VISPROVE: LEARNING TYPED VISUAL PROVENANCE FOR SCIENTIFIC DOCUMENT QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual citations show where an answer is supported, but not how the cited content contributes to it. In scientific document question answering, a model may read a value directly, compress several observations, or combine evidence to infer a conclusion. We study typed visual provenance, where a system generates an answer together with source regions and explicit support relations. We introduce **ReViE**, a dataset of 3,941 questions over complete scientific documents, with page, section, and region annotations labeled as *Quotation*, *Compression*, or *Inference*. We further propose **VisProve**, which augments joint answer and provenance generation with differentiable Spatial Evidence Reading and Evidence–Answer Relation Feedback. VisProve conditions next-token prediction on localized visual evidence and its estimated support relation through a gated norm-bounded residual. On the document-disjoint ReViE test set, VisProve improves Typed-F1 from 3.52 to 5.48 and Section-F1 from 45.99 to 53.24 over joint supervised fine-tuning with the same backbone, while Page-F1 changes only marginally. These results suggest that coupling visual grounding with support relation modeling mainly improves fine-grained provenance, while the low absolute Typed-F1 reveals a substantial gap between answer generation and precise visual evidence attribution.

1 INTRODUCTION

An answer to a scientific question is easier to assess when its evidence is visible. However, locating the correct table, figure, or text region does not fully explain how that evidence supports the answer. A model may read a single entry directly, summarize several observations, or compare multiple regions to derive a conclusion. These cases require different forms of verification. A bounding box therefore identifies where to look, but not how the cited region contributes to the answer. This distinction is especially important in scientific document question answering, where text, figures, and tables may provide complementary evidence across pages (Ma et al., 2024; Yu et al., 2026c).

Existing visual grounding benchmarks make supporting pages and regions inspectable (Yu et al., 2026c; Ma et al., 2026; Venna et al., 2026). In parallel, textual provenance research distinguishes how a source is used, including direct restatement, compression, and inference (Zhu et al., 2025; Wei et al., 2026b). These two lines of work capture complementary aspects of provenance, but they are usually modeled separately. We ask whether explicitly modeling both visual source locations and support relations can improve provenance beyond learning to serialize answers and citations jointly. Such relation labels make the proposed use of evidence explicit and evaluable, while remaining distinct from a claim that the evidence itself is correct.

We study *fine-grained typed visual provenance*. Given a question and the complete ordered sequence of document page images, a model generates an answer and a set of evidence items. Each item specifies a page, a section, a bounding box, and a support relation. We use *Quotation*, *Compression*, and *Inference* to distinguish information that is directly available in a region, information that is faithfully condensed from the evidence, and conclusions that require reasoning over the evidence. Fine granularity refers to the localization of regions rather than to sentence or claim decomposition of the answer. Spatial extent and semantic use are treated separately. For example, a directly stated sentence that spans two pages remains Quotation even though its evidence occupies multiple regions.

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.

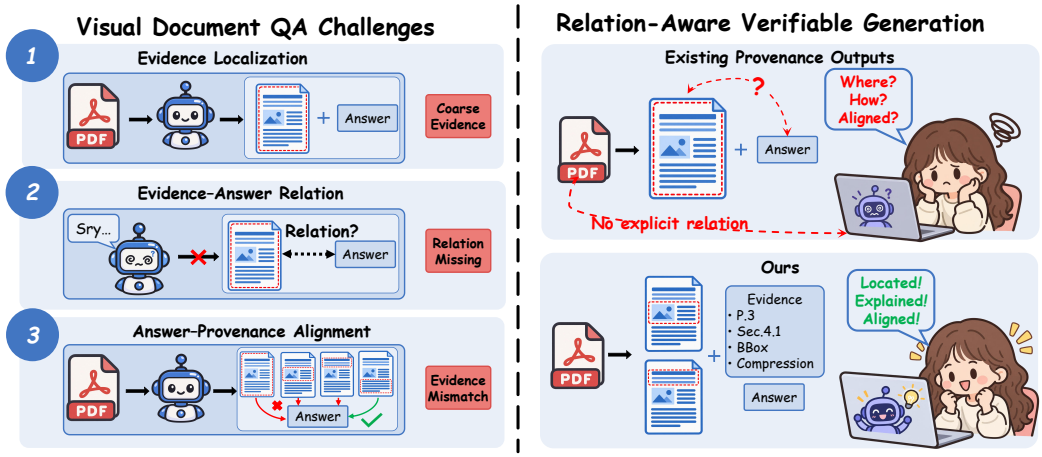

Figure 1: Overview of fine-grained typed visual provenance. The model links a complete answer to supporting visual regions and specifies how each region contributes to the answer.

To support this task, we construct **ReViE** from SciEGQA (Yu et al., 2026c). ReViE combines refined question-answering pairs from SciEGQA with newly constructed, visually verified multi-hop questions and a separate set of cross-page-contiguous questions. The dataset contains 3,941 instances linked to 8,523 visual evidence regions, with page, section, region, and support relation annotations under document-disjoint splits. We restore the complete ordered page sequence for every document, so models must locate sparse supporting evidence within the full document rather than reason only over preselected gold evidence.

We further propose **VisProve**, which couples evidence reading with support relation modeling during generation. Joint supervised fine-tuning (SFT) learns answer content and structured provenance, but it does not explicitly connect a predicted location to a differentiable representation of the visual content at that location. *Spatial Evidence Reading* (SER) addresses this gap by constructing an evidence representation from a page distribution and a latent spatial mask, both conditioned on the question and the current decoder state. *Evidence–Answer Relation Feedback* (ERF) maps this evidence representation to a distribution over support relations. VisProve then integrates the evidence and relation representations through a gated norm-bounded residual before the original language head. The model continues to generate both answers and provenance fields autoregressively, and its internal location and relation estimates guide generation without replacing the emitted citations.

Our main comparison is against joint supervised fine-tuning using the same backbone and training split. On the document-disjoint ReViE test set, VisProve improves Typed-F1 from 3.52 to 5.48 and Section-F1 from 45.99 to 53.24, while Page-F1 changes from 75.00 to 75.73. The larger gains at the section and typed levels suggest that the benefit is concentrated in finer provenance rather than coarse page localization. Feedback ablations also favor the full configuration for typed provenance, although improvements in answer quality are not uniform across metrics. Overall, these results suggest that explicit evidence reading and relation modeling provide complementary signals, while the low absolute Typed-F1 highlights the remaining difficulty of precise typed visual provenance.

Our contributions are summarized as follows.

- We introduce ReViE, which combines hierarchical visual source localization with region-level support relations under complete document input.
- We propose VisProve, which connects differentiable spatial evidence reading and support relation feedback to joint answer and provenance generation.
- We provide shared-backbone comparisons against joint SFT, quantify gains in fine-grained provenance, and expose the substantial remaining gap in precise typed visual provenance.

2 Related Work

Visual document question answering. Document QA has progressed from OCR- and layout-aware representations (Mathew et al., 2021; Xu et al., 2021; Powalski et al., 2021) to direct understanding of document images (Kim et al., 2022; Lee et al., 2023; Ye et al., 2023). Multi-page models address cross-page aggregation and visual compression (Blau et al., 2024; Hu et al., 2025), while long-document benchmarks examine understanding and reasoning over extended visual contexts (Ma et al., 2024; Deng et al., 2025). Retrieval, interactive navigation, and multi-step reasoning offer complementary strategies for accessing relevant evidence (Tanaka et al., 2025; Zheng et al., 2026; Lin et al., 2026; He et al., 2026). UniDoc-RL (Wang et al., 2026) combines retrieval, reranking, active visual perception, and reasoning within a reinforcement-learning framework, progressively narrowing evidence from documents to images and local regions. Complementing these evidence-acquisition pipelines, SCVER (Chai et al., 2026) enables decoder-state-dependent access to fine-grained visual features during autoregressive document parsing. Our setting takes the complete, ordered page-image sequence of a scientific document as input and requires an answer together with structured visual provenance, making evidence localization part of the output rather than only an intermediate processing step.

Visual evidence grounding and provenance. Visual grounding associates answers with spatial support, including page references and bounding boxes (Ma et al., 2025; Zhou et al., 2025; Ma et al., 2026). SciEGQA provides semantic-region evidence for scientific-document QA (Yu et al., 2026c), while M3Grounder formulates multi-span and multi-granular grounding as pixel-level segmentation (Venna et al., 2026). Grounding supervision also supports evidence-aware chart and document reasoning (Vogel et al., 2025; Xiong et al., 2026). Beyond localization alone, LAT (Liu et al., 2026b) and Chain of Evidence (Liu et al., 2026a) integrate attribution into reasoning or generation. DocSeeker (Yan et al., 2026) organizes document reasoning into analysis, evidence localization, and reasoning, while CoExVQA (Indrehus et al., 2026) explicitly localizes answer regions and restricts answer decoding to the grounded content. VinQA (Jang et al., 2026) interleaves cited visual elements with long-form answers, and DocTrace (Xiang et al., 2026) builds hierarchical evidence graphs with node-level provenance. More recently, DocAttriBench (De Grandis et al., 2026) constructs element-level grounding annotations using changes in answer perplexity after masking candidate document elements. VisProve focuses on jointly predicting hierarchical visual locations and the semantic support relation of each evidence region to the complete answer, while using relation feedback to condition generation.

Textual provenance and evidence–answer relations. Evidence-based text generation studies how generated content can be connected to supporting sources, with diverse choices of attribution granularity and evaluation (Schreieder et al., 2026). Sentence- and span-level attribution provide localized source links (Zhang et al., 2025; Hirsch et al., 2025), while ALiiCE examines fine-grained citation placement within generated sentences (Xu et al., 2025). Beyond identifying sources, TROVE annotates evidence-use relations, including *Quotation*, *Compression*, and *Inference* (Zhu et al., 2025). GenerationPrograms makes source transformations explicit through executable text operations (Wan et al., 2025), and GenProve jointly generates answers and sentence-level textual provenance with semantic support relations (Wei et al., 2026b). Extending this perspective to multimodal agents, TRACER links generated sentences to supporting tool observations and their evidence-use relations (Yu et al., 2026a). ReViE adopts relation-aware provenance for native scientific-document images while retaining complete-answer supervision: a complete answer is linked to a set of visual evidence regions rather than decomposed into individually attributed sentences. Each evidence item specifies a page, section, bounding box, and relation. The resulting task couples locating visual support with identifying whether it supplies directly stated information, a faithful condensation, or an inference supporting the answer.

3 ReViE: Relation-aware Visual Evidence

3.1 Task and Annotation Schema

Given a question q and an ordered sequence of document page images $D = (I_1, \ldots, I_M)$, the task is to generate an answer a and an evidence set $E = \{(p_k, s_k, b_k, r_k)\}_{k=1}^{K}$.

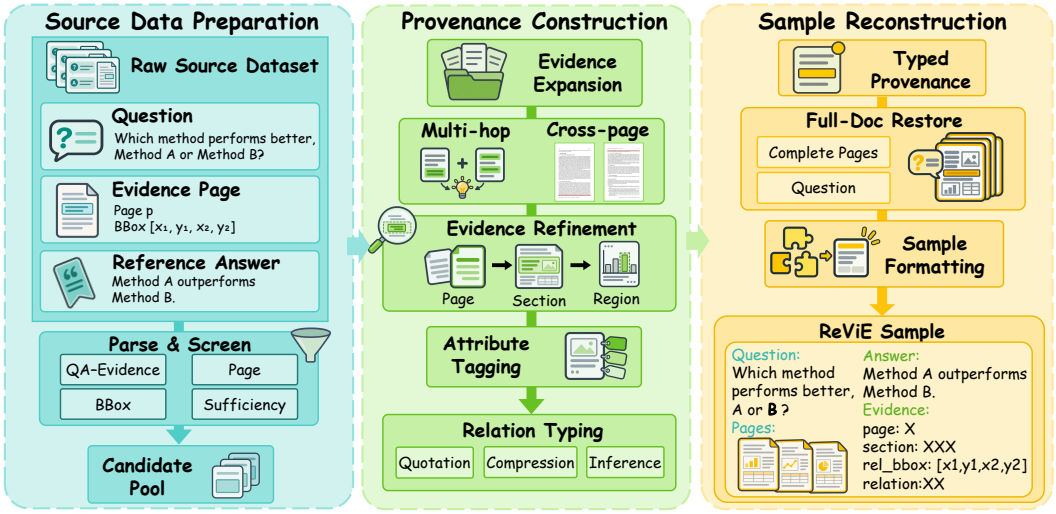

Figure 2: ReViE construction. Source QA pairs and newly generated, verified multi-hop QA pairs undergo fixed-QA evidence refinement. A later branch constructs additional QA pairs from contiguous evidence across adjacent pages. Supporting regions receive relation annotations, and complete ordered document pages are restored for every retained instance.

Each item records its 1-based page index p_k, section or subsection title s_k, bounding box b_k, and relation $r_k \in \{\textit{Quotation}, \textit{Compression}, \textit{Inference}\}$. The serialized fields are `page`, `section`, `rel_bbox`, and `relation`; box coordinates use a 0–1000 page-relative scale, and an unresolved section uses JSON `null`.

The evidence set supports the complete answer. Each region is labeled according to the answer content it contributes, and multiple regions can jointly support a conclusion. *Quotation* covers direct reading and minor rephrasing; *Compression* covers condensation of explicitly available information; and *Inference* covers conclusions requiring comparison, calculation, or reasoning (Zhu et al., 2025; Wei et al., 2026b). The schema associates evidence with the answer as a whole, with source granularity ranging from supporting sentences to necessary figure or table components.

3.2 Dataset Construction

ReViE combines source QA pairs from SciEGQA (Yu et al., 2026c) with newly constructed multi-hop and cross-page-contiguous QA pairs (Figure 2). GPT-5.6 Sol (OpenAI, 2026) assists evidence refinement and relation annotation, complemented by human annotator review and second-round expert validation to ensure evidence quality and annotation consistency.

QA preparation and multi-hop construction. We associate source annotations with page images and screen candidate QA–evidence associations for consistency, correct evidence pages, valid boxes, and sufficient support. In an independent image-direct stage, new questions and reference answers are generated from at least two visual evidence regions. A blind visual judge checks whether the complete answer genuinely requires combining multiple regions. Verified QA pairs are then fixed for subsequent refinement with their question–answer content preserved.

Fixed-QA evidence refinement. This path processes both source QA pairs and previously constructed, verified multi-hop QA pairs. It preserves the question, reference answer, evidence-page identities, and correspondence to the original evidence anchors. Refinement selects complete supporting sentences or necessary figure and table components within the supplied evidence crops. Individual regions contribute support for particular answer content, while the evidence set must collectively support the complete reference answer.

Cross-page-contiguous QA construction. A later branch constructs additional QA pairs from continuous supporting content across adjacent-page boundaries. Valid configurations include a sentence split across pages or semantically continuous adjacent sentences that jointly support the QA

pair. New questions and reference answers are grounded in the boundary evidence without outside knowledge. Cross-page continuity does not imply verified multi-hop status.

Annotation and finalization. Both later processing paths record the exact real section or subsection heading when it can be reliably established; otherwise, they use JSON `null`. Once a QA pair and its evidence are fixed, each region receives a support relation according to its contribution to the complete answer. Source units split across boxes are reconstructed before relation assignment, and uncertain cases enter a second-pass audit. The final quality-control pipeline retains 3,941 of 4,168 candidates after excluding 218 instances requiring box repair and nine with evidence mismatches. We restore complete ordered document pages and partition instances at the document level. Appendix A.2 details quality control, Appendix A.3 describes the construction protocol, and Figure 6 presents the relation-annotation instructions.

3.3 Dataset Analysis

Coverage and spatial structure. ReViE contains 3,941 instances, 1,612 documents, 31,981 unique page images, and 8,523 evidence regions across eight scientific domains. Its document-disjoint training, validation, and test splits contain 3,115, 417, and 409 instances. Single-page single-region (SPSR), single-page multi-region (SPMR), and multi-page multi-region (MPMR) instances account for 35.70%, 46.18%, and 18.12%, respectively. Figures, tables, or mixed evidence occur in 28.88% of instances.

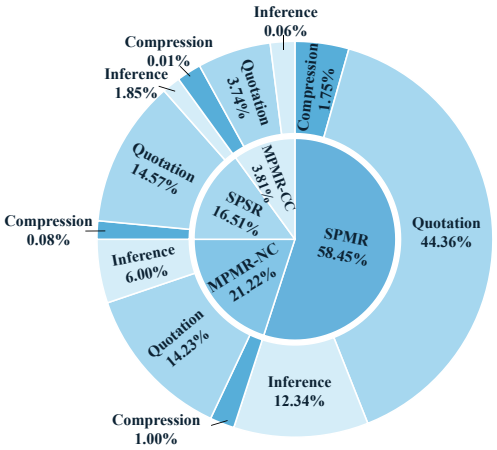

Figure 3: Evidence-region distribution by spatial structure and support relation. Percentages are over 8,523 annotated regions.

Support relations and sparsity. At the region level, *Quotation*, *Compression*, and *Inference* account for 76.91%, 2.84%, and 20.25%. Figure 3 separates MPMR instances without and with a cross-page-contiguity annotation: the former has the highest within-group *Inference* proportion, while the latter mainly contains *Quotation*. An instance has 2.16 evidence regions on 1.18 pages on average, within an 18.28-page input document. The task therefore requires locating a small set of supporting regions within a much longer document. Additional statistics appear in Appendix A.1.

3.4 Benchmark Comparison

Existing benchmarks evaluate multi-page document QA or spatial evidence grounding (Ma et al., 2024; Yu et al., 2026c; Venna et al., 2026). ReViE combines full scientific-document input with page, section, and region annotations and a semantic relation for each evidence item. It additionally records multi-region support, multi-page support, cross-page continuity, and multi-hop configurations. Table 7 compares the documented annotation schemes; these dimensions jointly describe task coverage and the granularity of available supervision.

4 VisProve

VisProve learns joint answer–provenance generation and augments it with localized visual reading and relation feedback. Figure 4 summarizes the framework. SER forms a spatial evidence representation from document features, ERF encodes its support relation, and a gated residual conditions the language head on both signals.

4.1 Joint Answer–Provenance Learning

We fine-tune Qwen3-VL-8B on a unified target containing the complete answer and its structured evidence set. The standard causal language-modeling objective supervises the entire target sequence. This SFT model provides the common generation backbone; its parameters are then frozen while

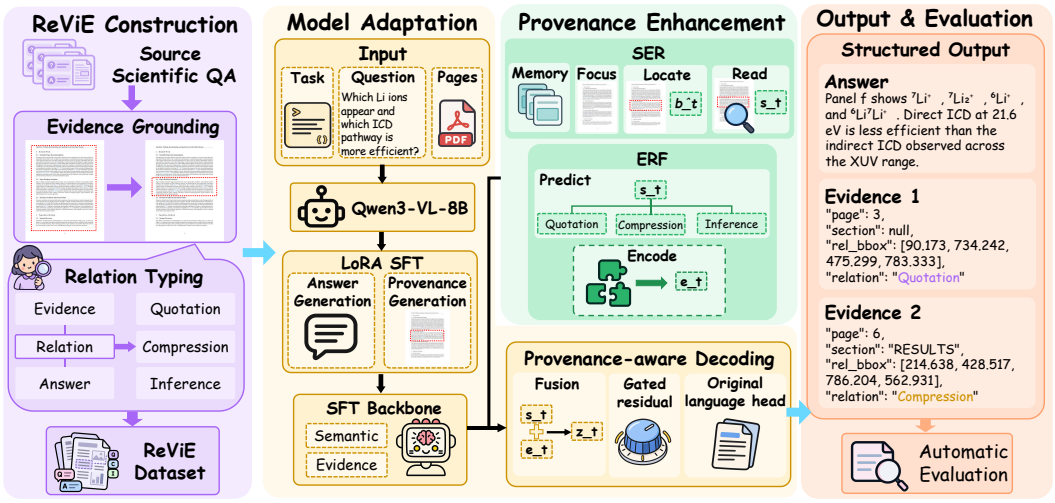

Figure 4: VisProve. Joint SFT learns answer–provenance generation. SER reads localized visual features, ERF encodes their support relation, and a gated residual combines the two representations before the original language head.

the provenance components are trained on the same data. Appendix B.1 specifies the SFT objective and configuration.

4.2 Spatial Evidence Reading

SER connects a latent location prediction to differentiable reading from the corresponding spatial feature grid. At generation step t, the decoder state h_t depends on the document, question, and generated prefix.

For each page j, frozen visual features are pooled into a 16×16 grid and projected to a $d = 128$-dimensional visual memory $\{v_{jk}\}$ with spatial positional encodings. The memory contains all document pages. Using h_t and the mean question embedding $\bar{e}_q$, SER computes a query and soft page distribution:

$$u_t = W_q \,\mathrm{LN}(h_t) + W_{\text{question}} \bar{e}_q, \qquad \pi_{tj} = \mathrm{softmax}_j \left(\frac{u_t^\top \bar{v}_j}{\sqrt{d}} \right), \tag{1}$$

where $\bar{v}_j$ is the mean visual-memory representation of page j.

A box predictor maps u_t to an ordered normalized box $\hat{b}_t$. Smooth boundary functions convert this box into spatial weights α_{tjk}, which pool local evidence alongside page-level context:

$$\begin{aligned} r_t^{\text{vis}} &= \sum_j \pi_{tj} \sum_k \alpha_{tjk} v_{jk}, \qquad c_t = \sum_j \pi_{tj} \bar{v}_j, \\ s_t &= \mathrm{GELU}\left(W_f [u_t;\, r_t^{\text{vis}};\, c_t] \right) + \mathrm{MLP}_{\text{geometry}}(\hat{b}_t). \end{aligned} \tag{2}$$

The spatial representation s_t combines local appearance, page context, the question-conditioned query, and box geometry. The internal box guides feature reading, while the final output coordinates are separately generated by the language head. Appendix B.2 defines the box parameterization and spatial weights.

4.3 Evidence–Answer Relation Feedback

Given the question-conditioned spatial representation s_t, ERF predicts a distribution over three support relations and maps it to a continuous feedback embedding for next-token prediction:

$$z_t^{\text{rel}} = W_{\text{rel}} s_t + b_{\text{rel}}, \quad \rho_t = \mathrm{softmax}(z_t^{\text{rel}}), \quad e_t = W_{\text{type}} \rho_t. \tag{3}$$

The entries of $\rho_t \in \mathbb{R}^3$ correspond to *Quotation*, *Compression*, and *Inference*. Since s_t incorporates both visual evidence and h_t, the relation estimate is conditioned on the question and generated prefix. The embedding e_t provides semantic feedback to generation; the final relation field is produced by the original language head.

4.4 Joint Fusion and Training Objective

Gated residual fusion. The full model combines the spatial and relation representations as

$$z_t = s_t + e_t. \tag{4}$$

It projects this representation to the decoder dimension and applies a bounded residual update:

$$\Delta h_t = g_t \lambda_t W_{\mathrm{up}} z_t, \qquad h'_t = h_t + \Delta h_t. \tag{5}$$

The learned gate g_t scales the update, and λ_t limits its pre-gated norm to 3% of $\|h_t\|_2$. The update is applied after the final decoder layer and before the original output head, which predicts the next token from h'_t. Appendix B.2 provides the gate and normalization formulas. Section 5.1 defines the feedback variants.

Training objective. With the SFT backbone frozen, the added components are optimized using the structured generation objective together with auxiliary provenance supervision:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{loc}}\, \mathbb{I}_{\mathrm{SER}}\, (\mathcal{L}_{\mathrm{page}} + \mathcal{L}_{\mathrm{box}}) + \lambda_{\mathrm{rel}}\, \mathbb{I}_{\mathrm{ERF}} \mathcal{L}_{\mathrm{rel}} \\ + \lambda_{\mathrm{field}} \mathcal{L}_{\mathrm{field}} + \lambda_{\mathrm{outside}} \mathcal{L}_{\mathrm{outside}} + \lambda_{\mathrm{gate}}\, \mathbb{I}_{\mathrm{SER}} \mathcal{L}_{\mathrm{gate}}.\end{aligned} \tag{6}$$

The indicators $\mathbb{I}_{\mathrm{SER}}$ and $\mathbb{I}_{\mathrm{ERF}}$ specify whether the corresponding components are optimized. The page and box losses supervise latent localization, the relation loss supervises the three-way relation prediction, and the remaining losses weight provenance fields and regularize residual feedback. Auxiliary supervision pairs serialized evidence items with their gold annotations in order and is applied at configuration-specific decoder positions. Gold annotations provide training targets; inference uses the question and document pages. Appendix B.2 specifies the supervision positions, trainable components, loss reductions, and weights.

5 Experiments

5.1 Experimental Setup

Data and models. We train on the 3,115-instance ReViE training split and evaluate on the document-disjoint 409-instance test split. Models receive the question and complete ordered document pages. General-purpose baselines comprise Qwen3-VL (8B and 32B) (Bai et al., 2025a), InternVL3.5 (8B and 38B) (Wang et al., 2025), MiniCPM-V4.5-8B (Yu et al., 2026b), Qwen2.5-VL-7B (Bai et al., 2025b), Eagle2.5-8B (Chen et al., 2025), GPT-4o-mini (OpenAI, 2024), Claude Haiku 4.5 (Anthropic, 2025), and Gemini 2.5 Flash-Lite (Google DeepMind, 2025). These models and the VisProve variants use the prompt in Appendix B.4.

Training and feedback variants. VisProve uses LoRA-based SFT (Hu et al., 2021) of Qwen3-VL-8B followed by provenance-component training with the SFT backbone frozen. The base model is included in the main comparison, while Default SFT provides the common backbone for the ablation variants. *w/o ERF* uses spatial feedback only ($z_t = s_t$) and disables ERF. *w/o SER* uses relation feedback only ($z_t = e_t$), retaining a frozen spatial reader to supply ERF's input but excluding s_t from direct residual feedback. The full model jointly optimizes and injects both representations. Auxiliary field losses follow the active components, as specified in Appendix B.2. All configurations generate their final fields through the language head, without post-hoc prediction repair.

Metrics. Answer quality is measured by ROUGE-L (Lin, 2004) and BERTScore-F1 (Zhang et al., 2019); localization by Page-F1, Section-F1, Evidence-IoU, and Region-F1; and typed provenance by Relation-F1 and Typed-F1. All are reported on a 0–100 scale. Typed-F1 evaluates all test instances using page-, section-, and relation-consistent region matching. Evidence-IoU, Region-F1, and Relation-F1 are conditional averages over each model's instances with an exactly correct evidence-page set; Section-F1 uses instances with a non-null reference section. Overall averages the answer, localization, and typed-provenance category scores with equal weight. A text-only judge scores answer quality on 50 shared questions using a 0–5 scale; its mean score is reported separately as Judge. Appendix C defines matching and aggregation.

Table 1: Main results on ReViE. E-IoU, R-F1, and Rel-F1 condition on exact page-set prediction; T-F1 uses all 409 test instances. Overall equally averages the three metric categories. Judge uses 50 shared questions. Light orange and light green mark the best and second-best values, respectively.

Model	Answer		Localization				Typed Prov.		Summary	
	R-L↑	B-F1↑	P-F1↑	S-F1↑	E-IoU↑	R-F1↑	Rel-F1↑	T-F1↑	Overall↑	Judge↑
Open-source models										
Qwen3-VL-8B	48.27	86.08	66.55	31.90	1.32	1.60	0.73	0.18	30.99	3.04
InternVL3.5-8B	26.63	84.26	21.81	14.35	0.54	0.76	0.66	0.00	21.71	2.66
MiniCPM-V4.5-8B	10.73	80.99	1.88	4.63	0.00	0.00	0.00	0.00	15.83	1.88
Qwen3-VL-32B	33.24	84.70	66.16	37.01	1.77	2.20	0.93	0.87	28.89	3.34
InternVL3.5-38B	4.98	79.91	4.38	3.77	1.48	2.36	2.36	0.00	15.54	2.64
Qwen2.5-VL-7B	44.39	88.09	32.29	12.04	0.02	0.04	0.00	0.04	25.79	2.44
Eagle2.5-8B	35.86	85.69	28.13	13.89	0.97	1.36	1.30	0.08	24.18	2.24
Closed-source models										
GPT-4o-mini	18.88	85.67	31.46	14.35	0.45	0.59	0.13	0.23	21.39	1.36
Claude Haiku 4.5	15.76	84.86	36.35	11.54	2.34	2.73	0.97	0.17	21.37	1.48
Gemini 2.5 Flash-Lite	23.41	84.52	73.81	32.10	6.59	7.08	6.68	3.21	29.60	3.42
VisProve (Ours)	**67.08**	**94.18**	**75.73**	**53.24**	**8.58**	**9.68**	**9.22**	**5.48**	**41.60**	**3.46**

Table 2: Ablation study on ReViE. Metric definitions, evaluation conditions, and best/second-best highlighting follow Table 1.

Model	Answer		Localization				Typed Prov.		Summary	
	R-L↑	B-F1↑	P-F1↑	S-F1↑	E-IoU↑	R-F1↑	Rel-F1↑	T-F1↑	Overall↑	Judge↑
VisProve (Ours)	67.08	**94.18**	**75.73**	**53.24**	**8.58**	**9.68**	**9.22**	**5.48**	**41.60**	3.46
w/o ERF	66.85	93.92	75.21	45.06	6.66	7.60	7.03	3.73	39.80	**3.56**
w/o SER	**67.82**	93.96	74.69	48.61	6.86	7.94	7.29	4.11	40.37	3.54
Default SFT	64.24	93.85	75.00	45.99	5.07	5.91	5.22	3.52	38.80	3.52

5.2 Main Results and Ablation Study

Table 1 compares VisProve with general-purpose open- and closed-source VLMs, while Table 2 reports the comparison with Default SFT and the two feedback variants.

Task adaptation and typed provenance. Tables 1 and 2 show gains from task-specific SFT and provenance feedback. Relative to the Qwen3-VL-8B base, Default SFT raises ROUGE-L from 48.27 to 64.24 and Typed-F1 from 0.18 to 3.52. VisProve further improves Typed-F1 to 5.48 and Section-F1 from 45.99 to 53.24, while Page-F1 increases only modestly from 75.00 to 75.73, indicating larger gains in finer provenance outputs than in page identification. In the main comparison, VisProve achieves the highest Overall score of 41.60 and leads all eight task-specific metrics.

Feedback configurations. Table 2 compares the feedback variants on a shared SFT backbone. Spatial feedback only (*w/o ERF*) reaches 3.73 Typed-F1, while relation feedback with a frozen spatial reader (*w/o SER*) reaches 4.11; combining both representations yields 5.48. VisProve also raises Overall from 38.80 for Default SFT to 41.60 and leads seven of the eight task-specific metrics, although *w/o SER* achieves the highest ROUGE-L and *w/o ERF* the highest Judge score. These results support the full configuration's advantage in typed provenance without implying uniform gains across answer-quality measures; the variants differ in feedback and associated training configurations rather than the SFT backbone.

5.3 Comparison with Existing Methods

Table 3 compares VisProve with VISA (Ma et al., 2025), Chain of Evidence (CoE) (Liu et al., 2026a), and UniDoc-RL-7B (Wang et al., 2026) in their shared answer–page–region output space. Their native inference procedures are evaluated with ROUGE-L, BERTScore-F1, Page-F1,

Evidence-IoU, and Region-F1. The comparison-specific score $\text{Overall}_{\text{shared}}$ equally weights answer quality and localization (Appendix C.3).

Table 3: Comparison in the shared answer–page–region output space. $\text{Overall}_{\text{shared}}$ uses two categories and differs from Overall in Table 1. E-IoU and R-F1 retain the page-exact conditioning. Light orange and light green mark the best and second-best values, respectively.

Method	Answer		Localization			Summary
	R-L↑	B-F1↑	P-F1↑	E-IoU↑	R-F1↑	$\text{Overall}_{\text{shared}}$ ↑
VISA (Ma et al., 2025)	11.65	85.53	7.50	0.02	0.02	25.55
CoE (Liu et al., 2026a)	10.45	84.57	30.42	1.37	1.55	29.31
UniDoc-RL-7B (Wang et al., 2026)	28.37	87.10	65.31	0.60	0.77	39.98
VisProve (Ours)	**67.08**	**94.18**	**75.73**	**8.58**	**9.68**	**55.98**

Results. VisProve has the highest values on all five shared metrics and reaches 55.98 $\text{Overall}_{\text{shared}}$. Among the baselines, UniDoc-RL achieves the highest answer-quality and page-level scores, while CoE obtains higher conditional region-overlap scores (Table 3). This comparison measures VisProve after ReViE-specific training against the existing methods under their native inference procedures; the shared-backbone comparison with Default SFT provides the task-trained reference point.

5.4 Qualitative Error Analysis

The examples in Appendix D show several discrepancies between answers and provenance. One prediction gives a correct chart-reading answer but cites adjacent prose. Another omits a supporting page and misidentifies two ion species. A third omits a legend region, gives an underspecified answer, and predicts a relation different from the annotation. These cases show that answer correctness does not necessarily imply accurate visual provenance, and that localization, evidence coverage, and relation typing can fail jointly. A mismatch with the annotated evidence also does not necessarily exclude other valid supporting regions. These observations motivate evaluating source location, evidence coverage, and semantic support alongside answer quality.

5.5 Consistency with Human Judgments

Human reviewers assess 220 answers from 11 models on the same 20 test questions, using the question, reference answer, and predicted answer. Reviewers score answers independently without access to model identities or automated scores, and the final human rating for each answer is the median of its individual ratings. The model-level mean ratings have Pearson correlation $r = 0.9898$ with Judge scores on those same questions (Figure 5). This analysis concerns agreement across 11 model averages in reference-based answer assessment. Appendix C.5 describes the independent ratings and aggregation procedure.

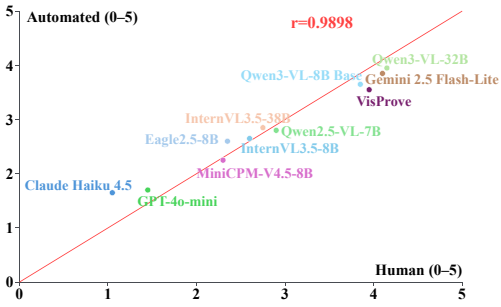

Figure 5: Consistency between LLM-as-Judge scores and human ratings across 11 evaluated models.

6 Conclusion

Visual provenance should reveal both where supporting evidence appears and how it contributes to an answer. ReViE makes source localization and support relations jointly evaluable, while VisProve couples differentiable spatial evidence reading with support relation feedback during generation. Using the same backbone and training split, VisProve improves typed provenance and section-level localization over joint SFT, although gains in answer quality are not uniform. Meanwhile, the low absolute typed provenance scores highlight a substantial gap between producing a correct answer and precisely attributing its visual support. These results support learning evidence localization and evidence use jointly, while treating provenance as an evaluation target distinct from answer quality.

AI USE STATEMENT

Generative AI tools were used to assist dataset construction, including QA construction, visual-evidence refinement, multi-hop verification, and evidence–answer relation annotation. The authors take responsibility for the final dataset, reported results, and conclusions.

REPRODUCIBILITY STATEMENT

Appendix A documents dataset construction, statistics, and quality-control procedures; Appendix B provides training and inference configurations; and Appendix C specifies metric definitions and automated and human evaluation protocols.

REFERENCES

Anthropic. Claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. Gram: Global reasoning for multi-page vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15598–15607, 2024.

Mingxu Chai, Chenyu Liu, Ziyu Shen, Jiazheng Zhang, Kaidi Zhang, Ruoyu Chen, Jun Long, Jihua Kang, Tao Gui, and Qi Zhang. State-conditioned visual evidence retrieval for fine-grained perception in document vision-language models. *arXiv preprint arXiv:2608.28698*, 2026.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *Advances in Neural Information Processing Systems*, 2025.

Luca De Grandis, Silvia Cappelletti, William Raccagni, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Docattribench: Benchmarking answer grounding in document visual question answering. *arXiv preprint arXiv:2609.20574*, 2026.

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.

Google DeepMind. Gemini 2.5 flash-lite. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite, 2025.

Liqi He, Zuchao Li, Hao Huang, and Ping Wang. Trace: Traversal retrieval-augmented chain of evidence for document understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9806–9825, 2026.

Eran Hirsch, Aviv Slobodkin, David Wan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. Laquer: Localized attribution queries in content-grounded generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15355–15370, 2025.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5817–5834, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kjetil Indrehus, Adrian Duric, Changkyu Choi, and Ali Ramezani-Kebrya. Towards self-explainable document visual question answering with chain-of-explanation predictions. *arXiv preprint arXiv:2605.06058*, 2026.

Young Rok Jang, Hyesoo Kong, Kyunghwan An, Jae Sub Huh, Gyeonghun KIM, and Stanley Jungkyu Choi. Vinqa: Visual elements interleaved long-form answer generation for real-world multimodal document qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41130–41139, June 2026.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jiahang Lin, Kai Hu, Binghai Wang, Yuhao Zhou, Zhiheng Xi, Honglin Guo, Shichun Liu, Junzhe Wang, Shihan Dou, Enyu Zhou, et al. Mm-doc-r1: Training agents for long document visual question answering through multi-turn reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29770–29783, 2026.

Peiyang Liu, Ziqiang Cui, Xi Wang, Di Liang, and Wei Ye. Chain of evidence: Pixel-level visual attribution for iterative retrieval-augmented generation. *arXiv preprint arXiv:2605.01284*, 2026a.

Shuochen Liu, Pengfei Luo, Chao Zhang, Yuhao Chen, Haotian Zhang, Qi Liu, Xin Kou, Tong Xu, and Enhong Chen. Look as you think: Unifying reasoning and visual evidence attribution for verifiable document rag via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32159–32167, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30154–30169, 2025.

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 C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, January 2021.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, pp. 732–747. Springer, 2021.

Tobias Schreieder, Tim Schopf, and Michael Färber. Attribution, citation, and quotation: A survey of evidence-based text generation with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30956–31000, 2026.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24827–24837. IEEE, 2025.

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

Venkata Kesav Venna, Sai Madhusudan Gunda, Jyothi Swaroopa Jinka, Hrithik Sagar Rachakonda, Anirudh Srinivasan, and Ravi Kiran Sarvadevabhatla. M3grounder: Mask-based multi-span and multi-granular grounding for document qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23685–23695, 2026.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. Refchartqa: Grounding visual answer on chart images through instruction tuning. In *International Conference on Document Analysis and Recognition*, pp. 523–537. Springer, 2025.

David Wan, Eran Hirsch, Elias Stengel-Eskin, Ido Dagan, and Mohit Bansal. Generationprograms: Fine-grained attribution with executable programs. *arXiv preprint arXiv:2506.14580*, 2025.

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. *arXiv preprint arXiv:2604.14967*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Hongchen Wei, Yuanzhe Wang, Bei Liu, Yifan Yang, Qi Dai, Ruichun Ma, Kai Qiu, Yunsheng Li, Dongdong Chen, Chong Luo, et al. Xl-docbench: Benchmarking evidence-grounded extra-long document understanding. *arXiv preprint arXiv:2608.00036*, 2026a.

Jingxuan Wei, Xingyue Wang, Yanghaoyu Liao, Jie Dong, Yuchen Liu, Caijun Jia, Bihui Yu, and Junnan Zhu. Genprove: Learning to generate text with fine-grained provenance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5027–5048, 2026b.

Le Xiang, Zhicheng Guan, Hong Chen, Xiaocong Lin, Zhenghua Lei, Teng Hu, Bolei He, and Long Zeng. Doctrace: Towards traceable long document vqa via hierarchical evidence graph reasoning. *arXiv preprint arXiv:2608.03292*, 2026.

Junyu Xiong, Yonghui Wang, Weichao Zhao, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. Docr1: Evidence page-guided grpo for multi-page document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11178–11186, 2026.

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.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. Aliice: Evaluating positional fine-grained citation generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 545–561, 2025.

Hao Yan, Yuliang Liu, Xingchen Liu, Yuyi Zhang, Minghui Liao, Jihao Wu, Wei Chen, and Xiang Bai. Docseeker: Structured visual reasoning with evidence grounding for long document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41140–41149, June 2026.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2841–2858, 2023.

Bihui Yu, Caijun Jia, Jing Chi, Xiaohan Liu, Yining Wang, He Bai, Yuchen Liu, Jingxuan Wei, and Junnan Zhu. Tracer: Verifiable generative provenance for multimodal tool-using agents. *arXiv preprint arXiv:2605.09934*, 2026a.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026b.

Wenhan Yu, Zhaoxi Zhang, Wang Chen, Guanqiang Qi, Weikang Li, Lei Sha, Deguo Xia, and Jizhou Huang. Sciegqa: A dataset for scientific evidence-grounded question answering and reasoning, 2026c. URL https://arxiv.org/abs/2511.15090.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. Longcite: Enabling llms to generate fine-grained citations in long-context qa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yuanlei Zheng, Pei Fu, Hang Li, Ziyang Wang, Yuyi Zhang, Wenyu Ruan, Xiaojin Zhang, Zhongyu Wei, Zhenbo Luo, Jian Luan, et al. Doc-v*: Coarse-to-fine interactive visual reasoning for multi-page document vqa. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45901–45923, 2026.

Yinan Zhou, Yuxin Chen, Haokun Lin, Yichen Wu, Shuyu Yang, Zhongang Qi, Chen Ma, and Li Zhu. Dogr: Towards versatile visual document grounding and referring. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3596–3606. IEEE, 2025.

Junnan Zhu, Min Xiao, Yining Wang, Feifei Zhai, Yu Zhou, and Chengqing Zong. Trove: A challenge for fine-grained text provenance via source sentence tracing and relationship classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11755–11771, 2025.

A Additional Dataset Details

A.1 Additional Dataset Statistics

ReViE contains 3,941 instances from 1,612 scientific documents, comprising 31,981 unique page images and 8,523 annotated evidence regions. Table 4 reports the split-level statistics. Splits are constructed at the document level, with no document or sample overlap across training, validation, and test sets.

Table 4: Split-level statistics of ReViE. Page counts refer to unique document pages within each split.

Split	Instances	Documents	Pages	Evidence regions
Train	3,115	1,296	25,439	6,738
Validation	417	161	3,106	921
Test	409	155	3,436	864
Total	**3,941**	**1,612**	**31,981**	**8,523**

Domain distribution. ReViE covers eight scientific domains. Table 5 reports both instance- and document-level distributions, showing broad coverage across computer science, economics, engineering, mathematics, physics, quantitative biology, quantitative finance, and statistics.

Table 5: Domain distribution of ReViE at the instance and document levels.

Domain	Instances	Instance %	Documents	Document %
Computer Science	493	12.51	200	12.41
Economics	330	8.37	169	10.48
Electrical Engineering and Systems Science	662	16.80	232	14.39
Mathematics	107	2.72	59	3.66
Physics	480	12.18	184	11.41
Quantitative Biology	555	14.08	273	16.94
Quantitative Finance	940	23.85	345	21.40
Statistics	374	9.49	150	9.31
Total	**3,941**	**100.00**	**1,612**	**100.00**

Spatial and reasoning structure. At the instance level, ReViE contains 1,407 SPSR instances (35.70%), 1,820 SPMR instances (46.18%), and 714 MPMR instances (18.12%). Among all instances, 604 (15.33%) are labeled as multi-hop and 162 (4.11%) as cross-page contiguous. These attributes describe different properties: multi-hop concerns the necessity of combining evidence regions, whereas cross-page continuity concerns supporting content across an adjacent-page boundary. The two labeled subsets do not overlap in the final dataset. This dataset-specific observation does not imply conceptual incompatibility: multi-hop evidence may span multiple pages without being labeled as cross-page contiguous.

Evidence modality, granularity, and relation distribution. Table 6 summarizes evidence modality at the instance level and evidence–answer relations at the evidence-region level. ReViE is predominantly text-based, while figures, tables, and mixed-modality evidence together account for 28.88% of instances. At the evidence-granularity level, 2,803 instances use sentence-level evidence, 836 use localized visual regions, 24 require complete visual regions, and 278 combine multiple granularities. *Quotation* is the dominant relation type, while *Inference* forms a substantial minority and *Compression* is comparatively rare. A total of 482 instances (12.23%) contain more than one evidence–answer relation type.

Evidence density and document length. Each instance contains 2.16 evidence regions and 1.18 unique evidence pages on average. The complete document input contains 18.28 pages on average (median 14), with a maximum of 234 pages. The median number of evidence regions per instance is 2, and 90% of instances contain at most four evidence regions. No reliable section heading can be

identified for 2,383 evidence regions (27.96%); these cases use JSON `null` in the final annotation schema.

Table 6: Evidence modality and evidence–answer relation distributions in ReViE. Modality statistics are computed over 3,941 instances, whereas relation statistics are computed over 8,523 evidence regions.

Evidence modality			**Evidence–answer relation**		
Modality	**Instances**	**%**	**Relation**	**Regions**	**%**
Text	2,803	71.12	Quotation	6,555	76.91
Figure	632	16.04	Compression	242	2.84
Table	212	5.38	Inference	1,726	20.25
Mixed	294	7.46			
Total	**3,941**	**100.00**	**Total**	**8,523**	**100.00**

A.2 Quality Control and Split Integrity

This section describes qualification, verification, and review procedures. The construction operations and annotation conventions are detailed in Appendix A.3.

Task-specific qualification. Candidate QA–evidence associations are screened for question–answer–evidence consistency, evidence-page correctness, bounding-box validity, and evidence sufficiency. Instances that do not provide reliable visual supervision are excluded from subsequent construction.

Visual and structural verification. An independent visual quality check examines answer support, evidence sufficiency, localization precision, section annotations, and structural metadata. For multi-hop candidates, a blind visual judge checks whether the complete answer genuinely requires combining information from multiple evidence regions. This construction-stage verification is distinct from the text-only answer evaluation in Appendix C.4. Cross-page-contiguous candidates are accepted only when the adjacent-page boundary content provides genuine continuous support.

Relation review and final exclusions. After the QA pair and its visual evidence are fixed, first-pass relation annotation may return `NEEDS_REVIEW` when the supplied evidence is unsupported or the relation is genuinely uncertain. Such cases undergo a separate second-pass audit, which may accept the existing evidence, flag a bounding-box repair requirement or evidence mismatch, or defer the instance for human review. The first-pass relation instructions are shown in Figure 6. Across the final quality-control pipeline, 227 of 4,168 candidate instances are excluded, including 218 requiring bounding-box repair and nine with evidence mismatches, yielding the final 3,941-instance dataset.

Human review and split integrity. The model-assisted construction process is complemented by human annotator review and second-round expert validation. We partition ReViE at the document level so that all instances from the same source document remain within a single split. The training, validation, and test sets contain 3,115, 417, and 409 instances, respectively, with no document or sample overlap across splits. Complete document pages are restored in their original order for every retained instance, and gold evidence pages are not used to preselect model inputs.

A.3 QA and Visual Evidence Construction Protocol

Image-direct multi-hop construction and QA freezing. An independent image-direct stage constructs new questions and reference answers from at least two visual evidence regions. After passing the multi-hop verification described in Appendix A.2, these QA pairs enter fixed-QA evidence refinement together with source QA pairs. Questions and reference answers are frozen for both groups during refinement.

Anchor-preserving evidence refinement. Here, "existing QA" refers to a QA pair established before refinement, not necessarily inherited from the source dataset. Refinement preserves evidence-page identities and correspondence to the original evidence anchors. Each anchor defines a supplied evidence crop within which one or more refined regions are selected. The returned region boundaries

may change, but the source anchor is not replaced and the question is not re-answered. Individual regions support the answer content they contribute, while the complete evidence set must collectively support the full reference answer.

Later cross-page-contiguous QA construction. A separate later branch constructs additional QA pairs from supporting content at adjacent page boundaries, without outside knowledge. It covers a sentence split across pages or semantically continuous adjacent sentences that jointly support the QA pair. In `NEW_CROSS_PAGE_QA` mode, an accepted instance has `cross_page_continuity_valid` set to `true`. Candidate preparation sets `verified_multihop` to `false` for this branch, so continuity alone does not confer a verified multi-hop label. This setting does not, by itself, establish that combining multiple regions is unnecessary.

Evidence completeness. Textual refinement requests complete supporting sentences. For figures and tables, it selects necessary cells, rows, headers, panels, legends, or other local components, retaining complete visual regions when required for sufficient support. When a supporting sentence spans a page boundary, the selected regions jointly cover the complete sentence across both pages; each page-local fragment need not independently form a complete sentence.

Section annotation. Both fixed-QA refinement and cross-page construction use the exact real section or subsection heading when it can be reliably established from the supplied visual context; otherwise, they use JSON `null`. An unresolved heading does not by itself invalidate an instance, and headings must not be fabricated to avoid a `null` value. In fixed-QA refinement, any supplied low-resolution page context is used to identify section headings, not to determine evidence coordinates.

Intermediate localization representations. Fixed-QA refinement returns `crop_bboxes` on a 0–1000 coordinate scale relative to each supplied evidence crop. The cross-page branch selects image-derived line identifiers for textual evidence, from which code computes bounding boxes, and uses page-relative coordinates for figure and table regions. These intermediate fields are distinct from the final `rel_bbox` field, which always uses a 0–1000 scale relative to the complete document page.

A.4 Relation Annotation Prompt

Figure 6 presents the first-pass evidence–answer relation instructions applied after QA construction and visual evidence refinement. At this stage, the question, complete reference answer, and supporting regions are fixed regardless of whether the QA pair was inherited from the source dataset or newly constructed. GPT-5.6 Sol assigns one of *Quotation*, *Compression*, or *Inference* to each evidence item. Cases marked as `NEEDS_REVIEW` follow the second-pass audit described in Appendix A.2.

The displayed output-format block specifies alternative field values and value ranges; it is a schema description rather than a literal JSON response. An actual response selects one allowed value for each field.

First-pass Relation Annotation Prompt

You annotate semantic provenance relations for an **EXISTING, already verified scientific visual-document QA sample**. The Question, complete Reference Answer, and every supplied Evidence region are trusted and **FROZEN**. Do not rewrite the answer, change evidence, add evidence, delete evidence, or assess bbox geometry.
For **EACH** supplied evidence image E1, E2, ..., decide how the visual content in that evidence region participates in supporting the complete Reference Answer. The red rectangle marks the exact evidence region; surrounding pixels are only small reading context.
Choose exactly one relation for each evidence region:

- **Quotation:** The portion of the Answer supported by this evidence can be read directly from visible text, a table cell/value, a label, or an explicit visual statement, with at most minor grammatical or wording changes.
- **Compression:** The supported Answer content faithfully condenses, summarizes, or reorganizes information explicitly present in this evidence (possibly together with other evidence), without introducing a new conclusion.

- **Inference:** The supported Answer content is not explicitly stated in this evidence and requires comparison, calculation, logical deduction, visual trend interpretation, or joint reasoning with other evidence regions.

Important rules:

1. Treat the complete QA pair as one sample. Do **NOT** split or rewrite it.
2. An evidence region need not support every clause in the complete Answer. Judge its relation to the particular Answer content that it actually supports.
3. If multiple evidence boxes are fragments of one continuous source unit because a sentence, table row, or visual element is split by a line, column, or page boundary, reconstruct that unit first and assign the **SAME** relation to every fragment. A directly stated sentence split across boxes remains Quotation; the physical split alone does not make it Compression or Inference.
4. If a conclusion can only be obtained by combining multiple evidence regions, label every region essential to that conclusion as Inference.
5. Multi-hop is not automatically Inference. Combining explicitly stated facts without deriving a new conclusion can be Compression.
6. If one region participates in more than one way, choose the relation required for the most substantive Answer content it supports, using the priority Inference $>$ Compression $>$ Quotation, and explain the mixed case.
7. If a supplied region does not actually support any part of the Answer, or the relation is genuinely indeterminate, return status `NEEDS_REVIEW`.

Output format.
Return JSON only with this exact top-level structure:

```
{   "status":  "VALID" or "NEEDS_REVIEW",
"overall_reason":  "brief sample-level assessment",
"relations":  [
{
"evidence_index":  1,
"relation":  "Quotation" or "Compression" or "Inference",
"supported_answer_content":  "the exact Answer content supported",
"justification":  "brief evidence-based reason",
"confidence":  0.0 to 1.0
}
]
}
```

Return exactly one `relations` item for every supplied evidence image, in index order. Do not include markdown fences or explanations outside the JSON object.

Figure 6: Prompt used for first-pass evidence–answer relation annotation with GPT-5.6 Sol. The prompt assigns a typed provenance relation to each verified visual evidence region before subsequent review and validation.

A.5 Benchmark Comparison Details

Table 7 compares ReViE with DocVQA (Mathew et al., 2021), MP-DocVQA (Tito et al., 2023), SlideVQA (Tanaka et al., 2023), MMLongBench-Doc (Ma et al., 2024), XL-DocBench (Wei et al., 2026a), GroundingDocQA-Bench (Venna et al., 2026), and SciEGQA (Yu et al., 2026c). We compare documented input settings and answer-linked evidence annotations across eight dimensions. GroundingDocQA-Bench is abbreviated as GDQA-Bench and is not conflated with the larger GroundingDocQA training resource.

Input and spatial criteria. Input distinguishes a single page (1), a capped multi-page context (Cap), and a full-document setting (Full). MP-DocVQA uses up to 20 pages, and SlideVQA retains the first 20 slides (Tito et al., 2023; Tanaka et al., 2023). SciEGQA offers a full-document setting alongside evidence-page and evidence-crop settings (Yu et al., 2026c). MR requires multiple answer-linked visual regions, not general OCR or layout boxes. MP denotes evidence on multiple pages; Contig. additionally requires explicit annotation of continuous support across adjacent page boundaries.

Localization and evidence sufficiency. Locator summarizes the evidence representation and its annotated hierarchy. XL-DocBench provides pages and supporting text snippets (Wei et al., 2026a); GroundingDocQA-Bench provides nested phrase-, line-, and block-level masks (Venna et al., 2026); and SciEGQA provides semantically coherent region boxes (Yu et al., 2026c). ReViE adds section

context and refines evidence to supporting sentences or necessary figure/table regions, retaining full regions when required. Section titles are actual document headings or `null`. Refinement seeks precise, collectively sufficient support; multiple valid evidence sets may exist for a question.

Table 7: Comparison of ReViE with seven visual document benchmarks. Columns summarize document context, visual evidence, and reasoning and relation supervision.

	Document		Visual evidence				Supervision	
Benchmark	**Sci.**	**Input**	**Locator**	**MR**	**MP**	**Contig.**	**Hop unit**	**Rel.**
DocVQA	—	1	—	—	—	—	—	—
MP-DocVQA	—	Cap	P	—	—	—	—	—
SlideVQA	—	Cap	P	—	✓	—	Slide	—
MMLongBench-Doc	○	Full	P	—	✓	—	—	—
XL-DocBench	○	Full	P+T	—	✓	—	Sect.	—
GDQA-Bench	—	1	M^h	✓	—	—	—	—
SciEGQA	✓	Full	P+B	✓	✓	—	—	—
ReViE (Ours)	✓	Full	P+S+B	✓	✓	✓	Region	Q/C/I

✓: explicitly provided; ○: scientific content in a mixed-domain corpus; —: not explicitly specified under our criteria. MR/MP: multi-region/multi-page evidence; Contig.: contiguous cross-page annotation. P/S/B/T/M: page/section/box/text snippet/mask; h: phrase–line–block hierarchy. Q/C/I: Quotation/Compression/Inference.

Reasoning and relation criteria. Hop unit reports the evidence granularity used in explicit multi-hop construction or verification: slides for SlideVQA bridge questions (Tanaka et al., 2023), chapter/section branches for XL-DocBench dependency checks (Wei et al., 2026a), and distinct visual regions for ReViE's image-direct multi-hop construction and blind visual verification. Multi-hop status is not inferred from MR, MP, or cross-page continuity alone. Rel. denotes per-evidence support relations, rather than question-level reasoning categories or answer-verification rules. ReViE associates each relation with an evidence item's contribution to the complete answer, without sentence- or claim-level answer decomposition.

B Implementation Details

B.1 SFT Configuration

VisProve is built on Qwen3-VL-8B (Bai et al., 2025a). We first perform LoRA-based supervised fine-tuning (Hu et al., 2021) on the 3,115-instance ReViE training split, while keeping the visual backbone frozen. The model is trained to generate a unified serialized target containing the complete answer together with its structured visual provenance.

Let $y = (y_1, \ldots, y_T)$ denote the complete assistant target. The SFT stage optimizes the standard causal language-modeling objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta \left(y_t \mid D, q, y_{<t} \right). \tag{7}$$

We use LoRA with rank 32, scaling factor 64, and dropout 0.05, applied to linear layers of the language model. Optimization uses AdamW (Loshchilov & Hutter, 2017) with cosine learning-rate scheduling and BF16 precision. The model is trained for three epochs in the main SFT stage with a learning rate of 1×10^{-4}. Training then continues for two epochs in the long-document continuation stage with a learning rate of 1×10^{-5}. The maximum sequence length is 16,384 and gradient clipping is set to 1.0.

The resulting SFT model serves as the common backbone for all provenance-enhancement variants. Its parameters are frozen during subsequent SER and ERF optimization.

B.2 SER AND ERF TRAINING

SER and ERF are optimized using the same 3,115-instance ReViE training split as SFT, with the SFT backbone kept frozen. All provenance-enhancement configurations are trained for two epochs using AdamW (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-4} and weight decay of 0.01. The variants use the same SFT backbone and differ in the optimized components, direct feedback, and associated auxiliary field losses.

Spatial visual memory. For each document page, frozen visual features are adaptively average-pooled into a 16×16 grid containing 256 spatial locations. Each grid feature is normalized, projected to $d = 128$ dimensions, and augmented with a two-dimensional positional encoding, yielding the visual-memory vectors v_{jk} used by SER. All document pages participate in the visual memory; no gold evidence page or external retriever is used to preselect the input.

Latent bounding-box parameterization. Given the SER query u_t, the model predicts

$$z_t^b = f_{\text{box}}(u_t) \in \mathbb{R}^4.$$

To ensure an ordered normalized bounding box, we define

$$l_t = \sigma(z_{t,1:2}^b), \qquad d_t = \sigma(z_{t,3:4}^b), \qquad h_t^b = l_t + (1 - l_t) \odot d_t, \tag{8}$$

and form $\hat{b}_t = [l_t; h_t^b]$ as two ordered corners on the internal $[0, 1]^2$ page plane.

SER converts $\hat{b}_t$ into a differentiable spatial mask over the visual-memory grid. Let c_{jk} denote the center coordinate of location k on page j. We compute

$$\tilde{\alpha}_{tjk} = \prod_{d \in \{x,y\}} \sigma(\kappa(c_{jk,d} - l_{t,d}))\, \sigma\big(\kappa(h_{t,d}^b - c_{jk,d})\big)\,, \tag{9}$$

followed by page-wise normalization:

$$\alpha_{tjk} = \frac{\tilde{\alpha}_{tjk}}{\sum_{k'} \tilde{\alpha}_{tjk'} + \epsilon}. \tag{10}$$

We set $\kappa = 30$, which controls the sharpness of the differentiable region boundary.

Norm-bounded residual feedback. For the representation z_t defined in Equation (4), we first project it to the decoder dimension:

$$\delta_t = W_{\text{up}} z_t. \tag{11}$$

Its magnitude is bounded relative to the original hidden state:

$$\lambda_t = \min\left(1,\, \frac{0.03\|h_t\|_2}{\max(\|\delta_t\|_2, \epsilon)}\right). \tag{12}$$

A learned gate further controls the residual contribution:

$$g_t = \sigma(W_g u_t + b_g), \qquad \Delta h_t = g_t \lambda_t \delta_t, \qquad h_t' = h_t + \Delta h_t. \tag{13}$$

Thus, before gating, the residual norm is bounded by 3% of $\|h_t\|_2$, while g_t may further reduce its contribution. The modification is applied after the final decoder layer and before the original output head.

Training objective. The objective in Equation (6) uses the following weights:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{gen}} + 0.1\,[\mathbb{I}_{\text{SER}}\,(\mathcal{L}_{\text{page}} + \mathcal{L}_{\text{box}}) + \mathbb{I}_{\text{ERF}}\mathcal{L}_{\text{rel}}] \\ + 0.25\mathcal{L}_{\text{field}} + 10\mathcal{L}_{\text{outside}} + 0.1\mathbb{I}_{\text{SER}}\mathcal{L}_{\text{gate}}.\end{aligned} \tag{14}$$

Supervision alignment and trainable components. Auxiliary targets are paired with gold evidence items in their serialized order after checking that the field counts agree. Field-value character spans are mapped to tokenizer positions. Let τ_b, τ_s, and τ_r denote the decoder positions immediately preceding the first coordinate token of `rel_bbox`, the first content token of `section`, and the first content token of `relation`, respectively. Table 8 summarizes the supervision positions and trainable components. In *w/o SER*, the frozen spatial reader still supplies s_t to ERF, but only e_t is

Table 8: Configuration-specific training of provenance-enhancement components. The SFT backbone remains frozen in all variants.

Configuration	Trainable components	Auxiliary supervision	Field CE values
w/o ERF	Spatial reader, geometric fusion, gate, and up-projection	Page and box at τ_b, with two loss contributions	`rel_bbox`
w/o SER	Relation classifier and relation embeddings only	Relation at τ_r	`section`, `relation`
VisProve	All provenance-enhancement components	Page, box, and relation at τ_r; page and box at τ_s	`rel_bbox`, `section`, `relation`

directly injected. Serialized page values are learned through generation cross-entropy, and there is no separate section-title classification head.

Loss computation and reductions. $\mathcal{L}_{\text{gen}}$ is causal cross-entropy averaged over the complete answer–provenance target and its end-of-sequence token. Each page or relation classification loss is averaged over evidence items; each box loss combines mean coordinate-level ℓ_1 error with mean $1 - \text{IoU}$ over evidence items. $\mathcal{L}_{\text{page}}$ and $\mathcal{L}_{\text{box}}$ sum the corresponding contributions listed in Table 8, without averaging over the number of contributions. In *w/o ERF*, the section-associated localization call is mapped to τ_b, so both page-and-box contributions are retained even though they use the same position.

Field masks and residual regularization. $\mathcal{L}_{\text{field}}$ averages causal cross-entropy over tokens overlapping the field-value spans listed in Table 8, including JSON `null` for unresolved sections and punctuation within the selected spans. Both generation and field losses supervise the prediction of each token from its preceding decoder state. The field-token mask also defines positive gate targets, with other output positions assigned negative targets. $\mathcal{L}_{\text{gate}}$ uses binary cross-entropy with equal weight for the positive and negative groups and is included only for *w/o ERF* and VisProve. $\mathcal{L}_{\text{outside}}$ averages the relative residual energy

$$\frac{\|\Delta h_t\|_2^2}{\|h_t\|_2^2} \tag{15}$$

over decoder positions predicting output tokens outside the selected fields. These positions include answer text and field names, which remain supervised by $\mathcal{L}_{\text{gen}}$. Prompt tokens do not contribute to the output losses.

B.3 Inference Configuration

At inference time, the model receives the question together with all pages of the corresponding document in their original order. Gold evidence pages, sections, bounding boxes, and relation labels are not provided. We use greedy decoding with a maximum of 768 newly generated tokens, without sampling or beam search. The answer and all provenance fields are generated autoregressively by the original language head.

For SER and ERF, the per-page visual budget is adjusted according to document length to control the total visual context: documents with at most 12 pages use up to 262,144 pixels per page, documents with 13–32 pages use up to 131,072 pixels per page, and longer documents use up to 65,536 pixels per page.

No post-hoc bounding-box calibration, section lookup, relation rewriting, or other prediction repair is applied. The latent bounding box predicted by SER is used only for internal differentiable visual reading and does not overwrite the bounding box generated in the final structured output.

The decoding and visual-budget settings in this subsection apply to VisProve and its two feedback variants. General-purpose baselines use the shared task prompt in Appendix B.4 with model-specific image processing, while VISA, CoE, and UniDoc-RL follow their native inference procedures as described in Section 5.3.

B.4 Unified Inference Prompt

Figure 7 shows the unified task prompt used for general-purpose open- and closed-source VLM baselines and for the SFT/SER/ERF variants. Each instance additionally provides all document pages in their original order together with the question. Reference answers and gold provenance annotations are not included at inference time. Specialized methods in Section 5.3 follow their respective native inference procedures.

Unified Inference Prompt

You are a scientific document question-answering assistant with visual provenance grounding.
The input images together form a single scientific document and are provided in page order. Answer the question using only the information contained in the document.
You must output both the answer and the minimum sufficient visual evidence that supports the answer. The evidence array must not be empty.
For each evidence item:

- `"page"` is the 1-based position of the corresponding page in the input image sequence.
- `"section"` must be the actual section or subsection title shown in the document. If it cannot be reliably determined, use the JSON value `null`, not the string `"null"`.
- `"rel_bbox"` must contain four numeric values in the range 0–1000, formatted as `[x1, y1, x2, y2]`, where $x1 < x2$ and $y1 < y2$. Coordinates may be integers or decimals; do not assume they must be integers.
- For textual evidence, the bounding box must cover the complete supporting sentence or sentences.
- For figures and tables, use the minimum sufficient visual region that supports the answer. Use the full figure or table only when necessary.
- `"relation"` must be exactly one of `"Quotation"`, `"Compression"`, or `"Inference"`.
- `"relation"` describes how this specific evidence item supports the complete answer. When multiple evidence items are provided, label each item separately.

Relation definitions:

- **Quotation:** the relevant answer information can be directly read from the evidence or is only lightly rephrased.
- **Compression:** the answer faithfully summarizes, condenses, or combines information explicitly present in the evidence without introducing a new logical conclusion.
- **Inference:** the answer requires comparison, calculation, ordering, logical deduction, conditional reasoning, or integration of evidence to derive a conclusion that is not directly stated.

Return strictly valid JSON with exactly two top-level fields, `"answer"` and `"evidence"`. Each evidence item must contain exactly `"page"`, `"section"`, `"rel_bbox"`, and `"relation"`.
Do not output Markdown, explanations, comments, or any text outside the JSON object.

Figure 7: Unified inference prompt used for general-purpose VLM baselines and VisProve variants. The prompt specifies complete-document input, structured answer–provenance output, visual localization requirements, and evidence–answer relation types.

C Evaluation Details

C.1 Metric Definitions

We evaluate models along three complementary dimensions. *Answer quality* is measured by ROUGE-L and BERTScore-F1. *Evidence localization* is evaluated using Page-F1, Section-F1, Evidence-IoU, and Region-F1, covering page identification, section localization, and fine-grained visual-region overlap. *Typed provenance* is measured by Relation-F1 and Typed-F1, which assess relation-consistent evidence prediction and complete typed provenance, respectively. All task-specific metrics are reported on a 0–100 scale, with higher values indicating better performance.

Answer quality. ROUGE-L is computed between the predicted and reference answers after basic Unicode and whitespace normalization. BERTScore-F1 measures their semantic similarity using contextual representations. Both metrics are evaluated over all 409 test instances.

Page and section localization. Page-F1 treats the predicted and reference page indices as sets and computes

$$\text{Page}-\text{F1} = \frac{2|\hat{\Pi} \cap \Pi|}{|\hat{\Pi}| + |\Pi|}, \tag{16}$$

where $\hat{\Pi}$ and Π denote the predicted and reference evidence-page sets, respectively. Page-F1 is averaged over all test instances.

Section-F1 similarly compares the sets of predicted and reference non-`null` section names. Section strings are normalized using Unicode normalization, case folding, whitespace normalization, and removal of leading numerical section identifiers. Section-F1 is evaluated only on instances containing at least one non-`null` reference section; a missing predicted section receives zero credit for such an instance.

Region and typed provenance metrics. Evidence-IoU, Region-F1, Relation-F1, and Typed-F1 use one-to-one matching between predicted and reference evidence regions. Bounding boxes are evaluated directly in the normalized coordinate space, and region similarity is measured by continuous IoU rather than thresholding at a fixed IoU value. The detailed matching procedure is described in Appendix C.2.

We additionally report a text-only LLM-as-Judge score on a shared 50-instance subset of the test set. The judge assigns an integer score from 0 to 5 to the predicted answer. This score is reported separately and is not included in the Overall metric.

C.2 Evidence Matching and Failure Handling

Let P and G denote the numbers of predicted and reference evidence items for an instance. For region-based metrics, we construct pairwise IoU scores between predicted and reference regions and perform maximum-weight one-to-one matching using the Hungarian algorithm. Unmatched predicted or reference evidence remains in the corresponding denominator, penalizing both over-prediction and missing evidence.

Evidence-IoU. Predicted and reference regions are matchable only when they occur on the same page. For instances whose predicted page set exactly matches the reference page set, Evidence-IoU is computed as

$$\text{Evidence}-\text{IoU} = \frac{\sum_{(i,j)\in\mathcal{M}} \text{IoU}(b_i, \hat{b}_j)}{\max(P, G)}, \tag{17}$$

where $\mathcal{M}$ is the maximum-weight one-to-one matching. The final Evidence-IoU is averaged over test instances satisfying the page-set exact-match condition.

Region-F1. Using the same page-consistent IoU matching, Region-F1 is

$$\text{Region}-\text{F1} = \frac{2\sum_{(i,j)\in\mathcal{M}} \text{IoU}(b_i, \hat{b}_j)}{P + G}. \tag{18}$$

Region-F1 is likewise averaged over instances whose predicted and reference page sets are identical.

Relation-F1. Relation-F1 additionally requires the predicted and reference evidence items to have the same relation label. Only same-page, relation-consistent pairs retain their IoU score before Hungarian matching:

$$\text{Relation}-\text{F1} = \frac{2\sum_{(i,j)\in\mathcal{M}_{\text{rel}}} \text{IoU}(b_i, \hat{b}_j)}{P + G}. \tag{19}$$

Like Evidence-IoU and Region-F1, Relation-F1 is averaged over page-exact test instances. This metric evaluates relation-consistent evidence matching rather than relation classification in isolation.

Typed-F1. Typed-F1 applies the strongest matching condition. A matched pair must agree on page, section, and evidence–answer relation, in addition to its continuous visual overlap:

$$\text{Typed}-\text{F1} = \frac{2\sum_{(i,j)\in\mathcal{M}_{\text{typed}}} \text{IoU}(b_i, \hat{b}_j)}{P + G}. \tag{20}$$

When the reference section is `null`, the section condition is treated as satisfied; when the reference section is non-`null`, the normalized predicted and reference section names must match. Unlike Evidence-IoU, Region-F1, and Relation-F1, Typed-F1 is evaluated over all 409 test instances rather than conditioning on exact page-set prediction.

Invalid and incomplete predictions. Predicted page indices outside the valid document range are discarded. A retained evidence item with an invalid or missing bounding box receives zero region overlap while remaining in the prediction count, and an invalid or missing relation label cannot satisfy relation-aware matching. Partially malformed structured outputs are evaluated using recoverable fields only; gold annotations are never used to repair model predictions.

C.3 Overall Score

Main evaluation. For the main results, we first average metrics within the three evaluation dimensions:

$$\begin{aligned} A &= \frac{\mathrm{ROUGE{-}L} + \mathrm{BERTScore{-}F1}}{2}, \\ L &= \frac{\mathrm{Page{-}F1} + \mathrm{Section{-}F1} + \mathrm{Evidence{-}IoU} + \mathrm{Region{-}F1}}{4}, \end{aligned} \tag{21}$$

and

$$T = \frac{\mathrm{Relation{-}F1} + \mathrm{Typed{-}F1}}{2}, \qquad \mathrm{Overall} = \frac{A + L + T}{3}. \tag{22}$$

This formulation gives equal weight to answer quality, evidence localization, and typed provenance regardless of the number of individual metrics in each category.

Existing-method comparison. VISA, CoE, and UniDoc-RL do not provide the section and typed-relation outputs required by the full ReViE evaluation. We therefore evaluate their shared answer and localization metrics using the same definitions as in the main evaluation, but compute a comparison-specific Overall score over only these shared dimensions:

$$A_5 = \frac{\mathrm{ROUGE{-}L} + \mathrm{BERTScore{-}F1}}{2}, \qquad L_5 = \frac{\mathrm{Page{-}F1} + \mathrm{Evidence{-}IoU} + \mathrm{Region{-}F1}}{3}, \tag{23}$$

$$\mathrm{Overall}_{\mathrm{shared}} = \frac{A_5 + L_5}{2}. \tag{24}$$

This comparison-specific score is used only in Table 3 and is not directly comparable to the three-category Overall score used in Table 1.

C.4 LLM-as-Judge Protocol

We use GPT-5.6 Luna as a blind, text-only judge on a shared 50-instance subset of the 409-instance ReViE test set. All evaluated systems are scored on the same set of questions. For each instance, the judge receives only the *Question*, *Reference Answer*, and *Predicted Answer*; document pages, evidence regions, section annotations, relation labels, provenance predictions, and model identities are not provided.

The judge evaluates the predicted answer jointly in terms of correctness, completeness, and faithfulness to the reference answer. It assigns a single integer score from 0 to 5 according to the following rubric:

- **5**: fully correct, complete, and faithful, with no material error or omission;
- **4**: the core answer is correct, with only a minor omission, imprecision, or wording issue;
- **3**: substantial correct content is present, but with one major error or omission, or several minor issues;
- **2**: some relevant content is correct, but key information is missing or materially incorrect;
- **1**: only weakly relevant content is present, with little usable answer to the question;
- **0**: the response is completely incorrect, contradictory, off-topic, empty, a refusal, or otherwise unusable.

Faithful paraphrases, equivalent scientific notation or units, and harmless rounding are accepted, and lexical overlap is not required. Each system's Judge score is the arithmetic mean over the shared 50 evaluated instances. The judge evaluates answer content only and does not inspect visual evidence or provenance. Its score is therefore treated as an auxiliary answer-quality signal rather than a provenance metric and is excluded from Overall.

Judge prompt. The complete system prompt and user template used for LLM-based answer evaluation are shown in Figure 8.

LLM-as-Judge Prompt

You are the fixed, blind, text-only answer evaluator for ReViE, a scientific-document question-answering benchmark.

INPUT BOUNDARY

You receive exactly three fields: Question, Reference Answer, and Predicted Answer. The Reference Answer is the authoritative evaluation target. Use no external facts. Never evaluate or infer pages, images, bbox coordinates, sections, evidence, provenance, model identity, or writing style beyond whether the answer is clear enough to assess.

SCORING PRINCIPLES

Evaluate the Predicted Answer jointly for:

1. **Correctness:** its substantive claims agree with the Reference Answer.
2. **Completeness:** it covers every part needed to answer the Question.
3. **Faithfulness:** it does not add material claims contradicted by, or unsupported by, the Reference Answer.

Accept faithful paraphrases, synonyms, reordered statements, equivalent scientific notation, equivalent units, and harmless rounding that preserves the meaning. Do not score by lexical overlap. Do not require an explanation when a short answer fully answers the Question. For multi-part questions, assess every requested part. Treat a material reversal, wrong entity/value, or contradiction as a major error. An empty answer, refusal, or unusable output scores 0.

INTEGER RUBRIC

- **5**: Fully correct, complete, and faithful; no material error or omission.
- **4**: Core answer is correct; only a minor omission, imprecision, or wording issue.
- **3**: Substantial correct content, but one major omission/error or several minor ones.
- **2**: Some relevant correct content, but key information is missing or materially wrong.
- **1**: Only weakly relevant content; almost no usable answer to the Question.
- **0**: Completely incorrect, contradictory, off-topic, empty, refusal, or unusable.

OUTPUT CONTRACT

Return JSON only, with exactly:

```
{"score": <integer 0..5>, "reasoning": "<brief auditable justification>"}
```

The justification must identify the decisive agreement, omission, or error in one or two concise sentences. Do not output markdown or additional keys.

USER MESSAGE TEMPLATE

Evaluate this answer.

Question:

```
{question}
```

Reference Answer:

```
{reference_answer}
```

Predicted Answer:

```
{predicted_answer}
```

Figure 8: Prompt used for blind text-only LLM-as-Judge evaluation. The judge receives only the question, reference answer, and predicted answer and assigns an integer score from 0 to 5.

C.5 Human Evaluation Protocol

We conduct human evaluation on a shared subset of 20 test questions across 11 models, yielding 220 model–question answer observations. The evaluated systems are Qwen3-VL-8B Base, InternVL3.5-8B, MiniCPM-V4.5-8B, Qwen3-VL-32B, InternVL3.5-38B, Qwen2.5-VL-7B, Eagle2.5-8B, GPT-4o-mini, Claude Haiku 4.5, Gemini 2.5 Flash-Lite, and VisProve. Ablation variants are not included in this analysis.

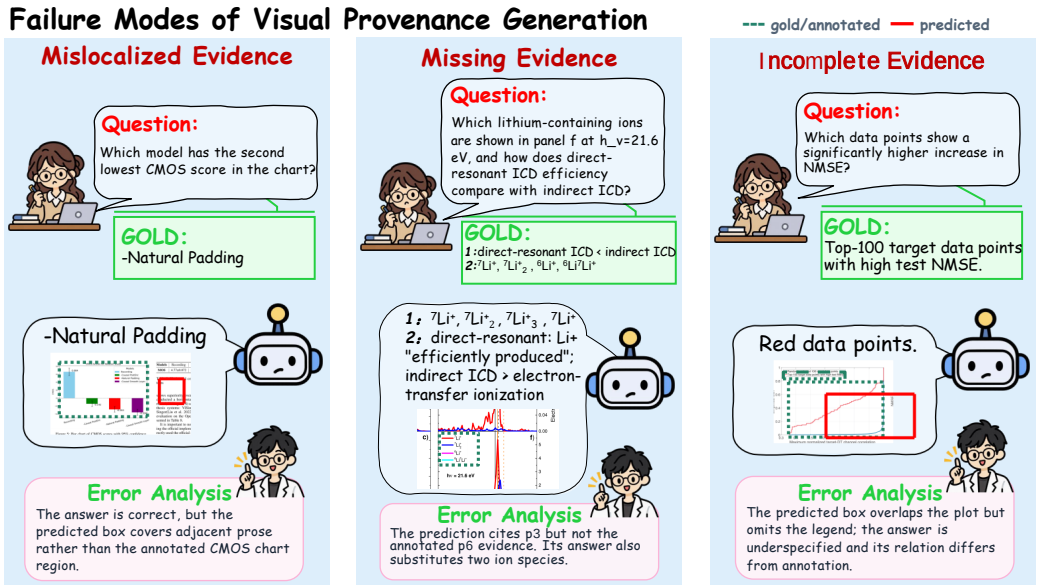

Figure 9: Representative prediction errors. Left: a correct chart-reading answer paired with a misplaced evidence box. Middle: incomplete cross-page evidence coverage accompanied by an incorrect identification. Right: incomplete evidence coverage, an underspecified answer, and a relation discrepancy.

Four expert reviewers participate in the evaluation. Reviewers assess answers independently using only the *Question*, *Reference Answer*, and *Predicted Answer*. Document pages, separately provided evidence or provenance annotations, model identities, and automated evaluation scores are not shown. Although provenance-like content may appear in a model's raw output, no separate evidence fields are explicitly provided to the reviewers.

Each reviewer assigns an integer score from 0 to 5 based jointly on correctness, completeness, and faithfulness to the reference answer. Here, faithfulness is assessed against the reference answer; reviewers do not inspect the source document. The scoring rubric is:

- **5**: fully correct and complete, with no material error or omission;
- **4**: the core answer is correct, with only minor issues;
- **3**: substantial correct content is present, but with a major omission or error;
- **2**: some relevant content is correct, but key information is missing or materially incorrect;
- **1**: the response is only weakly relevant and contains little usable answer content;
- **0**: the response is incorrect, off-topic, empty, or otherwise unusable.

All 220 answers are independently scored by three reviewers. A fourth reviewer independently re-evaluates 16 answers exhibiting larger disagreement among the initial three reviewers. No discussion-based consensus score is used. For the 16 re-evaluated answers, the final human score is the median of four ratings; for all remaining answers, it is the median of three ratings. Consequently, the final per-answer human score may take half-point values in the re-evaluated cases.

For consistency analysis, we compute each model's human score as the arithmetic mean over the same 20 questions. The LLM-as-Judge score is computed over these same model–question pairs, and each point in Figure 5 represents one model's mean human score and mean LLM-as-Judge score. Pearson's correlation across the 11 model-level pairs is $r = 0.9898$. The correlation is computed over 11 model-level pairs; instance-level agreement is a separate quantity.

D QUALITATIVE ERROR ANALYSIS

Figure 9 presents three representative failure patterns observed in VisProve predictions.

Region localization mismatch. The left case answers the CMOS ranking question correctly, but places the evidence box over adjacent prose. The reported region differs from the chart region supporting the answer.

Cross-page evidence omission. The middle case requires introductory text and a later figure. The prediction cites only the introductory page and misidentifies two ion species. Its output exhibits both incomplete evidence coverage and an answer error.

Evidence coverage and relation discrepancy. In the right case, a predicted box overlaps an annotated plot region but omits a separate legend region. The answer gives a color-based description with insufficient specificity, and its predicted relation differs from the annotation. This case combines evidence-coverage, answer-completeness, and relation discrepancies.

These examples distinguish several forms of disagreement with annotated support. Their interpretation is relative to the reference evidence sets, which may coexist with other valid support.