ITJOINT: BENCHMARKING AND DIAGNOSING MLLMS ON QUERY-DRIVEN IMAGE-TEXT JOINT EXTRACTION FROM LONG DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In domain-specific multimodal long documents, images and text jointly convey complex knowledge that cannot be fully captured by plain text alone. Structuring such information into usable multimodal data is therefore a practical demand. Motivated by this demand, we systematically formulate a new research problem, query-driven image-text joint extraction from multimodal long documents, and develop a two-level taxonomy covering challenges arising from user intent and document content. Further, we construct ITJoint, a high-quality, manually annotated benchmark for systematically evaluating and diagnosing MLLM capabilities on this problem, comprising 2,455 pages of domain-specific documents with numerous non-decorative images and 910 answer instances. Our evaluation reveals three key limitations of current MLLMs: they tend to over-rely on visual similarity for image-text association, struggle to exhaustively recover all associated images in one-to-many correspondences, and exhibit a clear gap between page localization and fine-grained image localization. To examine whether these bottlenecks can be mitigated at the system level, we further introduce Q2IT as a strong multi-agent baseline. Our analyses show that multi-agent approaches can compensate for MLLM weaknesses and better leverage their strengths through specialized external tools and structured strategy design, while remaining vulnerable to error propagation across stages. The complete dataset, prompts, and code are publicly available.[1]

1 INTRODUCTION

As foundation models have learned from massive amounts of web-scale data, their development is increasingly turning toward specialized documents that preserve knowledge accumulated through long-term human practice (Muennighoff et al., 2023; Villalobos et al., 2024). Such domain-specific multimodal long documents contain rich textual descriptions and images whose semantics are often tightly coupled and cannot be fully captured by plain text alone. However, transforming this unstructured multimodal information into structured image-text pairs still largely relies on time-consuming manual annotation. Therefore, query-driven image-text joint extraction provides a practical way to satisfy researchers' targeted needs and to transform domain-specific long documents into high-quality multimodal datasets (Gadre et al., 2023).

Existing image-text extraction methods mostly rely on predefined rules for exhaustive extraction and lack query-driven, on-demand capabilities (Clark & Divvala, 2016; Li et al., 2019; Siegel et al., 2018; Song et al., 2025; Tang et al., 2024; Pang et al., 2024). DocVQA returns textual answers to user queries (Mathew et al., 2021; Tito et al., 2023; Van Landeghem et al., 2023; Ma et al., 2024; Deng et al., 2025), while BBox-DocVQA further provides bounding boxes that localize evidence regions supporting the textual answers (Yu et al., 2025), seemingly addressing the above need. However, it represents a different research problem from the above need in two key aspects: (1) the text and bounding box form a containment rather than a complementary relation, and (2) the textual answer remains complete without the bounding box, whereas the multimodal data instance is incomplete without the corresponding image. Moreover, existing DocVQA datasets mainly consist of academic

[1]The complete resources are available at `https://anonymous.4open.science/r/FullResources-72C7/`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

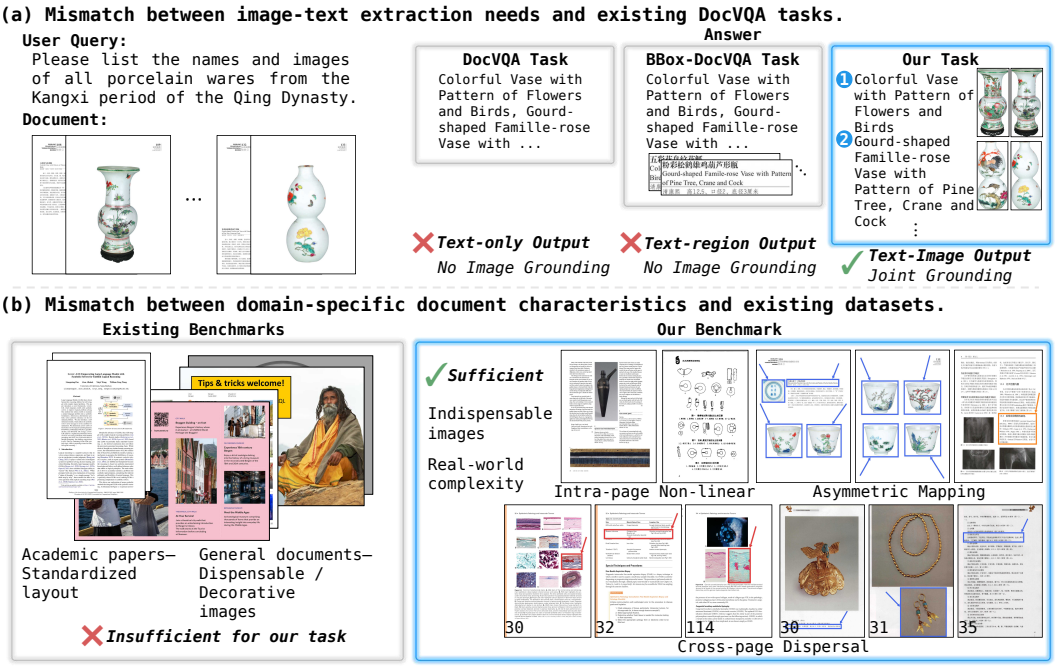


Figure 1: Mismatch between domain-specific image-text extraction needs and existing tasks and datasets. (a) DocVQA tasks mainly return textual answers or supporting evidence regions, whereas our task requires textual attributes paired with corresponding images. (b) Existing benchmarks mainly focus on standardized academic papers or general documents with dispensable or decorative images, which are insufficient for evaluating our task.

papers with standardized structures or general documents whose images are often decorative or auxiliary, and thus cannot adequately support the image-text annotations required by real-world domain-specific needs.

To this end, we systematically formulate a new research problem, query-driven image-text joint extraction from multimodal long documents. Given a user query and a long document, the task constructs answer instances by pairing query-relevant textual attributes with their corresponding images (Figure 1). However, this task poses challenges from both user intent and the document itself. From the perspective of user intent, real-world queries are not limited to clearly specified single-instance queries, but may also involve broadly referential set queries with an uncertain number of targets, requiring methods to balance precise extraction with comprehensive recall. From the document perspective, first, images and text within a page often go beyond linear arrangements and exhibit non-linear and multi-layer nested layouts, making rule-based association methods fragile in real-world domain-specific documents. Second, images and their descriptive text may be dispersed across pages rather than appearing adjacently on the same page, requiring cross-page tracking and aggregation with multi-level context. Third, image-text mappings often go beyond one-to-one correspondences, including one-to-many and text-without-image cases, requiring robust mapping and hallucination suppression. To enable systematic evaluation and diagnosis of MLLM capabilities on this problem, we design a two-level taxonomy spanning query- and instance-level challenges, and construct ITJoint, a high-quality manually annotated benchmark for query-driven image-text joint extraction from domain-specific multimodal long documents.

We systematically evaluate representative MLLMs on ITJoint, revealing three key limitations: they tend to over-rely on visual similarity for image-text association, struggle to exhaustively recover all associated images in one-to-many correspondences, and exhibit a clear gap between page localization and fine-grained image localization. To examine whether these bottlenecks can be mitigated at the system level, we further introduce Q2IT as a strong multi-agent baseline. Our analyses show that multi-agent approaches can compensate for MLLM weaknesses and better leverage their strengths through specialized external tools and structured strategy design, while remaining vulnerable to error

propagation across stages. *Moreover, these MLLM limitations differentiate ITJoint from existing DocVQA benchmarks, as such failure modes cannot be captured by existing DocVQA evaluations.*

In summary, our contributions are as follows:

- We systematically formulate a new research problem, query-driven image-text joint extraction from multimodal long documents. We further analyze its challenges from both user intent and the document itself, and design a two-level taxonomy for systematic evaluation.
- We construct ITJoint, a high-quality manually annotated benchmark designed for systematic evaluation and diagnosis of MLLM capabilities on this problem. ITJoint contains 2,455 document pages with an average of 1.77 non-decorative images per page and 910 answer instances.
- We systematically benchmark representative MLLMs on ITJoint and diagnose three key capability bottlenecks. To investigate system-level mitigation of these bottlenecks, we further introduce Q2IT as a strong multi-agent baseline, revealing both its benefits and limitations.

2 Related Work

Document Visual Question Answering and Benchmarks. DocVQA aims to answer user questions from document images (Mathew et al., 2021). Related benchmarks have evolved from single-page to multi-page documents (Tito et al., 2023), and further to multidomain, visually rich, and long-context multimodal document settings (Van Landeghem et al., 2023; Ma et al., 2024; Deng et al., 2025). Methods have also progressed from OCR- and layout-based document understanding models (Xu et al., 2021; Huang et al., 2022) to OCR-free models (Kim et al., 2021) and document-oriented multimodal large language models (Ye et al., 2023). BBox-DocVQA further introduces evidence bounding boxes for answer grounding (Yu et al., 2025). Nevertheless, existing DocVQA still centers on textual answers: even with bounding boxes, localization mainly serves as evidence grounding rather than extracting query-targeted textual attributes and corresponding image coordinates. Existing benchmarks are also dominated by structured academic papers or general documents with mostly auxiliary images. Therefore, existing DocVQA methods and benchmarks remain insufficient for query-driven image-text joint extraction.

Image-Text Pair Extraction from Documents. Another line of work extracts images and associated text from documents, including figure-caption extraction in scientific papers and multimodal data collection from domain-specific catalogs. Methods such as PDFFigures 2.0 (Clark & Divvala, 2016), PDFigCapX (Li et al., 2019), and DeepFigures (Siegel et al., 2018) extract figures, captions, and coordinates from academic papers, while later work studies subfigure-subcaption alignment in compound figures (Song et al., 2025). For domain-specific catalogs, PDFChatAnnotator (Tang et al., 2024) supports multimodal data collection from PDF-format catalogs through human-LLM collaboration, and VLM-based methods have been used to collect artifact images and attributes from archaeological catalogs (Pang et al., 2024). However, existing methods typically rely on predefined targets, fixed fields, or human-collaborative workflows, and cannot select target instances on demand from open-ended user queries.

3 Task Definition and Taxonomy

3.1 Task Formulation

Given a user query q and a long document $D = \{p_i\}_{i=1}^{N}$, the model is required to return a set of answer instances:

$$\mathcal{R} = \{z_j\}_{j=1}^{J}, \quad z_j = (e_j, t_j, v_j). \tag{1}$$

Here, i indexes the pages in D, and j indexes the answer instances in $\mathcal{R}$. The order of answer instances does not affect evaluation. Each answer instance z_j consists of an entity identifier e_j, a textual attribute value t_j satisfying the query condition, and an associated image set v_j:

$$v_j = \{(\pi_{j,m}, b_{j,m})\}_{m=1}^{M_j}, \tag{2}$$

where $\pi_{j,m}$ and $b_{j,m}$ denote the page index and bounding box of the m-th associated image, respectively. When textual evidence for an instance exists but no corresponding image is present, the system

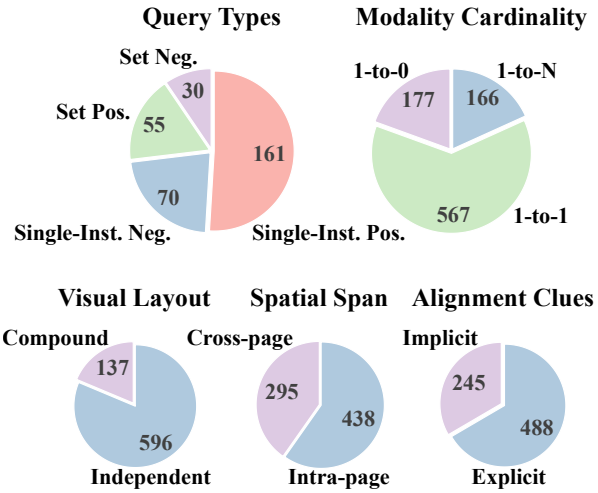


Figure 2: Distribution of ITJoint across taxonomy dimensions. Visual Layout, Spatial Span, and Alignment Clues apply only to non-1-to-0 instances.

Table 1: Summary statistics of ITJoint.

Statistic	Number
Documents	
Chinese / English	36.6 / 63.4%
Total pages	2,455
Avg. pages	59.88
Max / Min pages	178 / 22
Avg. images / page	1.77
Set Queries	
Total inst.	679
Avg. inst. / query	7.99
Max / Min inst. / query	62 / 1
Pos. / neg. inst. (neg. sets)	148 / 107
1-to-N Inst.	
Total images	450
Avg. images / inst.	2.71
Max / Min images / inst.	13 / 2

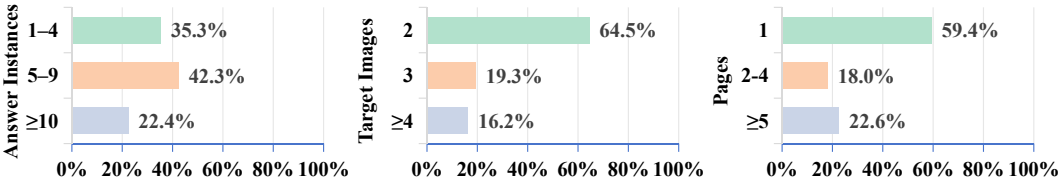


Figure 3: Detailed distributions of ITJoint. **Left:** Set-query size. **Middle:** Image cardinality of 1-to-N instances. **Right:** Image-text page distance in cross-page instances.

must still return t_j and explicitly output an empty image set, i.e., $v_j = \emptyset$, rather than omitting the instance or hallucinating an image match.

3.2 Challenge Taxonomy

To systematically characterize the challenges of producing $\mathcal{R}$ and support fine-grained evaluation, we define a two-level taxonomy over q and $z_j = (e_j, t_j, v_j)$.

Query-level taxonomy. To characterize challenges from user intent, we define two query-level categories. *Single-instance queries* require $J = 1$ and request image-text information for one specific answer instance. *Set queries* require all answer instances satisfying a given condition, where J is not known in advance. Based on this distinction, we further refine the conventional negative-query setting in existing benchmarks (Ma et al., 2024; Deng et al., 2025), where unanswerability is caused by the complete absence of textual evidence, into a task-specific negative setting where textual evidence exists but the corresponding image is missing.

Specifically, positive queries require answer instances to contain both the requested textual attributes and corresponding images. Negative queries differ across the two query types. For single-instance queries, the target instance has textual evidence but no corresponding image, resulting in a negative answer instance with $v_j = \emptyset$. For set queries, image-bearing and text-only instances coexist in $\mathcal{R}$, requiring the system to extract valid image-text pairs and explicitly identify instances with $v_j = \emptyset$.

Instance-level taxonomy. To characterize challenges from the document itself, we annotate each $z_j = (e_j, t_j, v_j)$ with fine-grained labels. For *modality cardinality*, instances are labeled according to M_j in Eq. 2: 1-to-0 if $M_j = 0$, 1-to-1 if $M_j = 1$, and 1-to-N if $M_j > 1$. Instances labeled as 1-to-0 contain textual evidence but no corresponding image and are therefore negative answer instances, while 1-to-1 and 1-to-N instances are positive answer instances.

For positive answer instances, we further annotate three dimensions that describe how t_j and v_j are associated in the document. For *visual layout*, instances are labeled as independent or compound, distinguishing whether the target image in v_j is a standalone image or is nested with other images in a compound image. For *spatial span*, instances are labeled as intra-page or cross-page according to whether the textual evidence and all image pages in v_j together involve a single page or multiple pages. For *alignment clues*, instances are labeled as explicit or implicit. Explicit alignment indicates that the document contains clear textual markers that directly establish the correspondence between t_j and v_j, whereas implicit alignment indicates that such markers are missing or insufficient, requiring the system to infer the correspondence using additional signals such as layout position, reading order, and visual content. Appendix E.2 provides visual examples of these instance-level labels.

4 ITJOINT BENCHMARK

4.1 DOCUMENT COLLECTION

Document collection follows a two-stage procedure. Candidate documents are obtained from public sources and recommendations by students or researchers with relevant domain backgrounds. They are first screened for domain relevance, diverse image-text relationships, and abundant non-decorative images, and then independently reviewed for their suitability for image-text annotation. Only documents passing both stages are included in ITJoint. In addition, the selected documents cover both English and Chinese, enabling evaluation of models' applicability across different language settings. In terms of domain distribution, ITJoint mainly covers medicine (39.0%) and archaeology (34.2%), together with biology (17.1%), art (7.3%), and agriculture (2.4%).

4.2 ANNOTATION WORKFLOW

We implement a three-stage annotation protocol, with one annotator conducting multi-round annotation and self-review, two independent checkers verifying the annotations, and one expert performing final adjudication. To support efficient and consistent annotation, we develop ITLabel. All three stages follow the same annotation guidelines. See Appendix A for ITLabel and annotation guidelines.

Notably, documents are inspected page by page throughout the workflow, especially for set queries. Each document is first assigned to one annotator, who formulates queries, identifies answer instances, annotates textual attributes, image regions, and taxonomy labels, and performs multiple rounds of self-review. The completed annotations are then independently checked by two checkers for validity, correctness, and completeness. Cases with unresolved disagreements are forwarded to a final reviewer for adjudication. Queries that remain ambiguous after final review are removed from ITJoint.

The initial single-annotator annotation and self-review stage lasted 3.5 months, resulting in 910 answer instances corresponding to 316 queries. During the subsequent checking and final review, three queries raised disagreements. One concerned a bounding-box annotation, which was retained after final review because the IoU threshold of 0.75 tolerates minor boundary deviations without affecting the evaluation result. A second case involved a set query for which checking identified an additional negative answer instance in another chapter. We therefore added a chapter constraint to the query, re-checked the answer set, and retained the query. The third case concerned a textual attribute, where the unit "cm" was added during final review for annotation rigor, after which the query was retained. Thus, all three disputed queries remained in ITJoint.

4.3 DATASET STATISTICS

Figures 2 and 3, together with Table 1, summarize the statistics of ITJoint. ITJoint contains 2,455 domain-specific document pages, with an average document length of 59.88 pages and 1.77 non-decorative images per page on average. It includes 910 answer instances across 316 queries. Among them, set queries correspond to 679 answer instances, with an average of 7.99 instances per query and a maximum of 62 instances for a single query. For 1-to-N cases, the 166 instances involve 450 images in total, with an average of 2.71 images per instance and a maximum of 13 images.

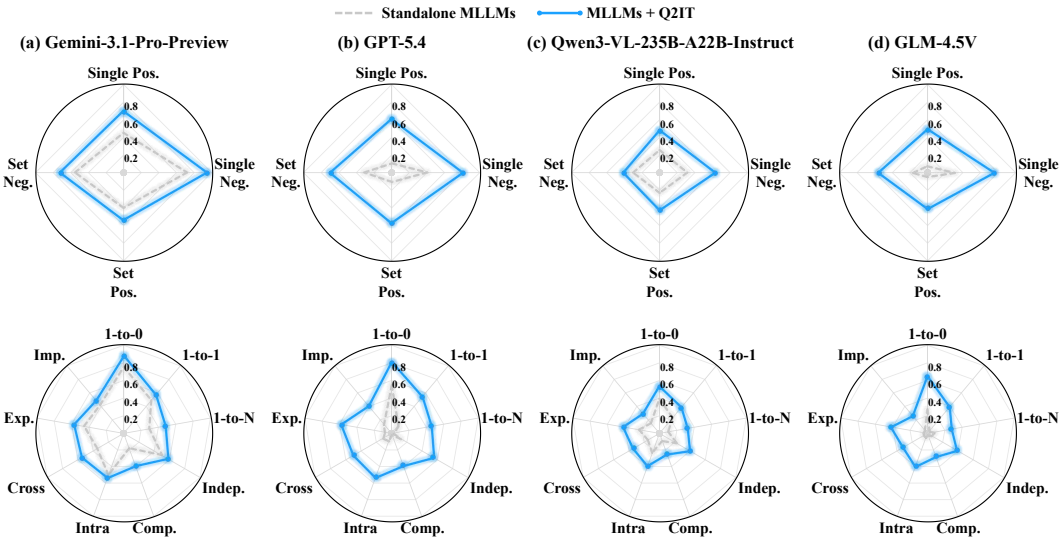


Figure 4: Main results on ITJoint for standalone MLLMs and Q2IT, a strong multi-agent baseline with the same MLLM backbones. See Appendix H.1 for the corresponding numerical results.

5 Experiments

In this section, we first evaluate standalone MLLMs to diagnose their limitations, and then introduce Q2IT as a multi-agent analysis vehicle to examine whether such system-level solutions can mitigate these limitations and what new challenges they introduce.

5.1 Experimental Settings

Metrics. We evaluate models in terms of image localization, text extraction, and image-text joint correctness. For image localization, we use an IoU threshold of $\tau = 0.75$ and define two settings: Strict and Soft. Strict requires the number of predicted image boxes to exactly match the number of gold boxes, with every matched box satisfying the IoU threshold. Soft relaxes the cardinality constraint, allowing a predicted instance to contain fewer boxes than the gold instance, but requiring every predicted box to match a gold box of the same instance with IoU $\geq \tau$. Text correctness is assessed using an LLM-as-a-Judge, and joint correctness requires both image localization and text extraction to be correct. We use Strict Joint as the main evaluation setting, reporting macro-F1 at the query level and accuracy at the instance level. The complete evaluation details are provided in Appendix G, with Soft results reported in Appendix D.

Baselines. We select four advanced multimodal large language models as baselines. For closed-source models, we use GPT-5.4 (OpenAI, 2026) and Gemini-3.1-Pro-Preview (Google, 2026). For open-source models, we use GLM-4.5V (Hong et al., 2025) and Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025). In the experiments, we refer to them as GPT, Gemini, GLM, and Qwen, respectively. We set the temperature of all models to 0.0 for reproducibility. Detailed input preprocessing and model-specific image limits are provided in Appendix B.

Strong Multi-Agent Baseline: Q2IT. Beyond standalone MLLMs, we introduce Q2IT as a strong multi-agent baseline to explore the potential of system-level solutions for this task, with implementation details provided in Appendix C. Q2IT consists of three sequential agents. The Evidence Agent designs three retrieval strategies, where an Evidence Planner selects the appropriate strategy to retrieve relevant document content and an Evidence Integrator extracts textual attributes and textual clues referring to target images. The Page Agent employs a Clue Analyzer to interpret these textual clues and select among multiple localization tools to identify candidate target pages. The Figure Agent follows a detect-then-select strategy, first generating candidate image regions and then selecting and binding the target regions to the corresponding textual attributes.

Table 2: Diagnostic results on standalone MLLMs and MLLMs within Q2IT. The 1-to-0 analysis reports Text and Joint accuracy and their gap, while the decomposed analysis reports Page, Image, and Joint performance.

Method	1-to-0 Analysis			Decomposed Analysis		
	Text Acc.↑	Joint Acc.↑	Gap↓	Page↑	Image↑	Joint↑
Standalone MLLMs						
Gemini-3.1-Pro-Preview	**0.849**	**0.744**	**0.105**	**0.881**	**0.518**	**0.507**
GPT-5.4	0.767	0.541	0.226	0.808	0.187	0.185
Qwen3-VL-235B-A22B-Instruct	0.714	0.387	0.327	0.776	0.268	0.263
GLM-4.5V	0.597	0.273	0.324	0.629	0.131	0.128
MLLMs within Q2IT						
Gemini-3.1-Pro-Preview	**0.904**	**0.876**	**0.028**	**0.908**	**0.793**	**0.722**
GPT-5.4	0.887	0.802	0.085	0.872	0.725	0.653
Qwen3-VL-235B-A22B-Instruct	0.778	0.528	0.250	0.786	0.553	0.492
GLM-4.5V	0.757	0.644	0.113	0.756	0.595	0.537

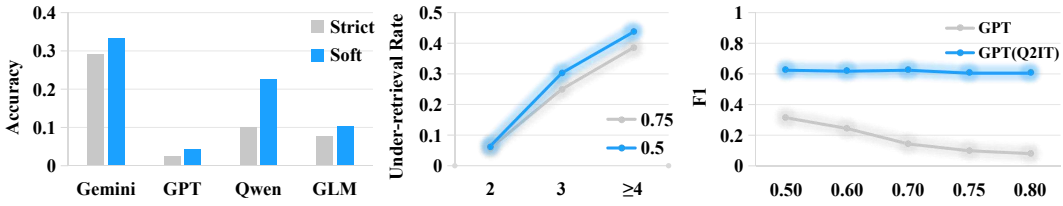


Figure 5: Diagnostic analyses on ITJoint. **Left:** Strict vs. Soft Joint on 1-to-N instances. **Middle:** Gemini under-retrieval rates by image cardinality at IoU 0.75 and 0.5. **Right:** GPT and GPT(Q2IT) F1 on single-positive queries across IoU thresholds.

5.2 Main Results

Figure 4 presents the main results under the Strict Joint setting. Overall, standalone MLLMs show limited performance on ITJoint, particularly on positive image-bearing and structurally challenging cases. The generally low scores indicate that query-driven image-text joint extraction from multimodal long documents remains challenging for current MLLMs and warrants further investigation. Beyond diagnosing standalone MLLMs, we further ask whether these limitations can be mitigated at the system level. To investigate this question, we introduce Q2IT as a strong multi-agent baseline that augments MLLMs with specialized external tools and structured workflows. As shown in Figure 4, Q2IT consistently achieves stronger performance across query- and instance-level categories, demonstrating the potential of multi-agent approaches for addressing this problem. Nevertheless, most category scores remain below 0.6, indicating that substantial challenges remain even under the multi-agent setting. These observations motivate two levels of further analysis: we first examine the capability limitations of standalone MLLMs revealed by ITJoint, and then investigate how multi-agent approaches mitigate these limitations, what challenges they still face, and what design and efficiency trade-offs they introduce.

5.3 What Limitations Does ITJoint Reveal in Standalone MLLMs?

Incomplete Recovery in One-to-Many Correspondence. Figure 5 (left) shows that Soft Joint consistently outperforms Strict Joint on 1-to-N instances. This raises a natural question: do MLLMs tend to retrieve only a subset of the target images in 1-to-N cases? We first examine the overall recovery performance across all 1-to-N instances, where Table 11 shows consistently low image-set recall, indicating that models recover only a limited portion of the required target images. To further investigate this issue, we group matched 1-to-N instances according to their ground-truth image cardinality and measure how often models predict fewer images than the ground truth. As shown in Figure 5 (middle), the under-retrieval rate increases substantially as the number of target images grows. Importantly, the same trend holds under IoU thresholds of both 0.75 and 0.5, suggesting that this pattern is not merely an artifact of strict bounding-box matching. These results indicate that MLLMs

Table 3: Q2IT vs. a simplified pipeline with the same tools.

Method	Cardinality			Layout		Span		Alignment	
	1-to-0	1-to-1	1-to-N	Indep.	Comp.	Intra	Cross	Exp.	Imp.
Simplified Pipeline	0.424	0.310	0.175	0.292	0.226	0.352	0.173	0.348	0.143
Q2IT	**0.802**	**0.534**	**0.452**	**0.545**	**0.387**	**0.530**	**0.495**	**0.574**	**0.400**

increasingly struggle to exhaustively recover all target images as one-to-many correspondence cardinality grows.

Insufficient Fine-Grained Localization within Pages. As shown in Table 2, standalone MLLMs perform substantially better in page localization than in image localization, suggesting that identifying the relevant page does not necessarily lead to accurate localization of the target image. However, this gap alone cannot determine whether the failure arises from selecting the wrong image within the page or from imprecise localization of the correct image. To further investigate this issue, we evaluate single-positive queries under different IoU thresholds. As shown in Figure 5 (right), the performance of standalone MLLMs consistently improves as the IoU threshold is relaxed, indicating that many predictions roughly identify the target image region but fail to localize it precisely under stricter criteria. These results confirm that fine-grained within-page localization remains an important bottleneck for current MLLMs.

Over-reliance on Visual Similarity. As shown in Table 2, on 1-to-0 instances, all standalone MLLMs show a clear gap of over 0.10 between Text and Joint accuracy, indicating that models often extract the correct textual information but still associate it with an image when no corresponding image exists. By further examining these failure cases, we observe that MLLMs tend to rely excessively on visual similarity for direct image-text association, while being less inclined to use evidence distributed across the document to verify or reject such associations (see Appendix E.3 for a visual example). While visual similarity is important for image-text association, reliable performance requires a better balance between visual matching and document-level verification reasoning. *One promising direction is to move beyond coarse answer-only supervision and introduce more fine-grained supervision over intermediate steps*, such as candidate identification, evidence gathering, and correspondence verification. Such structured supervision may encourage models to explicitly verify candidate associations before producing the final answer, rather than directly mapping a query to a seemingly plausible image-text pair.

5.4 How Do Multi-Agent Approaches Mitigate MLLM Limitations, and What Challenges Remain?

Compensating for Weaknesses while Leveraging Strengths. Table 2 shows that Q2IT brings much larger gains in image localization than in page localization across all backbones, suggesting that multi-agent workflows are particularly effective at compensating for capabilities where standalone MLLMs are weak. Specifically, Q2IT uses a detection tool to generate candidate regions and reformulates precise coordinate generation into candidate selection, allowing MLLMs to better leverage their general multimodal reasoning capabilities. This suggests that an important advantage of multi-agent approaches lies in using specialized external tools to compensate for model weaknesses while better leveraging their strengths.

Reducing Noise and Context Burden before Image Binding. Another source of improvement comes from progressively narrowing the document before image binding. The Evidence and Page Agents first reduce the long document to a small set of relevant pages, which not only filters out irrelevant or visually similar images but also shortens the multimodal context passed to the Figure Agent. This reduces both distraction from irrelevant content and the burden of processing long contexts, allowing more focused fine-grained image-text association. Consistently, Table 2 shows that Q2IT reduces the gap between Text and Joint accuracy on 1-to-0 instances across all backbones, indicating fewer incorrect image bindings.

Structured Strategy Design beyond Tool Aggregation. The above analyses show the benefits of specialized external tools, but an important question is whether these gains can be achieved by simply combining the same tools. To examine this, we construct a simplified pipeline that uses the same retrieval model, Grounding DINO, NMS, and MLLM-based box selection as Q2IT, but removes its structured multi-agent strategy. As shown in Table 3, Q2IT consistently outperforms the simplified pipeline across all categories, indicating that simply combining specialized tools is insufficient and that structured strategy design provides additional benefits. These results suggest that external tools and structured strategies play complementary roles in multi-agent systems: external tools compensate for specific weaknesses of MLLMs, while structured strategies organize these capabilities into a more effective solution for the target task. See Appendix I for component-level ablations.

Remaining Challenge: Error Propagation across Stages. While the staged multi-agent workflow enables flexible and task-specific designs, its non-end-to-end structure is inherently vulnerable to error propagation. Once errors or omissions occur at an early stage, downstream agents are inevitably affected and have limited opportunities to recover the missing information. To quantify this effect, we conduct Oracle Page and Oracle Evidence experiments with GPT-5.4 (Table 4). Replacing the predicted pages with ground-truth image pages improves Joint F1 on single-positive queries from 0.608 to 0.781, while further providing ground-truth evidence and target-image pages increases it to 0.882. On set-positive queries, Oracle Evidence similarly improves Joint F1 from 0.571 to 0.801. These results show that the effectiveness of staged multi-agent systems depends strongly on the quality of intermediate outputs, highlighting error propagation as an important challenge for reliable multi-agent solutions.

Table 4: Oracle analysis of error propagation using GPT-5.4 as the backbone.

Setting	Single Pos.	Set Pos.
Q2IT	0.608	0.571
Oracle Page	0.781	–
Oracle Evidence	0.882	0.801

5.5 Additional Experimental Analyses

Appendix F.1 presents our efficiency and cost analysis, showing that standalone MLLMs and multi-agent approaches exhibit different efficiency bottlenecks: direct inference requires less runtime and fewer model calls, whereas multi-agent approaches reduce token consumption at the cost of additional coordination overhead. Appendix F.2 examines the effect of document length across four models, showing a consistent decline in performance as document length increases. Appendix F.3 analyzes standalone Gemini across different image-text page distances, showing that Strict Joint Accuracy consistently decreases as the distance increases and quantitatively characterizing the increasing difficulty introduced by longer cross-page dependencies. Appendix F.4 presents a small cross-domain case study on a geology document, providing complementary evidence that similar image-text association and fine-grained localization difficulties can also arise outside the domains covered by ITJoint. Appendix F.5 evaluates the reliability of our LLM-based evaluator, showing 0.977 and 0.975 agreement with human judgments under the standalone and multi-agent settings, respectively.

6 Conclusion

We systematically define a new research problem, query-driven image-text joint extraction from multimodal long documents, and develop a two-level taxonomy to characterize its challenges from user intent and document content. We further construct ITJoint, a high-quality manually annotated benchmark for this task, and systematically evaluate representative MLLMs together with Q2IT as a strong multi-agent baseline. Our analyses reveal that current MLLMs tend to over-rely on visual similarity for image-text association, struggle to exhaustively recover all associated images in one-to-many correspondences, and exhibit clear limitations in fine-grained image localization. We further find that multi-agent approaches can compensate for MLLM weaknesses and better leverage their strengths through external tools and structured strategy design, while careful control of error propagation across stages remains important for reliable multi-agent systems.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing the manuscript. All AI-assisted revisions were reviewed by the authors, who take responsibility for the final content.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 143–152, 2016.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, 2025.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.

Google. Gemini 3.1 Pro Preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`, 2026. Accessed: 2026-05-21.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 4083–4091, 2022.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2, 2021.

Pengyuan Li, Xiangying Jiang, and Hagit Shatkay. Figure and caption extraction from biomedical documents. *Bioinformatics*, 35(21):4381–4388, 2019.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026. Accessed: 2026-05-21.

Honglin Pang, Yi Chang, Tianjing Duan, and Xi Yang. An archaeological catalog collection method based on large vision-language models. *arXiv preprint arXiv:2412.20088*, 2024.

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pp. 223–232, 2018.

Jifeng Song, Arun Das, Ge Cui, and Yufei Huang. Figex: Aligned extraction of scientific figures and captions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16558–16571, 2025.

Yi Tang, Chia-Ming Chang, and Xi Yang. Pdfchatannotator: A human-llm collaborative multi-modal data annotation tool for pdf-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 419–430, 2024.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2579–2591, 2021.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

Wenhan Yu, Wang Chen, Guanqiang Qi, Weikang Li, Yang Li, Lei Sha, Deguo Xia, and Jizhou Huang. Bbox docvqa: A large scale bounding box grounded dataset for enhancing reasoning in document visual question answer. *arXiv preprint arXiv:2511.15090*, 2025.

A Benchmark Details

A.1 Annotation Details

ITLabel Annotation Tool. To improve annotation efficiency and consistency, we develop **ITLabel**, a visual annotation tool specifically designed for query-driven image-text joint extraction from long documents. As shown in Figure 6, ITLabel provides a unified document-level annotation interface, rather than treating each page as an isolated image. It supports page browsing, OCR-based text extraction, keyword search, candidate-query revision, textual-attribute entry, image bounding-box annotation, and query-level and instance-level taxonomy labeling.

A key feature of ITLabel is its support for cross-page, multi-instance, and multi-image annotation. For a single query, annotators can create multiple answer instances; for a single answer instance, they can further link the textual attribute value to one or more image boxes across different pages. This design enables efficient annotation of set queries, cross-page cases, and 1-to-N image-text mappings. The OCR-based text extraction function also allows annotators to directly reuse document text when filling textual attributes, reducing manual transcription effort.

Document Selection Rules. We collect long documents from public sources and recommendations by students or researchers with relevant domain backgrounds to cover realistic document scenarios faced by researchers when acquiring image-text information. Document selection follows four principles: (1) **image richness**, where documents should contain abundant images that are indispensable for conveying core domain knowledge rather than merely decorative; (2) **domain diversity**, where documents should cover diverse domain-specific fields, such as medicine, agriculture, archaeology, biology, and art; (3) **language coverage**, where documents should cover different language settings, including English and Chinese; and (4) **document length**, where each document should contain a continuous page range from tens to over one hundred pages to form realistic long-range retrieval scenarios.

Query Generation Rules. Query generation follows seven principles: (1) **objectivity**, where the target attribute of each query should have a deterministic answer in the document; (2) **clarity**, where the query should be clearly formulated so that annotators can understand the requested information without ambiguity; (3) **diversity**, where queries should cover diverse language styles and attribute dimensions while avoiding repetitive templates; (4) **non-redundancy**, where similar queries about the same entity should be avoided; (5) **distribution balance**, where the number of involved instances and the distribution of categories should be kept as balanced as possible; (6) **self-containment**, where each query should be independently understandable without relying on additional context; and (7) **multimodal dependency**, where each query must require joint extraction of textual attributes and corresponding images rather than purely textual question answering.

Answer Annotation Rules. For each query, annotators locate all entities satisfying the query condition in the document, record their textual attribute values, annotate image bounding boxes on the corresponding pages, and assign query-level categories and instance-level attributes. Answer annotation follows four rules: (1) **entity specificity**, where each answer instance must refer to a specific individual entity rather than a generic concept; (2) **textual faithfulness**, where textual attribute values should be directly copied from the document and should not be manually summarized or inferred without evidence; (3) **explicit negative annotation**, where entities with textual evidence but no corresponding image must be explicitly recorded with an empty image set; and (4) **independent image boxes**, where each independent visual entity should be annotated with a separate bounding box, and multiple distinct images should not be merged into one box.

For set queries, annotators are required to exhaustively identify all instances satisfying the query condition in the document. Particular attention is paid to valid instances involving cross-page relations, implicit alignment, or weak visual clues, so as to avoid missing valid answer instances.

B Baseline Implementation Details

Each document page is rendered as a 144-DPI PNG image. The page images and the query are provided to each baseline in a single call whenever possible. When the number of pages exceeds the

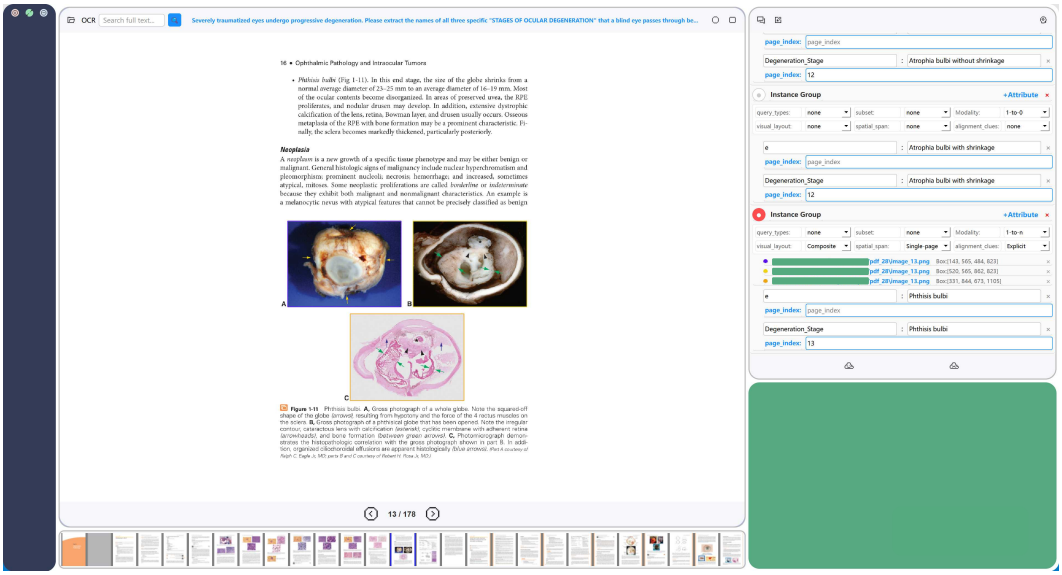


Figure 6: Interface of ITLabel.

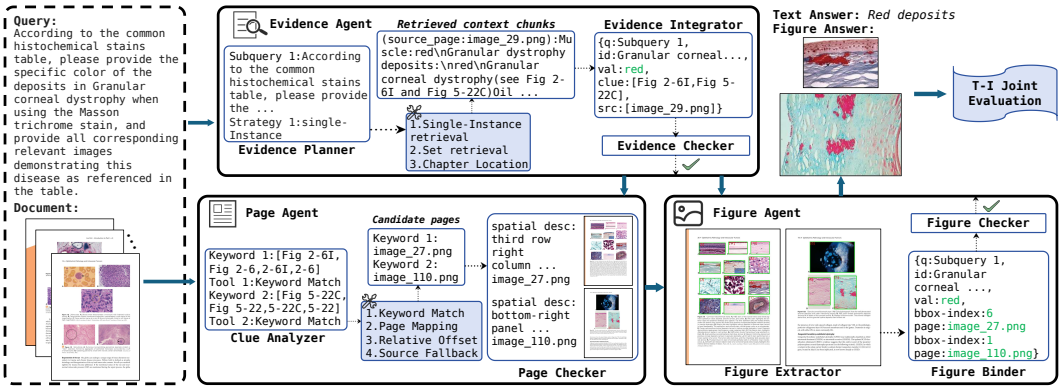


Figure 7: Overview of Q2IT, a strong multi-agent baseline designed to assess the potential of multi-agent systems for query-driven image-text joint extraction from multimodal long documents.

model's single-call image limit, we follow the stitching strategy of Ma et al. (2024), arranging pages into composite images using a three-column grid and mapping the predicted coordinates back to the original page coordinates. Due to API limitations, GPT, GLM, and Qwen support up to 50 images per call, whereas Gemini supports up to 16.

C Implementation Details of Q2IT

Q2IT is introduced as a strong multi-agent baseline, with the specific purpose of assessing the potential of multi-agent systems for query-driven image-text joint extraction from multimodal long documents. As shown in Figure 7, given a query q and a document D, Q2IT progressively constructs the answer set R through three stages: evidence collection, page localization, and image-text binding, handled by the Evidence Agent, Page Agent, and Figure Agent, respectively.

Before the agent workflow, each page of D is rendered as a 144-DPI image. Page-level text is extracted and segmented into page-indexed chunks, which are indexed using sparse and dense retrieval. The chunks are additionally clustered with a Gaussian Mixture Model (GMM) to improve recall for set queries. We use Qwen3-Embedding-8B to encode the text chunks. For hybrid retrieval, FAISS and BM25 each retrieve the top 30 candidates, which are fused and reranked by BAAI/bge-

reranker-v2-m3, with the top 10 chunks used for evidence extraction. For section-restricted queries, Q2IT expands a 30-page window after locating the section title and falls back to the next 50 chunks when page indices cannot be reliably parsed.

C.1 Evidence Agent

The Evidence Agent extracts candidate textual values and image-reference clues from the document. The Evidence Planner first decomposes the query q into subqueries $\{\tilde{q}_a\}_{a=1}^{A}$ and selects an appropriate retrieval strategy according to the query intent.

For single-instance queries, hybrid retrieval is used to obtain the top-k relevant chunks. For set queries, the top-k chunks are first retrieved, after which seed chunks are selected and expanded to their corresponding GMM clusters to improve coverage. When a subquery specifies a section range, the system first attempts to locate the corresponding section boundaries and falls back to retrieval if localization fails.

The Evidence Integrator then converts the retrieved contexts into a subquery-conditioned evidence collection:

$$E = \{(\tilde{q}_a, E_a)\}_{a=1}^{A}, \qquad E_a = \{e_{a,r}\}_{r=1}^{n_a}, \tag{3}$$

where

$$e_{a,r} = (id_{a,r}, val_{a,r}, clue_{a,r}, src_{a,r}). \tag{4}$$

Here, $id_{a,r}$ denotes the candidate instance identifier, $val_{a,r}$ is the candidate textual value, $clue_{a,r}$ contains image-reference clues, and $src_{a,r}$ records the source pages. The Evidence Checker removes irrelevant evidence and triggers supplementary retrieval when the textual information is missing or incomplete.

C.2 Page Agent

The Page Agent locates candidate target pages for each evidence item through a clue-driven localization strategy. Different types of clues are handled using different localization tools: explicit figure labels are localized through keyword matching, printed page numbers through page-index mapping, relative page references through offset computation, and intra-page spatial references by directly using the evidence page. If clue-based localization fails, the system falls back to entity-based full-document matching and subsequently to the source pages recorded by the Evidence Agent.

The Page Checker further verifies the candidate pages and their neighboring pages, producing verified target pages together with coarse spatial descriptions:

$$L = \{L_a\}_{a=1}^{A}, \qquad L_a = \{\ell_{a,r}\}_{r=1}^{n_a}, \tag{5}$$

where

$$\ell_{a,r} = \{(\rho_{a,r,u}, loc_{a,r,u})\}_{u=1}^{U_{a,r}}. \tag{6}$$

Here, $\rho_{a,r,u}$ denotes a verified target page and $loc_{a,r,u}$ denotes its coarse spatial description for subsequent image selection.

C.3 Figure Agent

The Figure Agent adopts a detect-then-select strategy to perform fine-grained image localization and image-text binding. For each verified page $\rho_{a,r,u}$, the Figure Extractor first applies Grounding DINO (Liu et al., 2024) to generate candidate image regions, followed by non-maximum suppression, area filtering, and validity filtering.

The Figure Binder then reformulates precise bounding-box generation as in-page candidate selection. Candidate regions are numbered on the original page, and the MLLM selects the target region based on the corresponding evidence item $e_{a,r}$, subquery $\tilde{q}_a$, and coarse spatial description $loc_{a,r,u}$. This formulation preserves page-level contextual information, including captions, labels, and layout relations.

Finally, the Figure Checker removes duplicate instances and merges complementary results obtained from different subqueries. The verified textual values and selected image regions are combined into the final answer set:

$$R = \{z_j\}_{j=1}^{J}, \qquad z_j = (e_j, t_j, v_j). \tag{7}$$

Table 5: Additional results of standalone MLLMs and Q2IT as a strong multi-agent baseline on ITJoint under the Soft Joint setting.

Model	Single Pos.	Single Neg.	Set Pos.	Set Neg.	1-to-0	1-to-1	1-to-N	Indep.	Comp.	Intra	Cross	Exp.	Imp.
Standalone MLLMs													
Gemini-3.1-Pro-Preview	**0.466**	**0.714**	**0.410**	**0.545**	**0.744**	**0.482**	**0.333**	**0.509**	**0.183**	**0.505**	**0.364**	**0.458**	**0.429**
GPT-5.4	0.099	0.393	0.107	0.310	0.541	0.113	0.042	0.118	0.007	0.091	0.105	0.080	0.131
Qwen3-VL-235B-A22B-Instruct	0.269	0.309	0.256	0.305	0.387	0.230	0.227	0.258	0.101	0.265	0.171	0.258	0.176
GLM-4.5V	0.074	0.304	0.056	0.158	0.273	0.034	0.102	0.057	0.015	0.048	0.051	0.045	0.057
MLLMs within Q2IT													
Gemini-3.1-Pro-Preview	**0.727**	**0.943**	0.583	**0.722**	**0.876**	**0.563**	0.596	**0.592**	**0.475**	0.550	**0.600**	0.607	**0.498**
GPT-5.4	0.649	0.805	**0.624**	0.701	0.802	0.534	**0.621**	0.579	0.453	**0.555**	0.556	**0.623**	0.420
Qwen3-VL-235B-A22B-Instruct	0.509	0.623	0.476	0.425	0.528	0.382	0.473	0.424	0.309	0.406	0.396	0.444	0.318
GLM-4.5V	0.535	0.752	0.490	0.580	0.644	0.388	0.452	0.416	0.343	0.418	0.380	0.459	0.290

D Additional Results under the Soft Joint Setting

In the main experiments, we report results under the Strict Joint setting, which requires the predicted image set to exactly match the gold image set. While this setting evaluates complete image-text joint extraction, it may be overly conservative in scenarios where a subset of correct images is sufficient for downstream use.

Therefore, we additionally report results under the Soft Joint setting in Table 5. Soft Joint relaxes the image-box cardinality constraint by allowing the prediction to contain fewer boxes than the gold instance. For image-bearing instances, the prediction must contain at least one correctly matched box, and every predicted box must match a gold box of the same instance. For 1-to-0 instances, the prediction is still required to return an empty image set. Thus, Soft Joint rewards partially complete but correct image retrieval while continuing to penalize irrelevant or hallucinated images. Compared with Strict Joint, Soft Joint generally yields higher scores, particularly for 1-to-N instances. Importantly, the overall performance patterns across models and benchmark categories remain broadly consistent, indicating that the main observations drawn from ITJoint are not sensitive to the choice between the two evaluation settings.

E Illustrative Examples

E.1 Ground-Truth Annotation Example

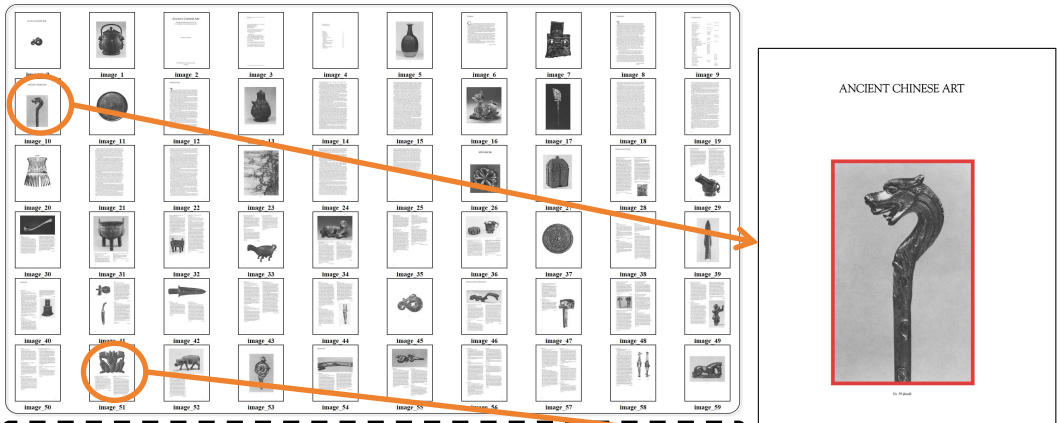
Figure 8 shows the ground-truth annotation format for a single-instance query in ITJoint. The query requires a textual answer and its corresponding target image. The annotation records the `text_answer`, the `image_answer` with the target bounding box, and the query-level and instance-level labels. This example further illustrates a cross-page case, where completing the answer instance requires information from multiple pages. For readability, we present a compact version of the annotation and omit implementation-specific fields such as local file paths.

E.2 Instance-Level Taxonomy Examples

Figure 9 provides examples of the instance-level taxonomy in ITJoint. Each panel corresponds to one answer instance and shows its labels along modality cardinality, visual layout, spatial span, and alignment clues. Red boxes indicate relevant textual evidence or target image regions.

Panel (a) shows an answer instance associated with multiple target images, and is therefore labeled as 1-to-N. The textual evidence and all target images together involve multiple pages, so the instance is also labeled as cross-page. Since the target images are nested in compound visual layouts and the document provides explicit label references linking the text to the images, it is further labeled as compound and explicit.

Panel (b) shows a 1-to-1 instance, where one textual attribute value corresponds to a single target image. Although the target image itself is independent, the textual evidence and the image together



Query:
The text describes a silver hairpin modeled to resemble a "wooden staff". According to the text, this dragon-head staff form is an attribute of which specific figures? Also, state the length of this hairpin (in cm), and provide its image.

Ground Truth:
text_answer:
Attribute_Of = Taoist sages and immortals;
Length = 25.1 cm
image_answer:
image_10.png, bounding box shown in red
labels:
single-instance / positive;
1-to-1 / independent / cross-page / explicit

Figure 8: Ground-truth annotation example for a single-instance query in ITJoint.

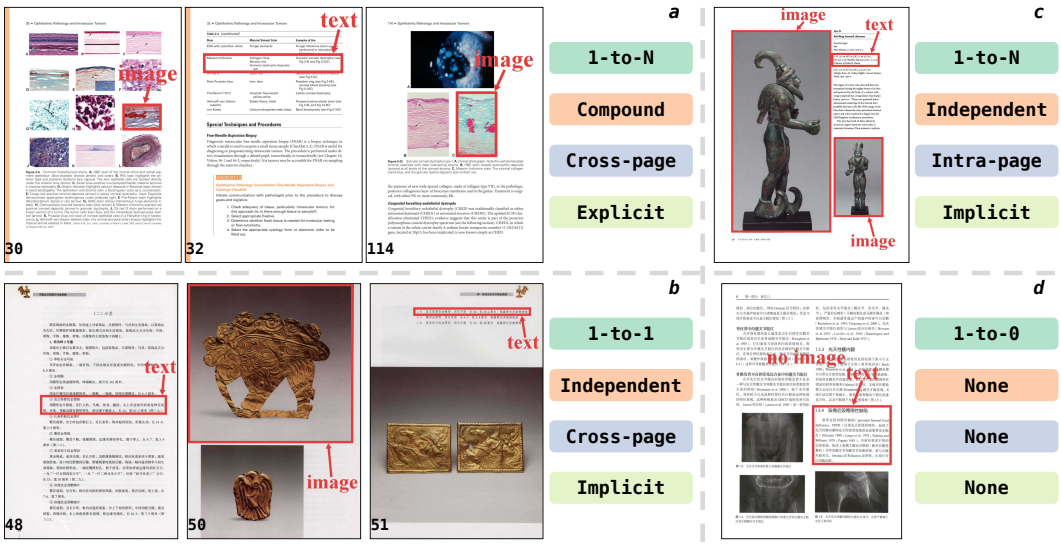


Figure 9: Examples of the instance-level taxonomy in ITJoint.

involve multiple pages, requiring cross-page association. The correspondence is implicit because the image-text relation cannot be established solely through an explicit marker and must be inferred from the surrounding context.



Source Page

Query:
The text details the principle of Self-glazing clay bodies. Please explain the specific role of fluxes during the firing of the body, and extract the image showing the surface effect after firing.

Ground truth:
"text_answer":"**the fluxes combine with the silica and alumina in the body to create a thin layer of glaze on the surface of the ware.**",
"image_answer": []

Q2IT output:
"text_answer":"When the piece is fired, **the fluxes combine with the silica and alumina in the body to create a thin layer of glaze on the surface.**",
"image_answer": []

Baseline output:
"text_answer":"In self-glazing clay bodies, soluble fluxes migrate during drying from the interior of the clay body to the surface. After the piece is fired, **these fluxes combine with the silica and alumina in the body to form a thin surface layer of glaze.** This creates a glazed or semi-glazed fired surface as the characteristic surface effect.",
"image_answer": [{"page": "image_10.png","bbox": [0.09, 0.07, 1.0, 0.55]}]

Figure 10: Example of image-text alignment hallucination on a 1-to-0 instance.

Table 6: Efficiency comparison using GPT-5.4.

Metric	Standalone MLLM	Q2IT
Tokens / Query	81,770.0	66,452.7
Time / Query (s)	133.6	191.5
API Calls / Query	1.03	24.54
Failure Rate	0.63%	0.32%

Panel (c) illustrates an intra-page instance, where the textual evidence and all target images together involve only a single page. The target images are standalone visual regions rather than parts of a compound figure, so the instance is labeled as independent. Since multiple target image regions are associated with the same textual attribute value, it is labeled as 1-to-N. Its alignment is implicit, as the association relies on page layout and local context rather than a direct textual reference.

Panel (d) shows a 1-to-0 instance. The document contains textual evidence for the target instance, but no corresponding image is present. Therefore, the system is expected to return the textual attribute value with an empty image set, i.e., $v_j = \emptyset$, rather than hallucinating an image region. Since no image is associated with the instance, the visual layout, spatial span, and alignment-clue labels are not applicable.

E.3 Image-Text Alignment Hallucination Example

Figure 10 shows a representative image-text alignment hallucination on a 1-to-0 instance. The document contains textual evidence for the queried entity, but no corresponding target image exists. Under direct inference, GPT correctly identifies the textual answer but still assigns an image region, resulting in correct text extraction but incorrect joint prediction. Notably, clear additional evidence on the same page indicates that the selected image refers to a different concept, yet the model still establishes an incorrect image-text association.

F Additional Experimental Analyses

F.1 Efficiency and Cost Analysis

Table 6 shows that direct MLLM inference requires substantially less runtime and fewer model calls, but consumes more tokens because the entire long document is processed directly. This suggests that standalone MLLMs still need more efficient long-context processing and multimodal token utilization to reduce the computational cost of handling long documents. In contrast, Q2IT reduces token consumption through staged processing and tool use, but requires longer runtime and substantially

Table 7: Performance of different models across different document lengths.

Document Length	#Docs	#Queries	Avg. Pages	Gemini	GPT	GLM	Qwen
22–42	15	110	35.07	0.577	0.244	0.190	0.372
46–68	14	108	58.64	0.522	0.168	0.150	0.243
72–178	12	98	92.33	0.402	0.134	0.033	0.123

Table 8: Strict Joint Accuracy of standalone Gemini under different image-text page distances.

Distance	**Strict Joint Accuracy**
Same page	0.507
1 page	0.338
2–4 pages	0.277
$\geq$5 pages	0.254

more model calls due to multi-agent coordination. This indicates that multi-agent approaches still need more efficient agent interaction and tool execution. Overall, the two paradigms exhibit different efficiency bottlenecks, and future systems should seek a better balance among task performance, token consumption, and inference overhead.

F.2 Effect of Document Length

We further analyze model performance across documents of different lengths by dividing the documents into three length groups and evaluating four models on each group. As shown in Table 7, all models exhibit a consistent decline in Strict Joint F1 as document length increases. This trend further highlights the growing challenge posed by longer document contexts.

F.3 Effect of Cross-Page Distance

We further analyze standalone Gemini with respect to image-text page distance. As shown in Table 8, Strict Joint Accuracy consistently decreases as the distance between textual evidence and the corresponding images increases. While longer cross-page dependencies are intuitively more challenging, ITJoint enables this effect to be quantitatively characterized, revealing a graded increase in difficulty as relevant multimodal information becomes more widely separated across the document.

F.4 Cross-Domain Case Study

To examine whether the diagnostic patterns observed on ITJoint also appear in other domains, we conduct a small case study using standalone GPT-5.4 on a long geology document, a domain not covered by ITJoint, with 13 manually annotated queries.

As shown in Table 9, several similar patterns can be observed. For the seven 1-to-0 instances, GPT correctly extracts the textual answer in all cases, while only four achieve correct joint predictions, showing that incorrect image associations can still occur despite correct textual extraction. Among the six image-bearing instances, GPT identifies at least one correct target page in five cases and the complete target-page set in three cases, but none achieves a correct image set, again showing a clear gap between page localization and fine-grained image localization. Overall, this small-scale study provides complementary evidence that similar difficulties can also arise outside the domains covered by ITJoint.

F.5 Reliability of LLM-as-a-Judge

To assess the reliability of the LLM-based text evaluation, we manually verify the GPT-4o Judge decisions on the Gemini-3.1-Pro-Preview results using two independent human reviewers (Table 10). The evaluation covers 728 direct-inference results and 825 Q2IT results, achieving human agreement rates of 0.977 and 0.975, respectively. Further examination of the disagreements shows that the Judge

Table 9: Cross-domain case study with standalone GPT-5.4 on 13 manually annotated queries.

Diagnostic Aspect	**Metric**	**Result**
Image–text alignment	Text correct / Strict Joint correct ($n = 7$)	7/7 / 4/7
One-to-many cardinality	Predicted images $<$ gold images ($n = 4$)	1/4
Fine-grained localization	$\geq$1 target page / Exact page set / Strict image set ($n = 6$)	5/6 / 3/6 / 0/6

Table 10: Human validation of the GPT-4o Judge.

Setting	**Human Agreement**
Standalone MLLM	97.7%
Q2IT	97.5%

tends to be conservative, with most errors arising from correct answers being judged as incorrect. These results indicate that the LLM-based evaluation is highly consistent with human assessment.

G Evaluation Protocol

G.1 Metric Definitions

Since this task requires both image-region localization and textual attribute extraction, we evaluate models from three perspectives: image localization, text extraction, and image-text joint correctness. For each query, we first greedily match predicted instances with gold instances that contain images based on instance-level bounding-box overlap. After all image-bearing instances are matched, the remaining 1-to-0 gold instances are matched with the remaining predictions based on textual equivalence. For a matched 1-to-0 instance, image localization is considered correct only if the prediction also outputs an empty image set; otherwise, it is counted as an image-text alignment hallucination.

For image localization, we use an IoU threshold of $\tau = 0.75$ and define two settings: Strict and Soft. Strict requires the number of predicted image boxes to exactly match the number of gold boxes, with every matched box satisfying the IoU threshold. Soft relaxes the cardinality constraint, allowing a predicted instance to contain fewer boxes than the gold instance, but requiring every predicted box to match a gold box of the same instance with IoU $\geq \tau$. For text extraction, we adopt the LLM-as-a-Judge paradigm and assign each instance a binary correctness label. The joint metric requires both image localization and text extraction to be correct. We use Strict Joint as the main evaluation setting, reporting macro-F1 at the query level and accuracy at the instance level.

G.2 Why Greedy Matching Is Reliable

Different answer instances correspond to different target regions, and under the IoU threshold used in our evaluation, no predicted box can form a valid match with multiple gold instances simultaneously. Therefore, the matching of one answer instance does not affect the matching of another, and greedy matching does not affect the evaluation results.

G.3 Why Image-Bearing Instances Are Matched First

A correct prediction for a 1-to-0 instance must output an empty image set and therefore does not participate in the first-stage image-based matching. Conversely, a prediction with a non-empty image set cannot be correct for a 1-to-0 instance by definition. Therefore, matching image-bearing instances first cannot consume a correct 1-to-0 prediction or change its correctness.

Algorithm 1 Evaluation Procedure of ITJoint

Require: Gold instances $\mathcal{G}$, predicted instances $\hat{\mathcal{G}}$, IoU threshold τ
Ensure: Strict/Soft Image and Joint correctness, and query-level evaluation
Split $\mathcal{G}$ into image-bearing instances $\mathcal{G}^+$ and 1-to-0 instances $\mathcal{G}^0$
Mark all predicted instances as unmatched
for each $g \in \mathcal{G}^+$ **do**
    Select the unmatched prediction $\hat{g}$ with the largest number of valid box matches to g
    Match predicted and gold boxes one-to-one on the same page with IoU $\geq \tau$
    if at least one valid box match exists **then**
        Mark $\hat{g}$ as matched
        Let m, n_g, and n_p denote the numbers of matched, gold, and predicted boxes
        $\mathrm{StrictImg}(g) \leftarrow [m = n_g = n_p]$
        $\mathrm{SoftImg}(g) \leftarrow [m = n_p > 0]$
        $\mathrm{Text}(g) \leftarrow \mathrm{TextJudge}(g, \hat{g})$
    else
        $\mathrm{StrictImg}(g), \mathrm{SoftImg}(g), \mathrm{Text}(g) \leftarrow 0$
    end if
    $\mathrm{StrictJoint}(g) \leftarrow \mathrm{StrictImg}(g) \wedge \mathrm{Text}(g)$
    $\mathrm{SoftJoint}(g) \leftarrow \mathrm{SoftImg}(g) \wedge \mathrm{Text}(g)$
end for
for each $g \in \mathcal{G}^0$ **do**
    Match g with a remaining prediction $\hat{g}$ based on textual equivalence
    if a textual match exists **then**
        Mark $\hat{g}$ as matched
        $\mathrm{Text}(g) \leftarrow 1$
        $\mathrm{StrictImg}(g), \mathrm{SoftImg}(g) \leftarrow [\hat{v} = \emptyset]$
    else
        $\mathrm{Text}(g), \mathrm{StrictImg}(g), \mathrm{SoftImg}(g) \leftarrow 0$
    end if
    $\mathrm{StrictJoint}(g) \leftarrow \mathrm{StrictImg}(g) \wedge \mathrm{Text}(g)$
    $\mathrm{SoftJoint}(g) \leftarrow \mathrm{SoftImg}(g) \wedge \mathrm{Text}(g)$
end for
Compute query-level Precision, Recall, and F1 from the numbers of correct, gold, and predicted instances
Compute instance-level accuracy over gold instances
Compute Page Precision, Recall, and F1 from the gold and predicted page sets

G.4 Detailed Evaluation Procedure

For completeness, we provide the detailed evaluation procedure used in ITJoint. Given the gold answer set $\mathcal{G}$ and the predicted answer set $\hat{\mathcal{G}}$ for a query, evaluation first establishes one-to-one correspondences between predicted and gold answer instances, and then evaluates image, text, and joint correctness. The complete procedure is summarized in Algorithm 1.

For each query, let C_q, N_q^{gold}, and N_q^{pred} denote the numbers of correct, gold, and predicted answer instances, respectively. Query-level Precision, Recall, and F1 are computed as

$$\mathrm{Prec}_q = \frac{C_q}{N_q^{\mathrm{pred}}}, \qquad \mathrm{Rec}_q = \frac{C_q}{N_q^{\mathrm{gold}}}, \qquad F1_q = \frac{2\,\mathrm{Prec}_q\,\mathrm{Rec}_q}{\mathrm{Prec}_q + \mathrm{Rec}_q}. \tag{8}$$

The reported query-level score is the macro-average of $F1_q$ across queries. Instance-level accuracy is the proportion of gold instances that satisfy the corresponding correctness criterion.

For page-level evaluation, we collect all pages containing gold image regions into a gold page set $\mathcal{P}_q^{\mathrm{gold}}$ and all pages containing predicted image regions into a predicted page set $\mathcal{P}_q^{\mathrm{pred}}$. Page Precision and Recall are computed by set overlap:

$$\mathrm{Prec}_q^{\mathrm{page}} = \frac{|\mathcal{P}_q^{\mathrm{gold}} \cap \mathcal{P}_q^{\mathrm{pred}}|}{|\mathcal{P}_q^{\mathrm{pred}}|}, \qquad \mathrm{Rec}_q^{\mathrm{page}} = \frac{|\mathcal{P}_q^{\mathrm{gold}} \cap \mathcal{P}_q^{\mathrm{pred}}|}{|\mathcal{P}_q^{\mathrm{gold}}|}, \tag{9}$$

Table 11: Image-set recovery on all 1-to-N instances.

Model	Precision	Recall	F1
Gemini-3.1-Pro-Preview	0.434	0.420	0.421
GPT-5.4	0.105	0.097	0.099
Qwen3-VL-235B-A22B-Instruct	0.252	0.194	0.212
GLM-4.5V	0.144	0.123	0.129

Table 12: Additional results of standalone MLLMs and Q2IT as a strong multi-agent baseline on ITJoint under the Strict Joint setting.

Model	**Single Pos.**	**Single Neg.**	**Set Pos.**	**Set Neg.**	**1-to-0**	**1-to-1**	**1-to-N**	**Indep.**	**Comp.**	**Intra**	**Cross**	**Exp.**	**Imp.**
Standalone MLLMs													
Gemini-3.1-Pro-Preview	**0.447**	**0.714**	**0.400**	**0.545**	**0.744**	**0.482**	**0.291**	**0.498**	**0.182**	**0.502**	**0.344**	**0.452**	**0.412**
GPT-5.4	0.099	0.393	0.104	0.310	0.541	0.113	0.024	0.113	0.007	0.084	0.105	0.074	0.131
Qwen3-VL-235B-A22B-Instruct	0.249	0.309	0.225	0.298	0.387	0.230	0.100	0.226	0.093	0.235	0.149	0.231	0.147
GLM-4.5V	0.074	0.304	0.046	0.158	0.273	0.034	0.078	0.050	0.015	0.046	0.041	0.043	0.045
MLLMs within Q2IT													
Gemini-3.1-Pro-Preview	**0.692**	**0.943**	0.535	**0.714**	**0.876**	**0.563**	**0.476**	**0.577**	**0.394**	**0.539**	**0.549**	0.572	**0.486**
GPT-5.4	0.608	0.805	**0.571**	0.691	0.802	0.534	0.452	0.545	0.387	0.530	0.495	**0.574**	0.400
Qwen3-VL-235B-A22B-Instruct	0.476	0.623	0.422	0.407	0.528	0.382	0.321	0.395	0.250	0.388	0.338	0.410	0.286
GLM-4.5V	0.485	0.752	0.403	0.557	0.644	0.388	0.277	0.383	0.277	0.395	0.315	0.416	0.257

with Page F1 computed accordingly. This metric evaluates whether the model identifies the correct pages independently of fine-grained bounding-box localization.

H Full Numerical Results

H.1 Results under the Strict Joint Setting

Table 12 provides the complete numerical results corresponding to the main results in Figure 4.

I Additional Ablation Analysis of Q2IT

The comparison with the simplified pipeline in the main paper shows that the gains of Q2IT cannot be attributed solely to the use of the same external tools. We further analyze how different stages of the multi-agent workflow contribute to its behavior. Unless otherwise specified, all analyses use GPT-5.4 as the backbone.

I.1 Evidence Retrieval

We first examine the set-retrieval strategy used by the Evidence Agent. Removing this strategy reduces performance on set-positive queries from 0.571 to 0.464 and on set-negative queries from 0.691 to 0.503. This suggests that expanding evidence beyond individually retrieved chunks is particularly important for queries that require identifying multiple answer instances.

I.2 Page Localization

We next analyze the Page Agent by removing the Page Checker and the multi-strategy page-search procedure. As shown in Table 14, removing the Page Checker reduces Page F1 from 0.872 to 0.799, with particularly clear degradation on cross-page and implicit-alignment instances. When the dedicated page-search procedure is further removed and only source pages from the Evidence Agent are used as candidate pages, Page F1 decreases to 0.740 and cross-page accuracy drops from 0.495 to 0.234. These results indicate that explicit page localization and verification are important when relevant text and images are distributed across different pages.

Table 13: Ablation of the set-retrieval strategy in the Evidence Agent using GPT-5.4 as the backbone.

Setting	Set Pos.	Set Neg.
Q2IT	**0.571**	**0.691**
w/o Set Retrieval	0.464	0.503

Table 14: Additional ablation analysis of page localization and fine-grained image localization in Q2IT using GPT-5.4 as the backbone.

Setting	1-to-0	1-to-1	1-to-N	Indep.	Comp.	Intra	Cross	Exp.	Imp.	Page F1
Page Localization										
Q2IT	**0.802**	**0.534**	**0.452**	**0.545**	**0.387**	**0.530**	**0.495**	**0.574**	**0.400**	**0.872**
w/o Page Checker	0.531	0.457	0.355	0.453	0.350	0.491	0.349	0.510	0.282	0.799
w/o Page Checker + Page Search	0.548	0.436	0.307	0.431	0.299	0.523	0.234	0.469	0.282	0.740
Fine-Grained Image Localization										
Q2IT	**0.802**	**0.534**	**0.452**	**0.545**	**0.387**	**0.530**	**0.495**	**0.574**	**0.400**	0.872
w/o Detect-then-Select	0.774	0.170	0.078	0.166	0.073	0.151	0.146	0.121	0.204	**0.888**

I.3 Fine-Grained Image Localization

We replace the Figure Agent's detect-then-select strategy with direct bounding-box generation by the MLLM, while keeping the Evidence and Page Agents unchanged. The performance on 1-to-0 instances, which does not require positive image localization, changes only slightly from 0.802 to 0.774, whereas the scores on 1-to-1 and 1-to-N instances drop sharply from 0.534 to 0.170 and from 0.452 to 0.078, respectively. Notably, the MLLM in this setting generates coordinates only over a small set of localized candidate pages, yet its image-localization performance remains close to that of standalone inference over the full document. This further highlights fine-grained localization as a weakness of current MLLMs. The multi-agent workflow mitigates this weakness by introducing external detection tools and adopting a detect-then-select strategy, replacing direct coordinate generation with candidate-region selection and allowing the MLLM to focus on visual reasoning and selection.