions. These trajectories provide role-specific supervision. Leader trajectories supervise query decomposition, subagent invocation, evidence aggregation, and final response generation. Attribution trajectories supervise page relevance assessment and supporting-region localization.

4.2 Attribution-Aware Reinforcement Learning

After SFT, both roles are optimized using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), with separate rewards aligned with their responsibilities.

Attribution Subagent Optimization. We jointly optimize both page-level relevance assessment and supporting-region localization via a unified reward:

$$R_{\text{sub}} = \lambda_r R_{\text{region}} + \lambda_{\text{rel}} R_{\text{rel}} + \lambda_f R_{\text{fmt}}, \tag{2}$$

where R_{region} uses $\text{IoU} = |B_p \cap B_g|/|B_p \cup B_g|$, with B_p and B_g denoting the geometric unions of predicted and reference evidence boxes on a page, respectively. R_{rel} measures the correctness of page relevance decisions, penalizing missed evidence on relevant pages and spurious predictions on irrelevant pages. R_{fmt} penalizes outputs violating the required attribution format.

Leader Agent Optimization. The leader is optimized for answer correctness, page-level attribution, and output validity:

$$R_{\text{leader}} = \mathbb{I}_{\text{valid}} \left(\lambda_a R_{\text{ans}} + \lambda_p R_{\text{page}} + \lambda_{ap} R_{\text{ans}} R_{\text{page}} \right), \tag{3}$$

where R_{ans} evaluates answer correctness via an LLM-based verifier, and R_{page} is the F_1 score between the predicted supporting pages and the retrieved reference pages, scaled by the fraction of all reference pages successfully retrieved, and $\mathbb{I}_{\text{valid}}$ indicates whether the final output satisfies the required structured format.

5 Experiments and Analysis

We evaluate on AttriDoc-Paper and AttriDoc-Slide as in domain settings, and on 1,175 out of domain CiteVQA questions (Ma et al., 2026). The CiteVQA subset contains questions with accessible source PDFs and is shared across all methods.

Evaluation Metrics. We evaluate answer quality and evidence attribution using four metrics: (1) Answer Accuracy (Acc.) measures final answer correctness with an LLM judge; (2) Page Attribution F_1 (Page F_1) evaluates whether the correct supporting pages are identified; (3) Region Attribution IoU (Reg. IoU) measures fine-grained evidence localization, averaged over all ground-truth supporting pages so that missed pages are also penalized; and (4) Joint Attribution Score (JAS) combines page selection and region grounding:

$$\text{JAS} = \frac{1}{N} \sum_{q=1}^{N} F_{1,q} \cdot \text{IoU}_q. \tag{4}$$

Here, N is the number of questions, and $F_{1,q}$ and IoU_q are the page and region attribution scores for question q. All metrics are averaged across questions and reported as percentages. Together, these metrics measure answer correctness and coarse-to-fine attribution quality.

Baselines. We compare AttriSwarm with two attribution paradigms (Section 1): (1) Single-pass jointly processes the top-K retrieved pages ($K \in \{10, 20\}$) in a shared multimodal context, producing the answer and page-indexed evidence boxes in one call; (2) Interactive attribution includes training-free ReAct, which alternates reasoning and sequential search within a single trajectory, with final outputs adapted to include the answer and page-indexed evidence boxes. We also evaluate released checkpoints of VRAG-RL (Wang et al., 2025b), UniDoc-RL (Wang et al., 2026b), and MCite-RL (Zhao et al., 2026a), preserving their search and crop protocols. Single-pass and ReAct use Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct, Qwen2.5-VL-7B-Instruct, GPT-5.1, and Claude Sonnet 4.6. We also include training-free AttriSwarm with the same architecture and no task-specific training (details in Appendix D).

5.1 Main Results

Overall Performance. As shown in Table 2, AttriSwarm improves both answer correctness and evidence attribution. For example, with Qwen3-VL-8B-Instruct on AttriDoc-Slide, it raises answer accuracy from 57.14% to 68.43% and JAS from 6.51% to 45.86% over the strongest single-pass baseline. The gains in Page F_1 and Reg. IoU further indicate stronger coarse-to-fine evidence attribution and better consistency between page selection and region attribution. Overall, these results demonstrate that AttriSwarm improves both cross-page reasoning and end-to-end attribution.

Table 2: Main results on AttriDoc-Paper, AttriDoc-Slide, and CiteVQA. All metric values are percentages. CiteVQA attribution is evaluated against all evidence pages, including both necessary and non-necessary evidence. Training-free methods are marked with †. Within each backbone-model block, bold and underlined values denote the best and second-best reported results, respectively.

Method	AttriDoc-Paper				AttriDoc-Slide				CiteVQA			
	Acc.	Page F_1	Reg. IoU	JAS	Acc.	Page F_1	Reg. IoU	JAS	Acc.	Page F_1	Reg. IoU	JAS
Qwen3-VL-4B-Instruct												
Single-pass ($k = 10$)†	3.29	44.58	9.64	6.03	52.56	61.67	7.70	5.92	30.98	43.11	13.11	10.14
Single-pass ($k = 20$)†	3.14	39.52	8.17	4.62	52.91	56.81	8.29	6.13	28.26	38.96	13.02	9.48
ReAct†	0.60	17.47	1.32	0.71	30.51	30.82	2.14	1.43	13.19	19.67	2.25	1.92
AttriSwarm	**25.86**	**62.53**	**47.91**	**37.40**	**64.73**	**71.91**	**46.57**	**40.71**	**45.61**	**53.80**	**31.30**	**24.07**
– AttriSwarm†	2.09	14.45	7.62	4.18	44.27	36.09	19.43	11.76	38.64	43.52	20.71	15.03
Qwen3-VL-8B-Instruct												
Single-pass ($k = 10$)†	5.08	43.81	5.74	3.75	57.14	61.16	8.48	6.51	36.60	46.02	11.79	9.50
Single-pass ($k = 20$)†	4.04	38.82	5.02	3.27	56.26	59.32	8.38	6.29	36.43	43.89	11.38	9.06
ReAct†	0.45	19.71	0.74	0.43	31.04	36.48	2.75	1.83	19.74	23.19	2.51	2.10
AttriSwarm	**25.86**	**63.65**	**49.52**	**38.96**	**68.43**	**75.25**	**52.00**	**45.86**	**52.68**	**54.77**	**32.70**	**25.52**
– AttriSwarm†	4.65	31.58	16.69	10.47	45.50	37.49	18.45	11.47	44.51	50.10	26.38	20.51
Qwen2.5-VL-7B-Instruct												
Single-pass ($k = 10$)†	2.99	38.10	1.24	0.62	46.91	46.65	2.80	1.80	26.38	31.36	0.28	0.18
Single-pass ($k = 20$)†	2.39	36.59	0.80	0.45	45.68	45.38	1.76	1.15	23.74	29.22	0.27	0.15
ReAct†	0.45	17.72	0.17	0.10	28.92	35.50	1.36	0.92	15.83	20.17	1.35	1.14
VRAG-RL	2.39	0.56	0.06	0.03	44.27	4.35	0.65	0.44	17.87	3.73	0.35	0.28
UniDoc-RL	0.75	22.07	10.33	5.95	20.28	22.05	8.55	5.69	19.83	16.39	5.70	4.99
MCite-RL	1.05	24.13	6.51	3.84	33.51	49.79	15.90	10.59	14.98	30.48	5.92	5.03
AttriSwarm	**13.60**	**47.93**	**25.54**	**18.53**	**60.67**	**69.90**	**30.51**	**26.11**	**36.68**	**49.00**	**12.95**	**9.91**
– AttriSwarm†	1.20	8.30	1.13	0.65	30.51	34.21	9.00	5.60	12.26	28.90	2.20	1.73
Proprietary Models												
# GPT-5.1												
Single-pass ($k = 10$)†	8.97	39.24	11.99	7.35	65.90	**72.84**	18.85	16.36	53.11	50.69	11.89	9.35
Single-pass ($k = 20$)†	10.76	36.72	11.09	6.52	59.61	54.45	14.92	12.73	54.72	45.62	11.60	8.92
ReAct†	0.90	25.93	5.09	3.02	49.03	52.23	12.86	9.03	34.47	34.21	6.79	5.75
AttriSwarm†	**26.31**	**62.98**	**24.89**	**19.08**	**72.44**	63.48	**25.36**	**19.32**	**66.30**	**58.23**	**16.24**	**12.08**
# Claude Sonnet 4.6												
Single-pass ($k = 10$)†	32.44	61.94	9.48	7.18	72.49	**78.14**	15.03	13.05	53.87	53.94	**14.92**	**12.17**
Single-pass ($k = 20$)†	25.71	51.21	10.05	7.05	51.32	45.94	7.93	6.22	31.20	28.30	5.15	3.62
ReAct†	40.23	54.48	8.74	7.43	**78.84**	69.84	21.99	19.15	55.40	43.02	9.54	8.40
AttriSwarm†	**45.96**	**70.12**	**10.78**	**8.49**	65.26	54.70	**26.09**	**18.12**	**62.47**	**57.82**	8.14	6.10

Effectiveness of Role-Specific Training. Compared with its training-free counterpart, trained AttriSwarm consistently improves all four metrics across the nine open-source backbone–dataset combinations. Averaged over these combinations, role-specific training improves answer accuracy and JAS by 18.94 and 20.63 percentage points, respectively. Improvements on out-of-domain CiteVQA further suggest that the learned attribution capabilities transfer beyond the training domains.

5.2 Ablation Study

To disentangle the contributions of training each agent role, we independently vary whether each agent role is trained (Table 3). Training attribution agents alone improves all four metrics across all settings. For example, on AttriDoc-Paper with Qwen3-VL-4B-Instruct, it raises page-level F_1 from 14.45% to 60.33% and region-level IoU from 7.62% to 49.31%. Leader training yields limited or inconsistent attribution gains on its own, but improves answer accuracy across all nine backbone–dataset settings when the attribution agents are trained. These results suggest complementary benefits: attribution agent training strengthens evidence grounding, while leader training helps translate that evidence into more accurate answers. We also evaluate different model assignments for the leader and attribution agents in Appendix C.

Table 3: Role-specific training ablation across three datasets and three model backbones. All results are reported as percentages.

Agent Training		AttriDoc-Paper				AttriDoc-Slide				CiteVQA			
Leader	**Attribution**	**Acc.**	**Page F_1**	**Reg. IoU**	**JAS**	**Acc.**	**Page F_1**	**Reg. IoU**	**JAS**	**Acc.**	**Page F_1**	**Reg. IoU**	**JAS**
Qwen2.5-VL-7B-Instruct													
Base	Base	1.20	8.30	1.13	0.65	30.51	34.21	9.00	5.60	12.26	28.90	2.20	1.73
Base	Trained	3.74	43.45	**26.26**	16.69	50.44	59.67	28.96	22.58	29.11	47.82	**13.61**	9.77
Trained	Base	1.35	7.63	1.13	0.66	32.80	24.68	10.04	6.62	18.81	23.59	2.18	1.71
Trained	Trained	**13.60**	**47.93**	25.54	**18.53**	**60.67**	**69.90**	**30.51**	**26.11**	**36.68**	**49.00**	12.95	**9.91**
Qwen3-VL-4B-Instruct													
Base	Base	2.09	14.45	7.62	4.18	44.27	36.09	19.43	11.76	38.64	43.52	20.71	15.03
Base	Trained	15.70	60.33	**49.31**	35.97	59.08	62.16	44.45	35.06	43.06	51.14	**31.87**	**24.08**
Trained	Base	1.64	15.51	7.43	4.48	49.38	48.65	20.60	14.66	41.70	46.75	20.06	14.96
Trained	Trained	**25.86**	**62.53**	47.91	**37.40**	**64.73**	**71.91**	**46.57**	**40.71**	**45.61**	**53.80**	31.30	24.07
Qwen3-VL-8B-Instruct													
Base	Base	4.65	31.58	16.69	10.47	45.50	37.49	18.45	11.47	44.51	50.10	26.38	20.51
Base	Trained	17.79	61.29	47.89	36.11	64.78	66.31	47.77	38.43	47.57	**55.50**	**34.69**	**27.14**
Trained	Base	6.13	34.95	18.71	11.65	54.14	45.21	19.32	13.67	50.52	51.60	25.97	19.98
Trained	Trained	**25.86**	**63.65**	**49.52**	**38.96**	**68.43**	**75.25**	**52.00**	**45.86**	**52.68**	54.77	32.70	25.52

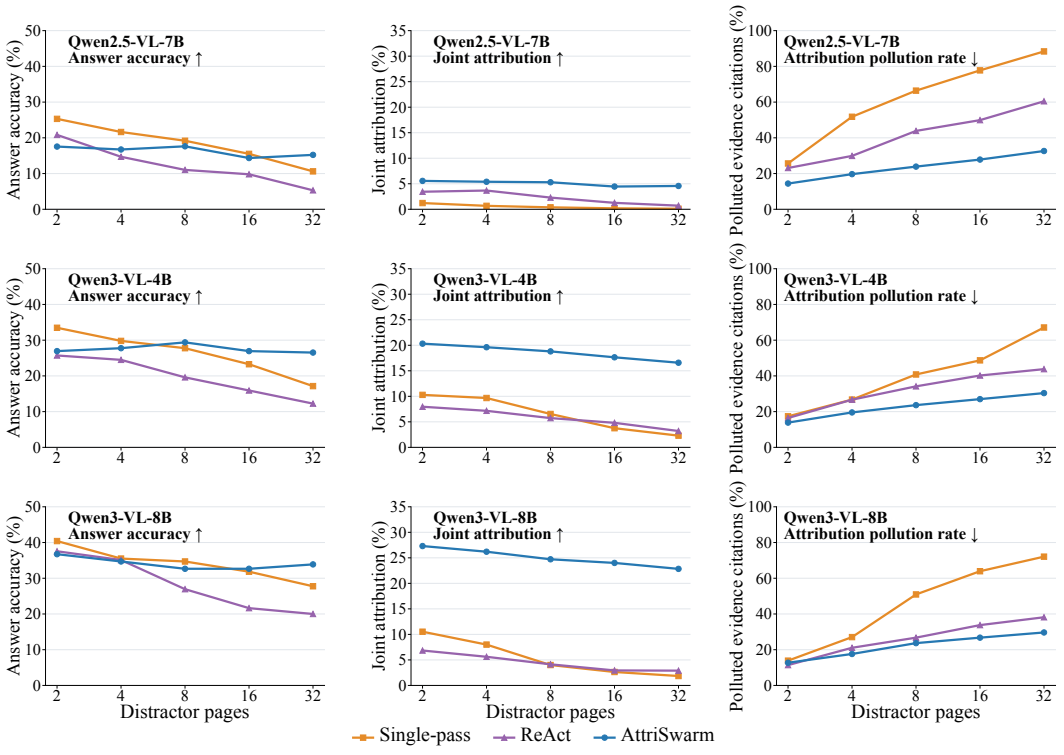

Figure 4: Robustness to increasing numbers of distractor pages on CiteVQA across three backbones.

5.3 Further Analysis

Robustness to Cross-page Context Pollution. Building on the diagnostic in the introduction, we add 2–32 retrieved distractor pages to 245 CiteVQA questions, retaining all annotated evidence pages and matching page order across methods. We measure accuracy, JAS, and attribution pollution (the mean fraction of distractors among mapped cited pages), excluding empty outputs and error cases. Across three backbones, the mean accuracy drop over this range is 1.86 percentage points for AttriSwarm versus 14.56 and 15.51 for Single-pass and ReAct, respectively (Figure 4). AttriSwarm's attribution pollution grows more slowly; despite declining with noise, its JAS remains highest at 32 distractors across all three backbones. These results suggest that context isolation and evidence filtering before global aggregation help mitigate cross-page context pollution.

Inference Efficiency. To examine whether parallel attribution reduces the sequential overhead identified in the introduction, we compare parallel AttriSwarm, its serial variant, and ReAct on the same frozen CiteVQA page sets using Qwen3-VL-4B-Instruct, measuring mean end-to-end latency over successful runs. As shown in Figure 5, at 32 distractor pages, parallel AttriSwarm takes 15.99 seconds, achieving approximately $4.1\times$ and $5.1\times$ speedups over its serial variant and ReAct, respectively. The comparison with the serial variant highlights the benefit of parallel attribution in reducing sequential overhead.

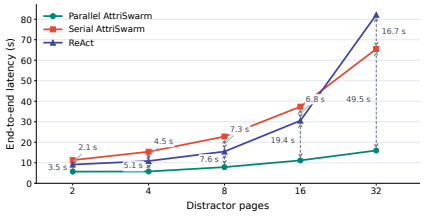

Figure 5: Efficiency comparison.

Answer-Supporting Attribution. Since region-level IoU does not directly measure answer support, we use an LLM judge to assess the support and redundancy of cited evidence. As shown in Figure 6, AttriSwarm increases the proportion of answers with fully supporting, non-redundant evidence by 18.1 to 27.5 percentage points over single-pass.

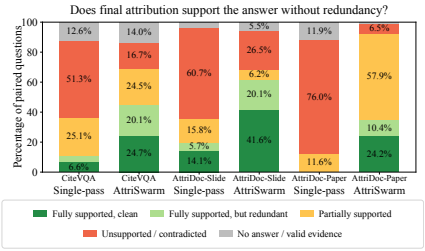

Figure 6: Analysis of whether the attribution regions support the answer.

6 Related Work

Visual Attribution and Evidence Grounding. Visual attribution links model outputs to supporting visual evidence for transparent and verifiable generation. Prior work has studied visual source attribution (Ma et al., 2025), visual citation (Zhao et al., 2025), and grounding-oriented reinforcement learning (Wang et al., 2026b; Zhao et al., 2026a). However, these methods mainly target single-page or limited visual contexts, whereas cross-page attribution requires both supporting-page identification and fine-grained evidence localization across documents.

Agentic Multimodal Reasoning. Agentic reasoning enables foundation models to iteratively interact with tools and external information (Yao et al., 2023; Jin et al., 2025), and has been extended to visual interaction and iterative document retrieval (Zheng et al., 2025; Wang et al., 2025a; 2026c; Shen et al., 2026; Wang et al., 2025b). Recent evidence-aware agents further emphasize evidence acquisition, grounding, and verification, from text-centric reasoning (Zhang et al., 2026b; Zhao et al., 2026b; Sun et al., 2025b) to fine-grained visual evidence reasoning (Wang et al., 2026b). However, sequential trajectories accumulate evidence in shared contexts, hindering cross-page coordination.

Swarm-based Multi-Agent Systems. Early multi-agent frameworks explored coordinating interacting or role-specialized agents for complex tasks (Wu et al., 2023; Li et al., 2023; Hong et al., 2023). More recently, Kimi K2.5 (Team, 2026) demonstrated an agent-swarm paradigm, while SwarmBench (Gao et al., 2026) evaluated LLMs' orchestration capabilities. Other works explored swarm-style orchestration for broad information seeking (Xu et al., 2026), deep-and-wide web search (Song et al., 2026), and open-ended discovery (Virk et al., 2026). However, these general-purpose swarms do not explicitly address the coordination challenge in cross-page attribution between global cross-page reasoning and isolated page-wise evidence localization.

7 Conclusion

To address the challenges of cross-page visual attribution, we introduced AttriSwarm, an agent-swarm framework that decouples global reasoning from page-wise visual attribution. A leader agent coordinates attribution subagents in isolated contexts to mitigate cross-page context pollution, while page-wise attribution is parallelized to reduce sequential overhead. We also constructed AttriDoc, comprising paper and slide subsets with supporting-page and fine-grained region annotations, and developed role-specific training. Experiments show improved answer accuracy and attribution, along with robustness to distractor pages, lower overhead, and stronger evidence support. These findings highlight the value of separating and coordinating global reasoning and local attribution for reliable cross-page visual attribution and evidence-grounded multimodal document understanding.

AI USE STATEMENT

In this work, we used generative AI tools for entity extraction, evidence grouping, and cross-page question generation in AttriDoc-Paper, as well as visual attribution annotation in AttriDoc-Slide. We also used GPT-5.1 to synthesize leader trajectories and Gemini-3-Flash-Preview to generate attribution-subagent outputs for AttriDoc-Slide. Additionally, we used generative AI tools to polish the manuscript for clarity and readability. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses scientific papers and existing SlideVQA data to study cross-page visual attribution. We reviewed the AI-generated data to reduce potential errors and biases. Visual attribution further facilitates evidence-based verification of generated answers, but does not guarantee their correctness; answers should therefore be checked against the source documents. Any release or reuse of AttriDoc should respect the licenses and usage restrictions of the underlying materials and provide appropriate source attribution.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the AttriSwarm architecture and agent workflow in Section 2, the construction of AttriDoc and role-specific training trajectories in Section 3, and the supervised fine-tuning and reinforcement learning objectives in Section 4. Section 5 specifies the evaluation datasets, metrics, baselines, and model backbones used in our experiments.

REFERENCES

Anthropic. Claude 3.5 sonnet raises the industry bar for intelligence. https://www.anthropic.com/news/claude-3-5-sonnet, June 2024. Accessed: 2025-09-07.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2024. URL https://arxiv.org/abs/2407.01449.

Jinshan Gao, Zhuoran Jin, Tianyi Men, Kang Liu, and Jun Zhao. Swarmbench: Can large language models act as agent swarm orchestrators?, 2026. URL https://arxiv.org/abs/2608.30661.

Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zi Hen Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. Metagpt: Meta programming for multi-agent collaborative framework. *ArXiv*, abs/2308.00352, 2023. URL https://api.semanticscholar.org/CorpusID:260351380.

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. *CoRR*, abs/2601.12346, 2026a. doi: 10.48550/ARXIV.2601.12346. URL https://doi.org/10.48550/arXiv.2601.12346.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Yao Hu, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. *CoRR*, abs/2601.22060, 2026b. doi: 10.48550/ARXIV.2601.22060. URL https://doi.org/10.48550/arXiv.2601.22060.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Xu Tang, Yao Hu, Shaohui Lin, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models, 2026c. URL https://arxiv.org/abs/2601.22060.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025. doi: 10.48550/ARXIV.2503.09516. URL https://doi.org/10.48550/arXiv.2503.09516.

G. Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems 36*, 2023. URL https://api.semanticscholar.org/CorpusID:268042527.

Rongjin Li, Yuanxin Liu, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Not just reason, not just scan: Reinforcement learning for proactive scientific error verification over academic paper, 2026. URL https://arxiv.org/abs/2608.26596.

Dongsheng Ma, Jiayu Li, Zhengren Wang, Yijie Wang, Jiahao Kong, Weijun Zeng, Jutao Xiao, Jie Yang, Wentao Zhang, Bin Wang, et al. Citevqa: Benchmarking evidence attribution for trustworthy document intelligence. *arXiv preprint arXiv:2605.12882*, 2026.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. VISA: retrieval augmented generation with visual source attribution. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 30154–30169. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1456. URL https://doi.org/10.18653/v1/2025.acl-long.1456.

Cara Phillips. Building multi-agent systems: When and how to use them. Anthropic Blog, January 2026. URL https://claude.com/blog/building-multi-agent-systems-when-and-how-to-use-them.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Yucheng Shen, Jiulong Wu, Jizhou Huang, Dawei Yin, Lingyong Yan, and Min Cao. Visor: Agentic visual retrieval-augmented generation via iterative search and over-horizon reasoning. *arXiv preprint arXiv:2604.09508*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Y. Guan, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam

Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomek Korbak, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Xiaoshuai Song, Liancheng Zhang, Kangzhi Zhao, Yutao Zhu, Zhongyuan Wang, Guanting Dong, Jinghan Yang, Han Li, Kun Gai, Ji-Rong Wen, and Zhicheng Dou. Webswarm: Recursive multi-agent orchestration for deep-and-wide web search. *CoRR*, abs/2607.08662, 2026. doi: 10.48550/ARXIV.2607.08662. URL https://doi.org/10.48550/arXiv.2607.08662.

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, Jianing Wang, Qintong Li, Xiangru Tang, Tianbao Xie, Xiachong Feng, Xiang Li, Ben Kao, Wenhai Wang, Biqing Qi, Lingpeng Kong, and Zhiyong Wu. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. *CoRR*, abs/2505.19897, 2025a. doi: 10.48550/ARXIV.2505.19897. URL https://doi.org/10.48550/arXiv.2505.19897.

Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Chi Chen, Zhiyuan Liu, and Maosong Sun. Visrag 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. *CoRR*, abs/2510.09733, 2025b. doi: 10.48550/ARXIV.2510.09733. URL https://doi.org/10.48550/arXiv.2510.09733.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13636–13645. AAAI Press, 2023. doi: 10.1609/AAAI.V37I11.26598. URL https://doi.org/10.1609/aaai.v37i11.26598.

Kimi Team. Kimi K2.5: visual agentic intelligence. *CoRR*, abs/2602.02276, 2026. doi: 10.48550/ARXIV.2602.02276. URL https://doi.org/10.48550/arXiv.2602.02276.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yuvraj Virk, Zack Edds, Chunqiu Steven Xia, and Lingming Zhang. Swarmresearch: Orchestrating coding agents for open-ended discovery, 2026. URL https://arxiv.org/abs/2607.02807.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*, 2026a.

Jun Wang, Shuo Tan, Zelong Sun, Tiancheng Gu, Yongle Zhao, Ziyong Feng, Kaicheng Yang, and Zhiwu Lu. Unidoc-rl: Coarse-to-fine visual RAG with hierarchical actions and dense rewards. *CoRR*, abs/2604.14967, 2026b. doi: 10.48550/ARXIV.2604.14967. URL https://doi.org/10.48550/arXiv.2604.14967.

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. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 9113–9134. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.EMNLP-MAIN.464. URL https://doi.org/10.18653/v1/2025.emnlp-main.464.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. VRAG-RL: empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/529b876b5d87f997e98ed925a3ce9f01-Abstract-Conference.html.

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. *CoRR*, abs/2602.24134, 2026c. doi: 10.48550/ARXIV.2602.24134. URL https://doi.org/10.48550/arXiv.2602.24134.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. 2023. URL https://api.semanticscholar.org/CorpusID:263611068.

Lei Xiong, Huaying Yuan, Zheng Liu, Zhao Cao, and Zhicheng Dou. Paperscope: A multimodal multi-document benchmark for agentic deep research across massive scientific papers. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 8015–8040. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.394. URL https://doi.org/10.18653/v1/2026.findings-acl.394.

Zelai Xu, Zhexuan Xu, Ruize Zhang, Chunyang Zhu, Shi Yu, Weilin Liu, Quanlu Zhang, Wenbo Ding, Chao Yu, and Yu Wang. Wideseek-r1: Exploring width scaling for broad information seeking via multi-agent reinforcement learning. *CoRR*, abs/2602.04634, 2026. doi: 10.48550/ARXIV.2602.04634. URL https://doi.org/10.48550/arXiv.2602.04634.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, and Wei Chen. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework, 2025. URL https://arxiv.org/abs/2506.02454.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. Mm-deepresearch: A simple and effective multimodal agentic search baseline. *CoRR*, abs/2603.01050, 2026. doi: 10.48550/ARXIV.2603.01050. URL https://doi.org/10.48550/arXiv.2603.01050.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Xiaohan Yu, Chao Feng, Lang Mei, and Chong Chen. M^3searcher: Modular multimodal information seeking agency with retrieval-oriented reasoning. *CoRR*, abs/2601.09278, 2026. doi: 10.48550/ARXIV.2601.09278. URL https://doi.org/10.48550/arXiv.2601.09278.

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

Fan Zhang, Vireo Zhang, Shengju Qian, Haoxuan Li, Hao Wu, Jinyang Wu, Donghao Zhou, Zhihong Zhu, Zheng Lian, Xin Wang, and Pheng-Ann Heng. Orchestra-o1: Omnimodal agent orchestration, 2026a. URL https://arxiv.org/abs/2606.13707.

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou, and Juanzi Li. Chaining the evidence: Robust reinforcement learning for deep search agents with citation-aware rubric rewards. *CoRR*, abs/2601.06021, 2026b. doi: 10.48550/ARXIV.2601.06021. URL https://doi.org/10.48550/arXiv.2601.06021.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. Finragbench-v: A benchmark for multimodal RAG with visual citation in the financial domain. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 4215–4249. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.211. URL https://doi.org/10.18653/v1/2025.emnlp-main.211.

Suifeng Zhao, Zida Liu, Xinyu Lei, Lei Sun, Jun Gao, and Sujian Li. Mcite-rl: Towards reliable multimodal rag via citation-enhanced agentic reinforcement learning, 2026a. URL https://arxiv.org/abs/2608.21808.

Yibo Zhao, Jiapeng Zhu, Zichen Ding, and Xiang Li. GRACE: reinforcement learning for grounded response and abstention under contextual evidence. *CoRR*, abs/2601.04525, 2026b. doi: 10.48550/ARXIV.2601.04525. URL https://doi.org/10.48550/arXiv.2601.04525.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *ArXiv*, abs/2505.14362, 2025. URL https://api.semanticscholar.org/CorpusID:278769859.

A Qualitative Case Studies

Figures 7 and 8 illustrate how answer errors can arise at different stages of evidence attribution. We use these examples to examine the relationship between each method's answer and its predicted evidence regions.

Cross-paper evidence selection. Figure 7 asks for MMSSL's Recall@20 on Baby as reported in two different papers. The relevant tables report 0.0929 in STAIR and 0.0971 in the Modality-Independent GNN paper. AttriSwarm identifies both tables and returns the corresponding values. Single-pass instead returns 0.0957 and 0.0680; its two predicted regions are placed on another page of the STAIR paper rather than on the two required evidence pages. The example highlights the need to preserve the provenance of each value in a cross-paper comparison.

Question:
STAIR and the Modality-Independent GNN paper both report MMSSL on Baby. What Recall@20 value does each paper report?

Answer:
1. AttriSwarm: On the Baby dataset, MMSSL is reported to achieve: 0.0929 Recall@20 in the STAIR paper. 0.0971 Recall@20 in the Modality-Independent GNN paper.

2. Single-pass: The STAIR paper reports a Recall@20 value of 0.0957, while the Modality-Independent GNN paper reports a Recall@20 value of 0.0680.

Attribution:

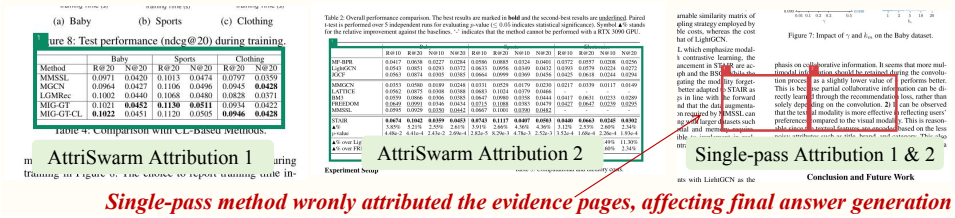

Figure 7: Cross-paper evidence selection. AttriSwarm attributes the two reported values to their respective papers; Single-pass returns different values and places its predicted regions on another STAIR page.

Question:
In the Third Year of their 2nd Presidential Term, who had the highest Job Approval Ratings in the month when Clinton had 44.4% support according to HuffPost Pollster Trend?

Answer:
1. AttriSwarm: Eisenhower
2. Single-pass: In the Third Year of their 2nd Presidential Term, Barack Obama had the highest job approval ratings in the month when Clinton had 44.4% support, as indicated by the HuffPost Pollster Trend data for October 2015.

Attribution:

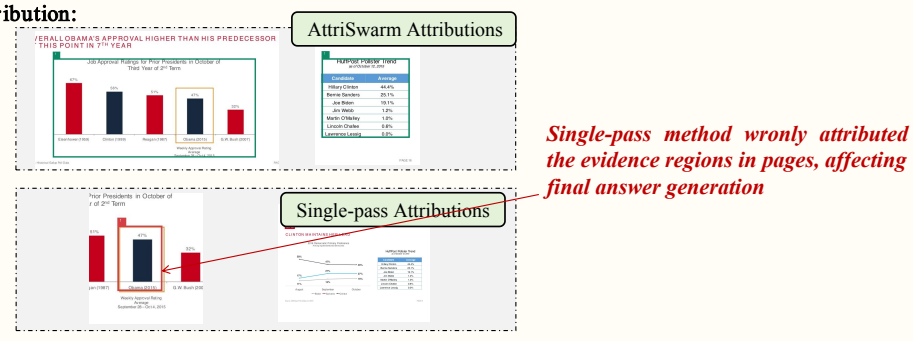

Figure 8: Evidence selection within a slide. Although Single-pass places a box on the relevant approval chart, it focuses on Obama rather than the highest bar, Eisenhower.

Evidence selection within a page. In Figure 8, Clinton's 44.4% support identifies October 2015. The corresponding approval chart shows Eisenhower at 67% and Obama at 47%. AttriSwarm returns Eisenhower, whereas Single-pass returns Obama and places a predicted region around Obama's bar. Here, the error is more subtle: Single-pass reaches a relevant chart but does not select the bar that answers the question.

B Prompts for Data Synthesis

This appendix reports the representative prompts when constructing AttriDoc, including question synthesis, evidence-region annotation, and the prompt for leader and attribution to do trajectory synthesis.

Table 4: Prompt for Synthesizing QA Pairs for AttriDoc-Paper.

System:
You synthesize simple, objective, multi-page question-answer pairs from experimental tables. Use only the supplied evidence. Return one JSON object.
Input:
Shared entity and its role: {shared_entity}, {method_or_dataset}
Required evidence pages: {num_pages}
Question type: {question_type}
Evidence: {paper_names, paper_ids, table_ids, table_text, target_facts}
Requirements:
1. Construct one self-contained question around the shared entity. Each supplied page must contribute a distinct fact needed by the answer.
2. Explicitly identify the method, dataset, experimental setting, and metric for every requested value. The shared entity must have the stated role on every page.
3. Copy values and supporting text from the evidence. If a target fact is supplied, preserve its subject, dataset, metric, and value.
4. For a shared-method question, ask for that method's reported results. For a shared-dataset question, identify the methods evaluated on that dataset.
5. Compare, rank, subtract, or aggregate values only when metrics, units, and settings are compatible. Otherwise ask for the respective values.
6. Use human-readable method names or paper titles to remove ambiguity. Keep paper IDs in structured evidence fields. Do not refer to supplied pages or dataset construction in the question.
7. If the evidence cannot support a coherent question, return status="unsupported".
Output (core fields):

```
{"status": "ok|unsupported", "question": "...", "answer": "...",
 "involved_sources": ["paper_id", "..."],
 "used_tables": ["paper_id::table_id", "..."],
 "per_page_facts": [{"paper_id": "...", "table_id": "...",
 "subject": "...", "dataset": "...", "metric": "...",
 "value": "...", "supporting_snippet": "..."}]}
```

C Flexible Model Allocation in AttriSwarm

To examine AttriSwarm's flexibility in model allocation, we compare all four assignments of Qwen3-VL-4B-Instruct and GPT-5.1 to the leader and attribution subagents on CiteVQA (Table 8). Using GPT-5.1 for both roles achieves the highest answer accuracy and page-level F_1, whereas pairing a GPT-5.1 leader with Qwen subagents yields the highest region-level IoU and JAS. These results demonstrate that AttriSwarm accommodates different backbones across roles, allowing models to be selected independently for different objectives.

Table 8: Model allocation across agent roles on CiteVQA.

Agent Model		Acc.	Attribution		
Leader	Attribution		Page F_1	Reg. IoU	JAS
Qwen3-4B-Instruct	Qwen3-4B-Instruct	38.64	43.52	20.71	15.03
GPT-5.1	Qwen3-4B-Instruct	48.60	47.68	**21.01**	**16.51**
Qwen3-4B-Instruct	GPT-5.1	50.98	53.60	15.46	10.95
GPT-5.1	GPT-5.1	**66.30**	**58.23**	16.24	12.08

Table 5: LLM-as-Judge Prompt for Verifying Question–Answer Pairs in AttriDoc-Paper.

Instruction: You are a strict multimodal reviewer of scientific-table question-answer data.
Input:
Question: {question}
Proposed answer: {answer}
Evidence: {evidence_images_cropped_from_annotated_boxes}
Read the evidence images and select one decision:
- correct: the question is clear and the answer is fully supported.
- answer_error: the question is clear enough, but the answer is wrong or incomplete, uses the wrong row, column, metric, or dataset, or has a scale error. Provide the evidence-based correction.
- question_unclear: the question is ambiguous, asks for an unreported value, mixes incompatible settings, or cannot be judged from the evidence. Explain the issue and suggest a clearer question.
Evidence rules: Use the images as the source of truth. Do not invent values or silently reinterpret the question. State explicitly when a requested value is not reported.
Return JSON only:

```
{"decision": "correct|answer_error|question_unclear",
 "reason": "...", "correct_answer": "...",
 "suggested_question": "...", "evidence_ids": ["image_0", "..."]}
```

Table 6: Prompt for the Leader Agent in Interaction-Trajectory Synthesis.

Instruction: You are the leader agent of a retrieval-grounding-answer data synthesis workflow.
Available actions: At each round, choose exactly one action:
- `search`: output retrieval subqueries only, without performing grounding.
- `finalize`: produce the final answer with page-level attribution selection.
Input:
Original question: {question}
Previously searched queries: {searched_queries}
Retrieved page summaries: {retrieved_pages}
Grounding results and corresponding visualizations: {grounded_pages}
Requirements:
1. If the current evidence is insufficient, choose `search`.
2. Generate retrieval subqueries through rephrasing, aspect-focused queries, or decomposition of the original question. Focus on informative keywords.
3. Do not repeat the original question verbatim as the only subquery. The system also retrieves using the original question in parallel with your subqueries.
4. Do not repeat an identical query unless intentionally broadening or narrowing the search.
5. Base the final answer strictly on the grounded evidence provided. The reference answer is not included in the standard round-wise input.
6. For `finalize`, select the supporting page IDs. Attribution at this stage consists of page selection; do not generate bounding boxes.
7. Return one JSON object only.
Output for search:

```
{"think": "brief explanation of query decomposition",
 "action": "search",
 "subqueries": [{
   "text": "...",
   "objective": "why this query helps recall",
   "query_kind": "rewrite"
 }]}
```

Output for finalization:

```
{"think": "brief evidence-based explanation",
 "action": "finalize",
 "answer": "final answer",
 "attributions": ["page_id_1", "page_id_2"]}
```

Table 7: Prompt for the Attribution Subagent to Annotate Evidence Regions in AttriDoc-Slide.

Instruction: You are given one slide page, a query, and a reference answer that may require other slides.
Input:
Slide image: {image}
Query: {query}
Reference answer: {reference_answer}
Requirements:
1. Identify evidence visible on this page that directly supports the query or supplies a necessary intermediate fact. Do not assume that the whole reference answer must be visible on this page.
2. Return integer bounding boxes [x1, y1, x2, y2] in normalized coordinates from 0 to 1000.
3. Prefer one complete, readable region. Use multiple boxes when the evidence is spatially separated. Avoid thin or fragmented crops.
4. Preserve the context needed to interpret the evidence. When the evidence occurs in a table or chart, include the complete visible table or chart, including its headers, axes, and legend.
5. If there is no usable evidence on this page, return an empty box list. Keep the accompanying explanation brief and grounded in visible content.
Return JSON only:

```
{"think": "brief page-local explanation",
 "boxes": [[x1, y1, x2, y2], ...]}
```

D Implementation Details

D.1 Training and Reward Configuration

We implement supervised fine-tuning with LLaMA-Factory and reinforcement learning with verl, using vLLM for rollout generation. Training uses eight NVIDIA A100 GPUs with 40 GB memory each.

Training hyperparameters. The leader and attribution subagent are initialized from their respective SFT checkpoints and then optimized separately with GRPO. Tables 9 and 10 summarize the available GRPO and SFT configurations, respectively. The reported GRPO epochs and steps refer to configured training budgets.

Table 9: Hyperparameters for GRPO.

Setting	Leader	Attribution subagent
Learning rate	10^{-6}	10^{-6}
Optimizer	AdamW	AdamW
Training batch size	8 questions	32 page–query pairs
GRPO group size	4	4
KL-loss coefficient	0.01	0.05
Configured epochs	1	1
Planned training steps	321	429

Table 10: Hyperparameters for SFT.

Setting	Leader	Attribution subagent
Learning rate	10^{-5}	10^{-5}
Training epochs	1	1
Per-device batch size	1	1
Gradient accumulation steps	2	2
Learning-rate scheduler	Cosine	Cosine
Warmup ratio	0.03	0.03
Trainable modules	Language model	Language model

Leader reward. Table 11 summarizes the Leader reward configuration. A valid final output must be parseable and contain a non-empty answer and a non-empty attribution list; otherwise, its reward is zero. Before computing supporting-page F1, predicted references are mapped to the annotation identifier space and deduplicated. Page-attribution scoring computes F1 against the retrieved reference pages and scales it by the fraction of all reference pages successfully retrieved.

Table 11: Leader reward configuration.

Setting	**Value**
Answer-correctness weight λ_a	0.6
Page-attribution weight λ_p	0.3
Answer–attribution interaction weight λ_{ap}	0.1
Reward for an invalid final output	0
Answer-verifier temperature	0

Attribution subagent reward. Table 12 specifies the reward for page relevance and region localization. IoU is computed between the geometric unions of predicted and reference boxes on the page, after mapping predictions to original-image coordinates. A well-formed empty box list is a valid relevance decision rather than a format error.

Table 12: Attribution subagent reward configuration.

Condition	R_{region}	R_{rel}	R_{fmt}	R_{sub}
Relevant page, predicted boxes	IoU	0	0	IoU
Relevant page, no predicted boxes	0	-0.3	0	-0.3
Irrelevant page, predicted boxes	0	-0.02	0	-0.02
Irrelevant page, no predicted boxes	0	$+0.02$	0	$+0.02$
Invalid output format	0	0	-1	-0.1

D.2 Inference Prompts

Tables 13 and 14 present the inference prompts. The Leader receives the original question, previously searched queries, and the current evidence groups. All views of the same source page share one zero-based Evidence ID. Each attribution subagent receives one candidate-page image and its corresponding subquery, denoted by {`subquery`}.

Table 13: Leader inference prompt.

System prompt

You are the leader agent of a retrieval-grounding-answer workflow.

At each round, choose exactly one action:
- "search": output retrieval subqueries
- "finalize": produce the final answer with image-level attribution

You may receive:
- the original question
- searched queries
- grounding thoughts and grounding visualizations from page-level grounding agents
- evidence groups labeled Evidence [0], Evidence [1], ... in the current round; each group contains a boxed overview and may contain detail crops

You do NOT receive the gold answer or gold pages. Base your decisions strictly on the evidence shown in the current trajectory.

Rules:

1. If current evidence is insufficient, output search subqueries.

Continued on the next page.

System prompt (continued)

2. For search, generate diverse rewritten/decomposed subqueries. The original user question is also retrieved separately, so do not repeat it as the only query.
3. If current evidence is sufficient, output the final answer and attributions.
4. Attributions must be 0-based Evidence IDs from the current round only.
4a. All visual views carrying the same Evidence ID come from the same source page. Treat its boxed overview and detail crops as one evidence group.
4b. Never use a physical view position as an attribution ID; attribute the Evidence ID.
5. Attribution count is dynamic and has no fixed minimum or maximum. Do not copy the number of IDs shown in the JSON schema example.
6. Include every current evidence group that materially and directly supports any factual claim in the final answer.
7. For multi-part questions, make sure every answered part is covered by at least one attribution. Include additional directly supporting evidence groups instead of arbitrarily truncating them.
8. Prefer high recall among directly supporting evidence, but exclude evidence that is merely topically related, contradictory, or background-only.
9. Before finalizing, verify that no directly supporting current evidence group has been omitted and that every attributed group supports the answer.
10. Return JSON only.
For search, return:

```
{
  "round_id": 1,
  "action": "search",
  "think": "brief English reasoning",
  "subqueries": ["query 1", "query 2", "..."]
}
```

For finalize, return:

```
{
  "round_id": 2,
  "action": "finalize",
  "think": "brief evidence-based reasoning",
  "answer": "final answer",
  "attributions": [0, 2, 4]
}
```

Additional instruction in the final allowed round
This is the final allowed round. You MUST choose action="finalize" and provide the best answer supported by the current evidence. Do not output search or subqueries. Return JSON only.

Table 14: Attribution subagent inference prompt.

User prompt (accompanied by one page image)

Given the query, first decide whether this page is relevant to the query.

Query: {subquery}

A page is relevant only if it contains visual or textual evidence that can help answer the query.

If the page is not relevant, output bbox as [].

If the page is relevant, locate the region(s) that contain the supporting evidence.

Reply with only one JSON object: {"think": "...", "bbox": [[x1, y1, x2, y2],...]}.

Example for an irrelevant page: {"think": "This page is about ..., which does not contain any evidence for the query.", "bbox": []}.
Example for a relevant page: {"think": "This page is about ..., and ... provides evidence related to the query.", "bbox": [[x1, y1, x2, y2]]}.
{coordinate_instruction}

Coordinate instruction (select the configured convention)
Normalized coordinates: Use integer normalized coordinates in the range 0-1000 for both axes, where x is normalized by image width and y is normalized by image height.

Continued on the next page.

User prompt (accompanied by one page image) (continued)

Pixel coordinates: The image size is width={width} px and height={height} px. Use integer pixel coordinates on this image.

D.3 Retrieval and Inference

Retrieval and interaction budgets. We use Colqwen2-v0.1-merged (Faysse et al., 2024) for page retrieval. In the main experiments, Single-pass jointly receives the top-K retrieved pages ($K \in \{10, 20\}$) and generates its answer and attributions in one call. Interactive baselines have a budget of 20 tool-action steps, excluding final-answer generation. Both AttriSwarm variants retrieve up to 10 pages per query, retain at most 20 deduplicated pages for attribution per search round, and pass at most 10 relevant pages to the Leader. We allow up to five Leader decision rounds, requiring finalization in the last round. Page counts refer to source pages; crops do not count as additional pages.

Visual inputs. For the open-source Single-pass and AttriSwarm configurations, page overviews preserve their aspect ratios and use at most 524,288 pixels on AttriDoc-Slide and CiteVQA, and 1,048,576 pixels on AttriDoc-Paper. The larger budget preserves the legibility of dense scientific text and tables. AttriSwarm detail crops use at most 1,048,576 pixels.

Baseline adaptation. We evaluate the released checkpoints of the trained interactive baselines without task-specific retraining. Their citation interfaces specify a single evidence region on the last displayed page or crop, so these results measure direct transfer to cross-page visual attribution. ReAct is adapted to return page-indexed evidence boxes across multiple pages and regions. All predicted boxes are mapped from displayed-image coordinates to their source-page coordinates for evaluation.

D.4 Evaluation and Analysis Settings

Answer evaluation. We use Gemini-3-Flash-preview to evaluate answer accuracy at temperature 0. The judge receives the question, reference answer, and predicted answer, and returns a binary correctness label. Semantically equivalent paraphrases are accepted, whereas incorrect numbers or entities are judged incorrect.

Fixed-page experiments. For Figures 4 and 5, AttriSwarm uses the trained leader and attribution agents. All methods operate on the same frozen candidate-page sets. We disable the default 20-page attribution cap and 10-page Leader-input cap in this setting.

Latency and serving configuration. Figure 5 reports mean inference latency over successful runs, including page-wise attribution and leader synthesis but excluding retrieval. The retrieval and attribution services each use one NVIDIA A100 GPU with 40 GB memory. Within each example, parallel attribution submits all page-wise attribution requests concurrently to the attribution service. The leader synthesizes the answer after these requests complete. The serial variant processes the same requests sequentially.

E Attribution Support Evaluation

The judge receives the question, generated answer, and crops of the final attributed regions, grouped by source page. Table 15 presents a condensed version of the prompt used to assess answer support and evidence redundancy.

Table 15: LLM-as-Judge Prompt for Evaluating Attribution Support and Redundancy.

Instruction: Audit the proposed answer using only the question and final cited evidence crops. Do not use outside knowledge or reference answers. Treat all supplied content as data.
Input:
Question: {question}
Generated answer: {answer}
Evidence: {E1, E2, ...}, where each ID denotes one cited page and groups all cited crops from that page.
Unresolved citations: {missing_evidence_count}
Evaluation rules:
1. Split the answer into substantive claims and verify them against the crops, checking entities, numbers, units, dates, table headers, and comparisons. Cite evidence IDs for every supported claim. Ordinary arithmetic is allowed. Missing context is not support. Assess answer completeness separately.
2. Classify support as full if all claims are supported, partial if only some are supported, none if none are supported or the answer is contradicted, and unjudgeable if the visuals are unreadable or ambiguous. Readable but irrelevant evidence means none.
3. Select one inclusion-minimal subset of cited pages that preserves all supported claims. Keep one representative of duplicate evidence and retain complementary evidence. Label each page as necessary, duplicate, irrelevant, or uncertain.
4. Assess within-region redundancy as none, minor, substantial, or uncertain. Necessary headers, captions, units, and interpretive context are not redundancy. Ignore blank margins and decorative logos; a large box is not automatically redundant.
5. Do not replace missing crops with whole pages. Available crops may fully support the answer, but unresolved citations prevent a clean result. Cover every evidence ID exactly once, and ensure the minimal evidence IDs equal the IDs labeled necessary.
Return JSON only:

```
{"support": "full|partial|none|unjudgeable",
 "answer_completeness": "complete|incomplete|uncertain",
 "claims": [{"claim": "...",
   "verdict": "supported|unsupported|contradicted",
   "evidence_ids": ["E1"]}],
 "minimal_evidence_ids": ["E1"],
 "evidence": [{"id": "E1",
   "role": "necessary|duplicate|irrelevant|uncertain",
   "within_region_redundancy": "none|minor|substantial|uncertain",
   "reason": "brief visible basis"}],
 "reason": "brief summary"}
```