RAG-SCHOLAR: CONTEXT-EFFICIENT RETRIEVAL OVER LARGE SCHOLARLY CORPORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large scholarly corpora contain many passages that match a query while repeating the same concepts. Under a limited context budget, these matches can displace the source-specific details needed for an answer. We present RAG-SCHOLAR, a hierarchical retrieval framework that separates source identification from answer evidence selection. Its graph-guided scholarly index combines explicit lexical, entity, and section-role features with a relation-aware graph neural network. Explicit features preserve query-specific source distinctions, while graph representations incorporate related descriptions from the scholarly neighborhood. Text-anchored encoding and constraint-preserving rank fusion integrate these signals across papers, sections, and local evidence. Scholarly-context-conditioned visual retrieval further discounts image regions that overlap with retained text. On a dataset constructed from 146 papers, RAG-SCHOLAR improves Fact F1 by 16.11 points over the strongest evaluated external text baseline with a 44.6-fold reduction (97.76%) in the combined preprocessing-and-answer token count. On multimodal questions, it improves Fact F1 by 3.08 points with 71.97% fewer answer-model input tokens. Across collections of 146–597 papers, hierarchical retrieval consistently outperforms flat retrieval in paper identification and factual answering. These results support selecting compact, source-specific evidence rather than accumulating independently relevant matches. The project page and source code are available at `https://rag-scholar.pages.dev`.

1 INTRODUCTION

Large-scale scholarly retrieval must search far more information than an answer model can read. Retrieval-augmented generation provides access to external documents (Lewis et al., 2020), but a larger searchable corpus does not automatically yield a more informative retrieved context (Zhao et al., 2024; Lee et al., 2025). Research papers share terminology, background explanations, and descriptions of established methods. A query can therefore retrieve more relevant material without obtaining more of the information needed for its answer. Under a fixed context budget, repeated matches compete directly with evidence that contributes a different detail (Lee et al., 2025; In et al., 2025). We study the resulting question: *How can retrieval select complementary evidence from a growing scholarly corpus within a fixed context budget?*

Consider a query comparing the retrieval mechanisms of papers that use the same underlying method. Several passages may accurately describe that method, yet the comparison depends on paper-specific differences in candidate construction, reranking, or the interactions depicted in pipeline figures. Repeating the common method description uses space needed for these differences. Increasing the number of matches can exacerbate this imbalance, while truncating the result can discard the complementary evidence (Zhao et al., 2024; Lee et al., 2025). Here, *complementary evidence* contributes a query-relevant detail not already represented in the retrieved context, including the source-specific information needed for a comparison.

Existing methods approach the context constraint through hierarchical and graph-based retrieval (Sarthi et al., 2024; Gutiérrez et al., 2025; Zuo et al., 2026), context compression (Jiang et al., 2024; Xu et al., 2024), and visual document or region retrieval (Faysse et al., 2025; Li et al., 2026). When candidates are ranked independently, several descriptions of the same concept can all score highly and occupy the space needed for another part of the answer. Compression can remove repetition

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.

from supplied candidates, but cannot recover complementary evidence excluded during retrieval. A query-relevant visual region can likewise repeat information already retained as text. The budget problem is thus not matching alone: *relevant repetition can crowd out complementary evidence*.

We propose RAG-SCHOLAR, a *hierarchical retrieval framework* that separates information used to identify relevant sources from evidence retained for answering. Paper and section relevance determine which local evidence competes for the context, with separate limits on each source's contribution. A paper does not obtain an unrestricted share of the candidate pool simply because it contains more matching passages. Shared concepts can instead guide source identification, leaving the answer context for the paper-specific details required by the query.

A *graph-guided scholarly index* supports this separation by combining explicit scholarly features with learned graph representations. Explicit features retain named source conditions and section roles; a relation-aware GNN connects descriptions expressed across neighboring document and entity nodes. These signals address different aspects of source identification: relational similarity captures shared ideas, but does not itself enforce which institution, dataset, or section role the query requires. We therefore integrate both rankings while preserving recognized source constraints. *Scholarly-context-conditioned visual retrieval* extends this principle to images, evaluating candidate regions against the retained text. The three components jointly determine which sources contribute evidence and which source-specific details deserve the limited context.

In end-to-end evaluation, RAG-SCHOLARimproves Fact F1 by **16.11 points** over CompAct, the strongest evaluated external text baseline on this metric, while reducing the combined preprocessing-and-answer token count from 11,181.32 to 250.97: a **44.6-fold reduction in the reported token count** (97.76%). On multimodal questions, it improves Fact F1 by **3.08 points** over M3DocRAG while using **71.97% fewer answer-model input tokens**. Our contributions are threefold:

- **Hierarchical retrieval framework.** We propose a hierarchical retrieval framework to address the competition between repeated matches and complementary evidence under a fixed context budget. By conditioning local evidence selection on its parent paper/section, the framework bounds each source's contribution and supports accurate scholarly answering with compact retrieved contexts.
- **Graph-guided scholarly index.** We introduce a graph-guided scholarly index that combines explicit scholarly features with text-anchored, relation-aware GNN representations to reconcile semantic matching with query-specific source requirements. Their rank fusion incorporates neighborhood information while preserving recognized source constraints, allowing shared concepts to guide retrieval without losing the paper-specific distinctions needed for answering.
- **Scholarly-context-conditioned visual retrieval.** We develop scholarly-context-conditioned visual retrieval to address the repetition of retained textual information in visual inputs. Conditioning region selection on the query and retained text, together with spatial constraints, yields compact visual evidence. In the crop ablation, it reduces visual tokens by 58.42% with a 0.22-point difference in aggregate Fact F1.

2 RELATED WORK

Hierarchical and graph-based retrieval. Document-grounded scholarly QA addresses questions requiring evidence from multiple parts of a paper (Lewis et al., 2020), as in QASPER (Dasigi et al., 2021). Summary hierarchies and entity–passage graphs support retrieval beyond independent text chunks (Sarthi et al., 2024; Gutiérrez et al., 2025; Gutiérrez et al., 2024), while multimodal document graphs connect heterogeneous evidence across pages through query-linked chunks or budget-controlled page and element selection (Zhang & Wu, 2026; Zuo et al., 2026). Yet answering within a relevant paper or searching a long document does not identify which papers merit detailed search in a large collection: passage or page relevance alone cannot establish that a paper jointly satisfies institution, method, and dataset constraints.

Context compression. Prompt compressors shorten text through token selection or learned extractive compression (Jiang et al., 2023; 2024; Pan et al., 2024). Retrieval-specific compressors select or summarize information before answer generation (Xu et al., 2024; Yoon et al., 2024). Without resolving paper selection, expanding related evidence can spend the limited context budget on locally relevant but partially qualifying papers, displacing needed evidence. Compression cannot recover

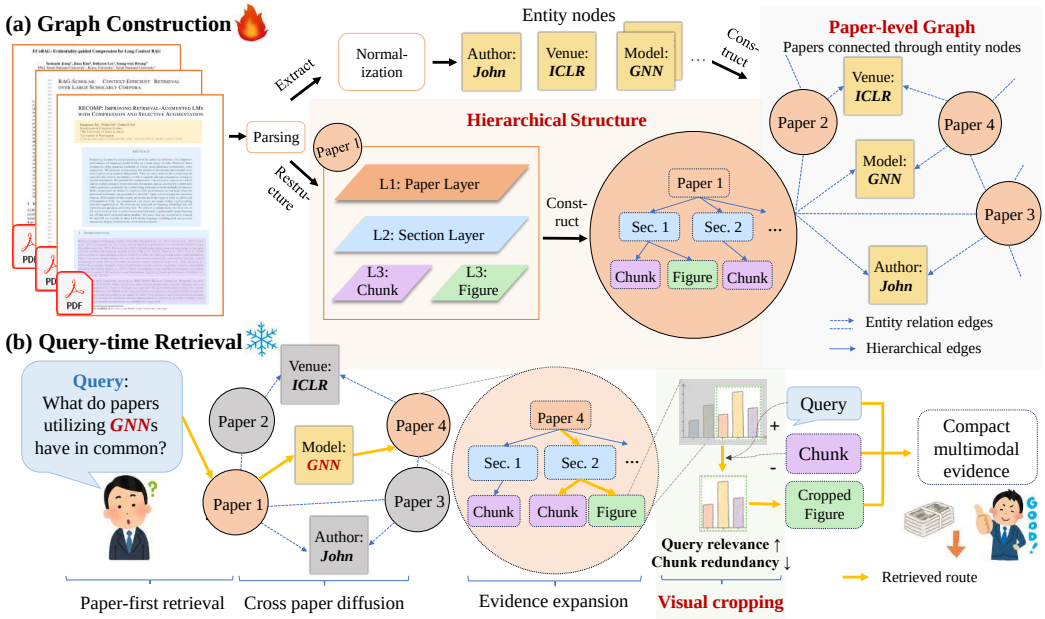


Figure 1: **Selecting complementary evidence under a fixed context budget.** (a) Graph construction organizes papers into document hierarchies linked through normalized scholarly entities, followed by GNN training to learn retrieval representations. (b) Query-time retrieval combines lexical and rule-based retrieval with GNN-based retrieval through weighted rank fusion. Hierarchical evidence selection and text-conditioned visual cropping produce compact multimodal context.

evidence absent from its retrieved input or avoid retrieval costs already incurred. Shorter answer-model contexts alone therefore imply neither better paper retrieval nor lower retrieval latency as the corpus grows, leaving selection of the right content before compression unresolved.

Visual document and region retrieval. Multi-document multimodal retrieval accesses evidence across pages and documents (Cho et al., 2025). Visual document retrieval and multimodal document understanding preserve information that text extraction may omit (Faysse et al., 2025; Yu et al., 2025; Han et al., 2025), while region-level retrieval and query-driven parsing reduce irrelevant visual content (Li et al., 2026; Wang et al., 2026). Yet spatial localization alone does not resolve redundancy with retrieved text: a query-relevant figure region may repeat a retained method description or omit a complementary detail needed for the answer.

3 METHOD

3.1 OVERVIEW

Let $\mathcal{D} = \{d_i\}_{i=1}^{N}$ be a collection of N scholarly papers with metadata, sections, and figures. For query q, retrieval supplies text $\mathcal{X}_q$, visual evidence $\mathcal{I}_q$, and source identifiers $\mathcal{Z}_q$ to the answer model g, which produces $\widehat{y}_q = g(q, \mathcal{X}_q, \mathcal{I}_q, \mathcal{Z}_q)$. Let τ_t and τ_v count serialized text and processed visual tokens, with respective budgets B_t and B_v. We study retrieval under

$$\tau_t(\mathcal{X}_q, \mathcal{Z}_q) \le B_t, \qquad \tau_v(\mathcal{I}_q) \le B_v. \tag{1}$$

As N grows, more passages can match the query without contributing additional answer details. The goal is therefore to retain complementary, source-specific evidence rather than enlarge the answer context; text-only queries have $\mathcal{I}_q = \varnothing$.

RAG-SCHOLAR separates *source identification* from *answer evidence selection* (Figure 1). Hierarchical retrieval limits each paper's and section's contribution to the local candidate pool. The graph-guided scholarly index combines explicit source and role features with learned neighborhood representations to rank these candidates. Scholarly-context-conditioned visual retrieval then selects

image regions relative to the retained text. Together, these components use broad corpus information to guide selection while reserving the answer context for the evidence actually needed.

3.2 Hierarchical Retrieval Framework

A relevant paper and a useful passage are different retrieval units. A passage may describe the requested method but belong to the wrong source; conversely, only a few passages in a relevant paper may support the answer. We therefore rank local evidence within its parent paper and section.

Within a search scope Ω, let L_{E} and L_{G} be the explicit-feature and graph candidate lists, and let $r_{\mathrm{E}}(v \mid \Omega, q)$ and $r_{\mathrm{G}}(v \mid \Omega, q)$ be candidate v's ordinal ranks. The scope is the collection, a paper, or a section, depending on the retrieval level. With graph weight $\alpha = 0.3$ and smoothing constant $k = 60$, the fused score is

$$S_{\mathrm{H}}(v \mid \Omega, q) = (1-\alpha)\frac{k+1}{k + r_{\mathrm{E}}(v \mid \Omega, q)} + \alpha\frac{k+1}{k + r_{\mathrm{G}}(v \mid \Omega, q)}. \tag{2}$$

An absent candidate contributes zero in that branch. Rank fusion (Cormack et al., 2009) combines scores with different scales without equating lexical-rule values with embedding similarities.

Let $\mathcal{U}_p(q)$ and $\mathcal{U}_s(p, q)$ be the unions of returned papers and sections within paper p. Let $\mathcal{A}(q)$ be the papers returned by an activated explicit-constraint mode, or $\mathcal{U}_p(q)$ otherwise. For paper and per-paper section limits K_p, K_s, the selected sets $\mathcal{P}(q), \mathcal{S}(p, q)$ are

$$\begin{aligned} \mathcal{P}(q) &= \mathrm{TopK}^{K_p}_{p \in \mathcal{U}_p(q) \cap \mathcal{A}(q)} S_{\mathrm{H}}(p \mid \mathcal{D}, q), \\ \mathcal{S}(p, q) &= \mathrm{TopK}^{K_s}_{s \in \mathcal{U}_s(p,q)} S_{\mathrm{H}}(s \mid p, q), \end{aligned} \tag{3}$$

where TopK^K retains up to K highest-ranked items. Analogous fusion selects chunk and figure sets $\mathcal{C}(s, q), \mathcal{F}(s, q)$, with per-section limits K_c, K_f. Their union forms the local evidence pool $\mathcal{E}(q)$:

$$\mathcal{E}(q) = \bigcup_{p \in \mathcal{P}(q)} \bigcup_{s \in \mathcal{S}(p,q)} \big(\mathcal{C}(s, q) \cup \mathcal{F}(s, q)\big). \tag{4}$$

For ordinary top-K queries, each paper contributes at most $K_s(K_c + K_f)$ items, and $|\mathcal{E}(q)| \leq K_p K_s(K_c + K_f)$. These limits control source concentration rather than total corpus-scoring cost. The context controller retains prioritized fields with their source identifiers. Prefetch, exhaustive-list handling, and context-budget implementation are specified in Appendices B.

3.3 Graph-Guided Scholarly Index

Shared concepts identify relationships between papers without making their evidence interchangeable. Our typed graph $G = (V, E)$ retains both concepts and provenance: V contains document and canonical entity nodes, and E contains structural and scholarly relations. Writing p_i for paper i, s_{ij} for its jth section, and $c_{ijk}, f_{ij\ell}$ for that section's chunks and figures, containment preserves

$$p_i \longrightarrow \{s_{ij}\}_j, \qquad s_{ij} \longrightarrow \{c_{ijk}\}_k \cup \{f_{ij\ell}\}_\ell. \tag{5}$$

Typed entity and citation links connect these sources through authors, institutions, methods, datasets, and other scholarly attributes. Canonicalization and confidence aggregation are detailed in Appendix B.

Relation-aware graph encoding. We use a two-layer GNN to incorporate document–entity relationships into node representations. A frozen text encoder supplies unit vectors $\mathbf{x}_v, \mathbf{e}_q \in \mathbb{R}^{768}$ for node v and query q. Let $\tau(v)$ be the node type, $\mathbf{t}_{\tau(v)} \in \mathbb{R}^{128}$ its learned embedding, and $W_{\mathrm{in}}, \mathbf{b}_{\mathrm{in}}$ a learned projection and bias. The initial hidden state is

$$\mathbf{h}_v^{(0)} = W_{\mathrm{in}}\mathbf{x}_v + \mathbf{b}_{\mathrm{in}} + \mathbf{t}_{\tau(v)}. \tag{6}$$

Let E_{msg} contain the original and reverse message edges. For $(u, r, v) \in E_{\mathrm{msg}}$, u is the source, v the target, r the relation, and $\bar{c}_{urv}$ its confidence normalized within relations and across active relation groups. In layer $\ell \in \{0, 1\}$, $W_n^{(\ell)}, W_s^{(\ell)}$ are neighbor and self projections, $\mathbf{b}_s^{(\ell)}$ is a bias, and $\mathbf{g}_r^{(\ell)}$ is

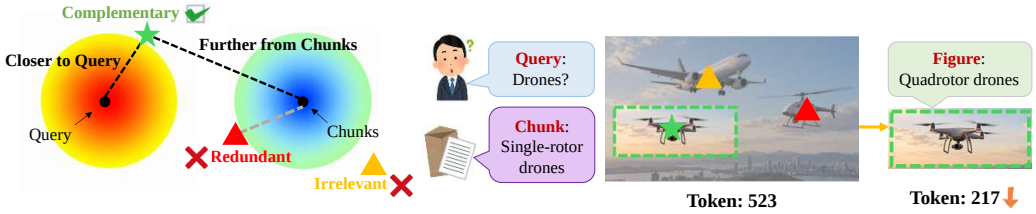


Figure 2: Illustration of redundancy-aware visual cropping. The semantic map combines query relevance with similarity to retained text, favoring a query-relevant target region while discounting regions that are irrelevant to the query or redundant with retrieved chunks.

a learned relation gate. With aggregate $\mathbf{m}_v^{(\ell)}$, elementwise product $\odot$, layer normalization LN, and GELU activation, propagation is

$$\mathbf{m}_v^{(\ell)} = \sum_{(u,r,v)\in E_{\text{msg}}} \bar{c}_{urv}\big(W_n^{(\ell)}\mathbf{h}_u^{(\ell)} \odot \mathbf{g}_r^{(\ell)}\big),$$
$$\mathbf{h}_v^{(\ell+1)} = \text{LN}\Big(\mathbf{h}_v^{(\ell)} + \text{GELU}\Big(W_s^{(\ell)}\mathbf{h}_v^{(\ell)} + \mathbf{b}_s^{(\ell)} + \mathbf{m}_v^{(\ell)}\Big)\Big). \tag{7}$$

Relation gates distinguish the messages carried by different scholarly links, while normalization balances their contributions. With learned output projection $W_o, \mathbf{b}_o$, residual scale $\eta = 0.25$, and norm denoting ℓ_2 normalization, the graph increment $\boldsymbol{\delta}_v$ and exported representation $\mathbf{z}_v$ are

$$\boldsymbol{\delta}_v = \text{norm}(W_o\mathbf{h}_v^{(2)} + \mathbf{b}_o), \qquad \mathbf{z}_v = \text{norm}(\mathbf{x}_v + \eta\boldsymbol{\delta}_v). \tag{8}$$

The residual anchors graph representations to the frozen query-embedding space. Masked reconstruction, visible-node anchoring, and contrastive alignment train the GNN without query labels. Normalization and training details are in Appendix C; figure features encode captions and OCR, not pixels.

Complementing explicit scholarly features. Graph representations capture neighborhood semantics, whereas explicit features retain lexical matches, named-source conditions, and section roles. Papers can be semantically close while differing in the dataset or institution required by a query; corpus-level reconstruction does not directly supervise these distinctions. We therefore combine the two rankings through equation 2, rather than replacing explicit features with graph similarity. The admissible set in equation 3 preserves recognized source constraints, allowing relational information to guide ranking within those constraints.

Source-balanced semantic support. For an active node-type channel t and paper p, let $\mathcal{V}_t(p)$ contain nodes of type t belonging or positively linked to p. The channel score $a_t(p,q)$ retains its strongest query-aligned support:

$$a_t(p,q) = \max_{v\in\mathcal{V}_t(p)} \mathbf{e}_q^\top \mathbf{z}_v. \tag{9}$$

An empty channel contributes no support. Channel rankings are combined by reciprocal rank fusion; local sections, chunks, and figures use their node cosines within the parent scope. Taking one maximum per type–paper pair prevents repeated supports from accumulating solely through their multiplicity. The resulting graph lists enter equation 2 alongside the explicit lists. Node representations are computed offline; queries score the stored vectors without GNN propagation. Channel weighting and online cost are given in Appendices C.4 and E.

3.4 Scholarly-Context-Conditioned Visual Retrieval

Figure retrieval identifies relevant sources; pixel selection determines what visual information should accompany the text. Let $\mathcal{S}(q) = \bigcup_{p\in\mathcal{P}(q)} \mathcal{S}(p,q)$ be the retained sections and $\mathcal{H}_q = (\mathcal{P}(q), \mathcal{S}(q), \mathcal{X}_q)$ the scholarly conditioning context. For figure f in section s_f, its candidate score S_f is the parent-conditioned fused score,

$$S_f(f \mid q, s_f) = S_{\text{H}}(f \mid s_f, q). \tag{10}$$

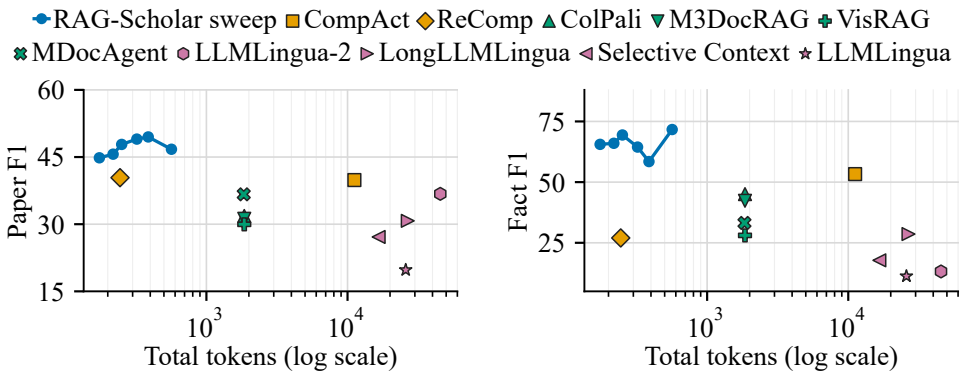


Figure 3: Text-retrieval quality–cost trade-off. Each point plots the per-query Total token cost from Table 1 against Paper F1 (left) or Fact F1 (right); the x-axis is logarithmic. RAG-SCHOLAR's retention sweep is connected in blue.

The explicit branch uses figure-text and query-specific source signals, while the graph branch uses the figure node's relational representation. Neither branch replaces pixel-level selection.

Let I_f be the figure image, $\mathcal{W}_f$ its candidate windows, and $\mathcal{T}_f(q)$ retained text snippets from its section. For $w \in \mathcal{W}_f$, $I_f[w]$ is the corresponding crop, and $|w|/|I_f|$ is its relative area. Let E_i, E_t be normalized CLIP image and text encoders (Radford et al., 2021). We define query alignment u, text overlap r, and region score S_w using nonnegative overlap and area penalties α_v, β_v:

$$u(w,q) = \langle E_i(I_f[w]), E_t(q) \rangle,$$
$$r(w,q) = \max_{t \in \mathcal{T}_f(q)} \langle E_i(I_f[w]), E_t(t) \rangle,$$

$$S_w(w \mid q, f, \mathcal{H}_q) = u(w,q) - \alpha_v r(w,q) - \beta_v \frac{|w|}{|I_f|}. \tag{11}$$

Thus, equally query-relevant regions can receive different scores when one repeats retained text. The area penalty favors a compact spatial extent rather than treating all relevant pixels as equally valuable, as illustrated in Figure 2.

Greedy selection retains high-scoring windows subject to overlap and total-area constraints. The selected visual input is processed by the answer model's image processor; its resulting token count, not area alone, determines visual cost. Appendix B.3 specifies these constraints and the remaining rendering settings. The final context combines selected text, visual evidence, and their paper and section identifiers.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and metrics. We construct a dataset from 146 scholarly papers, with 116 text queries and 30 multimodal queries. Paper-set F1 measures paper identification in the final answer; Fact F1 measures factual answering on 24 text and 15 visual-reasoning queries. Quality scores are macro-averaged percentages. Dataset construction and scoring details are in Appendices A and D.

Comparisons and cost. We compare compact-evidence, page-level visual-retrieval, and long-context compression methods. Pre. counts online preprocessing tokens; Input counts retained context delivered to the answer model; Total sums the two. Prep. measures online preparation time, excluding generation. The retention setting b is distinct from measured context length. Configurations and counting conventions are in Appendix E.

4.2 ANSWER QUALITY UNDER CONTEXT CONSTRAINTS

Table 1 shows that $b = 320$ attains 69.42 Fact F1 with 250.97 answer-model tokens. Against ReComp's 244.28 tokens, it gains 42.42 Fact F1 and 9.97 Paper F1 points. CompAct retains fewer

Table 1: Text-suite results: Paper F1 uses 116 queries; Fact F1 uses 24 factual queries. Costs are per-query means. Pre. counts the online tokens incurred when a method invokes an LLM to compress its retrieved context; it is zero when no such LLM-based compression is used. Total = Pre. + Input. Dashes indicate unreported times. The full sweep, including $b = 140, 180, 800$, and precision/recall results appear in Appendix F.

	Answer quality ↑		Token cost ↓			Online cost
Method / b	Paper F1	Fact F1	Pre.	Input	Total	Prep. (s)
RAG-SCHOLAR: retention settings						
$b = 100$	45.76	65.56	0	174.42	**174.42**	11.39
$b = 240$	48.79	66.00	0	218.20	218.20	11.38
$b = 320$	50.36	69.42	0	250.97	250.97	11.37
$b = 480$	**50.91**	64.44	0	321.36	321.36	**11.35**
$b = 640$	50.54	58.47	0	387.05	387.05	11.36
Uncapped	49.37	**71.67**	0	565.21	565.21	11.80
Compact evidence baselines						
CompAct *(EMNLP'24)*	39.86	53.31	11026.36	**154.96**	11181.32	46.77
ReComp *(ICLR'24)*	40.39	27.00	0	244.28	244.28	25.06
Page-level VLM retrieval						
ColPali *(ICLR'25)*	31.84	45.07	0	1854.68	1854.68	–
M3DocRAG *(ICCVW'25)*	31.21	42.30	0	1854.42	1854.42	–
VisRAG *(ICLR'25)*	29.97	28.05	0	1854.46	1854.46	–
MDocAgent *(arXiv'25)*	36.67	33.19	0	1838.84	1838.84	–
Long-context compression baselines						
LLMLingua-2 *(ACL Findings'24)*	36.78	13.19	42729.27	2687.00	45416.27	205.91
LongLLMLingua *(ACL'24)*	30.75	28.60	24421.79	2304.91	26726.70	448.14
Selective Context *(EMNLP'23)*	27.15	17.78	14106.06	2500.00	16606.06	–
LLMLingua *(EMNLP'23)*	19.78	11.25	23596.27	2234.35	25830.62	418.99

tokens, but RAG-SCHOLAR improves Fact F1 by 16.11 points while reducing preparation time from 46.77 to 11.37 seconds and using no recorded preprocessing tokens. Figure 3 visualizes these comparisons on a logarithmic total-token axis. Its sweep remains in the low-token region, while the higher-token baselines do not surpass its best Paper or Fact F1 settings. The sweep separates paper identification from factual coverage: $b = 480$ gives the highest Paper F1 (50.91), whereas uncapped retrieval gives the highest Fact F1 (71.67). The compact $b = 320$ setting uses 55.60% fewer answer-model tokens than uncapped retrieval, with a 2.25-point Fact F1 difference. Across the sweep, longer contexts do not consistently improve factual answering. Together with the similar-length ReComp comparison, this indicates that context length alone is a poor predictor of answer quality. The benefit of RAG-SCHOLAR is therefore a favorable quality–cost trade-off, rather than input reduction considered in isolation.

On the multimodal suite (Table 2), RAG-SCHOLAR reaches 100.00 Paper F1 and 60.00 Fact F1. Relative to M3DocRAG, it improves these scores by 37.56 and 3.08 points, respectively, while using 71.97% fewer answer-model tokens. The reduction spans both text and visual input. The smaller text input is accompanied by fewer visual tokens, rather than a compensating increase in the other modality. Gains in Paper F1 and Fact F1 also occur together, so the compact context supports both source identification and factual answering. Across the two suites, the favorable trade-off is not confined to either textual or visual input alone.

4.3 HIERARCHICAL RETRIEVAL UNDER CORPUS GROWTH

Figure 4 expands the collection from 146 to 597 papers while retaining the same queries and $b = 240$ setting. Hierarchical retrieval improves both quality metrics at every measured size and reduces preparation time by 2.58–3.14×. At 146 papers, it yields a 7.2-point Fact F1 advantage for the hierarchy. At 597 papers, the Paper F1 and Fact F1 margins are 12.4 and 33.0 points, respectively, while preparation takes 39.38 rather than 123.62 seconds. The hierarchy thus preserves stronger factual answering as the candidate collection expands, at lower online preparation cost. The preparation-

Table 2: Multimodal results: Paper F1 uses 30 queries; Fact F1 uses 15 reasoning queries. Total = Text + Visual. Precision/recall and scoring details are in Appendices G and D.

	Answer quality ↑		Answer-model input (tokens) ↓		
Method	Paper F1	Fact F1	Text	Visual	Total
RAG-SCHOLAR	**100.00**	**60.00**	**1090.70**	**217.80**	**1308.50**
ColPali *(ICLR'25)*	56.33	46.29	3265.37	1384.60	4649.97
M3DocRAG *(ICCVW'25)*	62.44	56.92	3271.37	1397.00	4668.37
VisRAG *(ICLR'25)*	44.67	25.33	3200.71	1381.06	4581.77

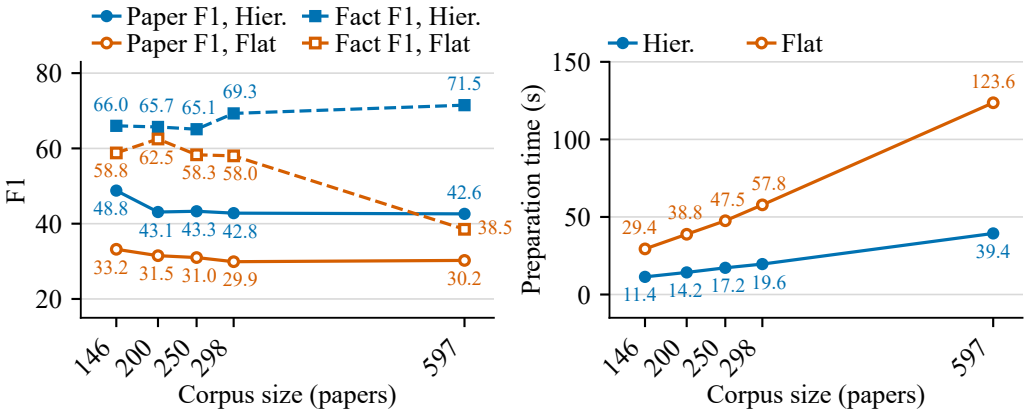


Figure 4: Corpus expansion at $b = 240$ with fixed queries. Left: Paper F1 and Fact F1; right: online preparation time. Hier./Flat denote hierarchical/flat retrieval, and each marker is labeled with its measured value.

time gap widens from 18.02 to 84.24 seconds between the smallest and largest collections. Over the same range, corpus size more than quadruples. These measurements indicate that broader search need not require a proportional increase in the evidence passed to the answer model.

Flat retrieval does not close the quality gap as more candidate papers become available. The hierarchy instead uses paper and section relevance to determine which evidence occupies the limited context.

4.4 Ablation of Retrieval Branches

Within the hierarchical retrieval framework, we examine how the graph and explicit-feature branches contribute to answer quality. We compare GNN-only retrieval, lexical-only retrieval, and their fusion at the same retention setting $b = 320$. GNN-only retrieval uses the graph branch described in Section 3.3; lexical-only retrieval uses the explicit-feature branch, which combines lexical matches, named-source conditions, and section-role features.

Table 3: Retrieval-branch ablation at $b = 320$. Paper F1 uses 116 text queries; Fact F1 uses the 24 factual queries. Scores are percentages.

Retrieval variant	Paper F1 ↑	Fact F1 ↑
GNN only	39.27	50.54
Lexical only (explicit features)	48.05	60.15
RAG-SCHOLAR (fused)	**50.36**	**69.42**

Table 3 shows that fusion achieves the highest scores on both metrics. Relative to GNN-only retrieval, it improves Paper F1 by 11.09 points and Fact F1 by 18.88 points. Relative to lexical-only retrieval, the gains are 2.31 and 9.27 points, respectively. This pattern supports the complementary role of graph representations in source selection alongside explicit scholarly features.

Table 4: Entity-node ablations: Full retains the index; w/o removes the named type. Parenthesized percentages report the relative change from w/o. Direct targets that type; Related includes compositions involving it. Pairs use entity-specific subsets. Complete metrics are in Appendix I.1.

	Direct queries				Related queries			
	Fact F1 ↑		Context tok. ↓		Fact F1 ↑		Context tok. ↓	
Entity	w/o	Full	w/o	Full	w/o	Full	w/o	Full
Author	62.50	**64.58** (+3.3%)	888.6	**652.4** (-26.6%)	62.50	**64.58** (+3.3%)	888.6	**652.4** (-26.6%)
Institution	8.33	**82.54** (+890.9%)	1056.0	**86.8** (-91.8%)	35.00	**83.52** (+138.6%)	1074.8	**458.9** (-57.3%)
Method	5.14	**84.68** (+1547.5%)	1048.8	**116.8** (-88.9%)	34.82	**83.31** (+139.3%)	1073.6	**485.1** (-54.8%)
Dataset	24.17	**86.82** (+259.2%)	1112.9	**102.6** (-90.8%)	47.39	**90.80** (+91.6%)	1089.6	**558.3** (-48.8%)
Model	7.50	**81.22** (+982.9%)	1109.0	**127.2** (-88.5%)	42.38	**86.73** (+104.6%)	1080.2	**525.9** (-51.3%)
Reference	1.52	**83.62** (+5401.3%)	1105.9	**708.2** (-36.0%)	1.52	**83.62** (+5401.3%)	1105.9	**708.2** (-36.0%)

Table 5: Visual-cropping ablation on the multimodal suite. Total excludes preprocessing. Factual precision/recall appears in Appendix I.2.

	Answer quality ↑		Answer-model input (tokens) ↓		
Variant	Paper F1	Fact F1	Text	Visual	Total
Visual selection: multimodal suite					
No crop	**100.00**	**60.22**	**1090.70**	523.87	1614.57
Crop (Ours)	**100.00** (+0.0%)	60.00 (-0.4%)	**1090.70** (+0.0%)	**217.80** (-58.4%)	**1308.50** (-19.0%)

4.5 Component Contributions to Context Use

Table 4 evaluates entity nodes on direct and related queries. Removing method nodes lowers Fact F1 from 84.68 to 5.14 on direct queries and from 83.31 to 34.82 on related queries, despite substantially longer contexts. Institution, dataset, model, and reference ablations show the same pattern: more retained text does not compensate for missing scholarly index signals. On related queries, removing dataset or model nodes roughly doubles context length while lowering Fact F1 by 43.41 and 44.35 points, respectively. This extends the observed quality–cost pattern to the broader query populations involving these entities.

Table 5 compares crop selection with uncropped visual input. Visual selection reduces visual tokens by 58.42% and total answer-model input by 18.96%. Paper F1 remains 100.00 and Fact F1 differs by only 0.22 points (0.4%). Localizing visual evidence therefore substantially reduces input while retaining nearly the same aggregate factual F1. The two variants retain the same average text input, so the savings arise entirely from the visual channel.

5 Conclusion

We presented RAG-Scholar for retrieving compact, complementary evidence from scholarly corpora with substantial conceptual overlap. Its central design separates source identification from the evidence retained for answering. Hierarchical retrieval controls source contributions; the graph-guided scholarly index combines explicit source and role features with learned relational representations; and visual retrieval selects regions relative to retained text. The experiments demonstrate a favorable quality–cost trade-off, consistent advantages over flat retrieval across the measured corpus sizes, and substantial visual-token reductions with nearly unchanged aggregate factual F1. These findings support a unified view of scholarly retrieval: shared concepts help identify relevant sources, while source-specific and cross-modal distinctions determine which evidence deserves the limited answer context.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript text and assist with code debugging. We did not use LLMs to formulate the core research ideas or design the proposed method. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docvqa: Multi-modal multi-page multi-document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 6237–6247, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/Findings/html/Cho_M3DocVQA_Multi-modal_Multi-page_Multi-document_Understanding_ICCVW_2025_paper.html.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, New York, NY, USA, 2009. URL http://doi.acm.org/10.1145/1571941.1572114.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL-HLT*, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/99e9e141aafc314f76b0ca3dd66898b3-Paper-Conference.pdf.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569. Curran Associates, Inc., 2024. doi: 10.52202/079017-1902. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ddc001d07ca4f319af96a3024f6dbd1-Paper-Conference.pdf.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. MDocAgent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025. URL https://arxiv.org/abs/2503.13964.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. URL https://arxiv.org/abs/2205.10803.

Yeonjun In, Sungchul Kim, Ryan A. Rossi, Mehrab Tanjim, Tong Yu, Ritwik Sinha, and Chanyoung Park. Diversify-verify-adapt: Efficient and robust retrieval-augmented ambiguous question answering. 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. 1212–1233, 2025. doi: 10.18653/v1/2025.naacl-long.56. URL https://aclanthology.org/2025.naacl-long.56/.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of EMNLP*, 2023. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of ACL*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17606–17619, 2025. doi: 10.18653/v1/2025.acl-long.861. URL https://aclanthology.org/2025.acl-long.861/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.11401.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. RegionRAG: Region-level retrieval-augmented generation for visual document understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):6662–6670, 2026. doi: 10.1609/aaai.v40i8.37597. URL https://ojs.aaai.org/index.php/AAAI/article/view/37597.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of EMNLP*, pp. 6342–6353, 2023. doi: 10.18653/v1/2023.emnlp-main.391. URL https://aclanthology.org/2023.emnlp-main.391/.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of ACL*, pp. 963–981, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8a2acd174940dbca361a6398a4f9df91-Paper-Conference.pdf.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web. ESWC 2018*, volume 10843 of *Lecture Notes in Computer Science*, pp. 593–607. Springer, 2018. doi: 10.1007/978-3-319-93417-4_38. URL https://link.springer.com/chapter/10.1007/978-3-319-93417-4_38.

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

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bda88ed2892f5e61c9a9bf215c566913-Paper-Conference.pdf.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. CompAct: Compressing retrieved documents actively for question answering. In *Proceedings of EMNLP*, 2024. doi: 10.18653/v1/2024.emnlp-main.1194. URL https://aclanthology.org/2024.emnlp-main.1194/.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-Based Retrieval-Augmented Generation on Multi-Modality Documents. In *International Conference on Learning Representations*, 2025. URL https://mlanthology.org/iclr/2025/yu2025iclr-visrag/.

Yongyue Zhang and Yaxiong Wu. MLDocRAG: Multimodal long-context document retrieval augmented generation. *arXiv preprint arXiv:2602.10271*, 2026. URL https://arxiv.org/abs/2602.10271.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. LongRAG: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, 2024. doi: 10.18653/v1/2024.emnlp-main.1259. URL https://aclanthology.org/2024.emnlp-main.1259/.

Yilong Zuo, Xunkai Li, Jing Yuan, Qiangqiang Dai, Hongchao Qin, and Ronghua Li. MAGE-RAG: Multigranular adaptive graph evidence for agentic multimodal RAG in long-document QA. *arXiv preprint arXiv:2606.15906*, 2026. URL https://arxiv.org/abs/2606.15906.

A Dataset Construction and Evaluation Suites

A.1 Scholarly Corpus

We construct a core collection of 146 open-access papers in computer science and natural language processing, covering retrieval-augmented generation, long-context modeling, document understanding, and graph-enhanced reasoning. Each PDF is parsed into sections, text chunks, and figure assets, which form the document hierarchy of the graph-guided scholarly index in Section 3.3. The index is built from MinerU `content_list` outputs, with paragraph and image JSONL records as a fallback. Cached GROBID parses provide additional paper metadata when available. Table 6 summarizes the resulting index.

Table 6: Core-graph node and edge statistics. Cross-paper edges are counted as relation records, whereas distinct paper pairs count each pair once across relation types and directions.

Node type	Count	Node type	Count
Paper	146	Author	336
Section	3,937	Institution	92
Text chunk	9,028	Reference	6,403
Figure	3,023	Method	306
Dataset	284	Task	2,304
Metric	10	Model	115
Venue	11		
Edge and paper-pair counts			
Total edges	60,062	Cross-paper edges	1,745
Distinct paper pairs	1,573		
Cross-paper relation	Count	Cross-paper relation	Count
Bibliographic coupling	1,231	Shared author	2
Resolved citation	59	Shared institution	8
Weak citation	20	Shared dataset	125
Shared method	138	Shared model	162

A.2 Query Suites and Reference Facts

The text suite contains 116 queries phrased as natural user requests, each assigned a unique `query_id`. We group them by information need in Table 7. Paper-set F1 is evaluated on all 116 queries. The multi-constraint and two explanatory categories form a fixed subset of 24 queries with 72 reference propositions for Fact Precision, Recall, and F1.

Table 7: Text-suite queries grouped by information need. The 24 multi-constraint and explanatory queries form the factual-evaluation subset.

Information need	Count	Information need	Count
Known-paper lookup	18	Research provenance	20
Technical focus	30	Experimental resources	12
Citation tracing	12	Multi-constraint search	12
Single-paper explanation	6	Cross-paper comparison	6
Total	116		

The multimodal suite contains 15 figure-lookup queries and 15 figure-grounded reasoning queries (Table 8). The reasoning subset is evaluated against 45 fixed core facts. All 30 queries contribute to Paper-set F1; only the reasoning subset contributes to factual scores. The queries cover pipelines, frameworks, workflows, embedding plots, charts, and heatmaps.

Table 8: Multimodal-suite categories.

Lookup category	Count	Reasoning category	Count
Pipeline	6	Workflow	7
Framework	5	Diagram	4
Long-context figure	2	Embedding plot	1
Workflow	1	Chart	1
Embedding plot	1	Heatmap	1
		Multimodal	1
Total	15	Total	15

A.3 Reference Construction and Subset Membership

All evaluation sets are manually authored, including the text and multimodal benchmark queries and their gold annotations. Human review verifies the queries, target paper sets, and expected answer content. Each query is associated with a gold paper set and, for factual queries, a set of reference propositions. Visual reasoning is anchored to the target visual and its local paper record. These records define what is credited by the metrics in Appendix D.

The reviewed query specifications and source-verified relations are organized into reference propositions. During human review, each paper–entity relation label is checked against the original paper and finalized independently of the entity extraction and relation construction used to build the evaluated index. Explanatory and visual-reasoning propositions follow the expected answer points, while compositional propositions express the paper–entity relations in the query metadata. Gold content is drawn from these benchmark specifications rather than the answers produced by the evaluated systems. Each proposition has a fact identifier linked to its query and, through the query record, to the source paper identifiers and titles. Relation facts also retain the entity-node identifier for record alignment, and visual queries provide the target modality and a caption hint locating the relevant figure.

System answers are assessed in the subsequent claim-annotation stage. Each claim is one independently verifiable proposition, with conjunctions split before assessment. Claims are linked to reference propositions through query and fact identifiers. When the reference set does not determine claim correctness, the claim is checked against the source paper. Claims verified as correct but not matched to a reference proposition are recorded as supported additional content and excluded from U_q. Visual claims are assessed against the target figure and its local paper record.

Entity ablations use two scopes for each removed type. Direct queries request that entity type alone; related queries include every question involving the type, including compositional constraints. The direct/related manifests contain 8/8 Author, 12/20 Institution, 12/19 Method, 12/23 Dataset, 12/21 Model, and 12/12 Reference queries.

B Supplementary Method Details

B.1 Entity Normalization and Confidence

For a mention m of type τ, let $k_\tau(m)$ denote the normalized key of its canonical name after type-specific alias resolution. Each existing entity $u \in V^\tau$ is indexed by the corresponding key $k_\tau(u)$, which is unique within its type. The normalization map is

$$\phi_\tau(m) = \begin{cases} u, & \text{if } u \in V^\tau \text{ and } k_\tau(u) = k_\tau(m), \\ u_{\text{new}}(m), & \text{otherwise.} \end{cases} \tag{12}$$

Here, $u_{\text{new}}(m)$ denotes a new entity node indexed by $k_\tau(m)$. Mentions that resolve to the same canonical key share an entity node while retaining their source associations. For document node v and entity u,

$$w(v,u) = \max_{m \in \mathcal{M}(v,u)} \text{conf}(m), \tag{13}$$

where $\mathcal{M}(v,u)$ contains the mentions resolved to u; an absent edge has zero weight. Paper, section, and figure nodes may carry entity edges. Maximum-confidence aggregation preserves the strongest mention support without summing repeated mentions as independent evidence. Semantic mentions are extracted using a type-specific alias table and dynamic regular-expression rules. Aliases are matched in decreasing order of length, and shorter matches contained within an accepted span are suppressed within the same entity type. Alias-table matches receive confidence 0.94, while dynamic patterns use rule-specific scores that can depend on the section role and source. The default thresholds for retaining mentions are 0.72 for section text, 0.78 for paper profiles comprising the title, abstract, and keywords, and 0.80 for figure captions and OCR text. These thresholds determine which extracted mentions contribute document–entity edges.

B.2 Cross-Paper Affinity

Let $\mathcal{N}_{\text{ent}}(p)$ contain non-reference entity neighbors and $R(p)$ contain cited-reference nodes. The affinity from shared entities and shared references is given by

$$\begin{aligned} w_\times(p_i,p_j) = &\sum_{u \in \mathcal{N}_{\text{ent}}(p_i) \cap \mathcal{N}_{\text{ent}}(p_j)} \lambda_{\tau(u)} \rho(u) h(c_{iu}, c_{ju}) \\ &+ \mu \sum_{r \in R(p_i) \cap R(p_j)} \rho(r) h(c_{ir}, c_{jr}), \end{aligned} \tag{14}$$

where $c_{iu} = w(p_i,u)$, $\rho(u) = \log(1 + N/\deg(u))$, and $\deg(u)$ counts incident papers. Confidence support is combined by

$$h(a,b) = \begin{cases} 2ab/(a+b), & a+b > 0, \\ 0, & a+b = 0. \end{cases} \tag{15}$$

The same inverse-frequency weighting applies to references. The non-reference sum uses paper-level author, institution, venue, method, dataset, and model links, with respective default weights of 3.0, 1.4, 1.0, 2.2, 2.4, and 1.8. Shared references use $\mu = 1.6$. In both terms, $\deg(u)$ counts distinct incident papers, retaining the highest-confidence connection when a paper has multiple links to the same entity. The sums include shared entities with $2 \le \deg(u) \le 10$ and shared references with $2 \le \deg(r) \le 16$. Bibliographic coupling uses informative reference identifiers, excluding generic reference keys and incomplete DOI fragments. Contributions are summed within each relation type for each paper pair, retaining records with weight at least 0.55. Shared-entity and bibliographic-coupling relations are symmetric; resolved and weak citations are constructed separately and stored from citing to cited papers. These relations provide cross-paper support for retrieval.

B.3 Visual Candidates and Window Constraints

Figure selection follows equation 10: the explicit and graph lists are fused within each selected section to obtain $\mathcal{F}(s, q)$. Corpus-wide figure matches can guide source selection through the explicit branch. The retained local figures supply the images for window selection below.

For a candidate window w, we use the query-relevance and redundancy terms in Section 3.4, with normalized area $a(w) = |w|/|I_f|$:

$$S_w(w \mid q, f, \mathcal{H}_q) = u(w, q) - \alpha_v r(w, q) - \beta_v a(w). \tag{16}$$

The coefficients are $\alpha_v = 0.75$ and $\beta_v = 0.08$. We detect a non-white content region with 8-pixel padding and generate candidate windows from 28 size–aspect templates on a 9×9 anchor grid. The full image and content region remain fallback candidates. We evaluate at most 360 candidate windows using CLIP ViT-B/32 against the query and up to ten retained text snippets associated with the figure's section. The redundancy term is zero when no snippet is available.

Candidates are considered in decreasing order of S_w and retained greedily. Selection retains at most three windows under the constraints

$$\begin{aligned} &\mathrm{IoU}(w, v) < \tau_{\mathrm{IoU}} \quad \text{for distinct } w, v \in \mathcal{W}_f^*, \\ &\sum_{w \in \mathcal{W}_f^*} a(w) \le A_{\max}, \end{aligned} \tag{17}$$

where $\tau_{\mathrm{IoU}} = 0.5$ and $A_{\max} = 0.55$. If no window satisfies these constraints, the highest-ranked non-maximum-suppressed candidate is retained.

The selected windows are rendered in `union` mode as a single enclosing crop at native resolution, and at most one image is attached to the answer model. Normalized window area guides spatial selection, while the enclosing crop also includes the pixels between selected windows and can therefore cover a larger area. Visual-token cost is measured from this rendered image using the answer model's image processor.

B.4 Retrieval and Context Assembly

The retrieval procedure implements the three components in Section 3:

1. Compute direct paper relevance S_0 from paper descriptors and canonical entities; form the anchor set and include relation support in S_p.
2. Retain the top K_p papers and the top K_s sections per paper. Construct local candidates from up to K_c chunks and K_f figures per section.
3. Apply the text-retention controller and retain paper/section identifiers with the selected text.
4. For visual queries, combine the configured local and global figure candidates; score figures and windows against the query and retained scholarly context.
5. Render the selected visual input, measure its processor cost, and assemble the retrieved context for the answer model.

The default retrieval limits are $K_p = 5$, $K_s = 2$, $K_c = 2$, and $K_f = 1$, with up to five entity matches and two cross-paper neighbors. During context assembly, the bundle is serialized as Markdown and checked against a 6,500-character cap and, when enabled, an estimated token cap. Pruning begins with entity matches that are neither exact matches nor retained as query evidence, followed by institution and author metadata, lower-priority selection evidence, cross-paper neighbors, and trailing chunks. Section entities and summaries are then removed before pruning empty or lower-ranked sections and, if needed, figures or papers according to the query modality. The bundle is reserialized after each removal.

In `text_plus_images` mode, the selected visual input is attached to the answer model. The answer model generates one answer per query and evaluated configuration, with retries for execution failures.

C GNN ARCHITECTURE AND TRAINING

This section supplements the graph-guided scholarly index with the frozen input features, relation-aware network, self-supervised objective, and graph-recall configuration.

C.1 FROZEN NODE FEATURES

We use the frozen `bce-embedding-base_v1` encoder with CLS pooling, 768-dimensional normalized outputs, a 256-token input limit, and encoding batches of 16. The query uses the same encoder. For papers, title, abstract, keywords, authors, institutions, and venue are encoded independently. Let $\mathcal{J}_p$ index the nonempty fields of paper p, $\mathcal{W}_{pj}$ be the nonoverlapping token windows of field j, and E_{txt} be the frozen encoder. With $\mathbf{x}_{pj}$ denoting the field vector, $\mathbf{x}_p$ the paper vector, and norm denoting ℓ_2 normalization, window and field pooling are

$$\mathbf{x}_{pj} = \text{norm}\left(\frac{1}{|\mathcal{W}_{pj}|}\sum_{w\in\mathcal{W}_{pj}} \text{norm}(E_{\text{txt}}(w))\right), \qquad \mathbf{x}_p = \text{norm}\left(\frac{1}{|\mathcal{J}_p|}\sum_{j\in\mathcal{J}_p}\mathbf{x}_{pj}\right). \tag{18}$$

Windows include all field tokens while respecting the encoder limit after special tokens. Equal field weighting prevents a long abstract from dominating author or institution metadata. Author, institution, and venue entities also use full-window encoding. Other nodes use their own descriptors with the 256-token limit: section title/summary/text, chunk text, figure caption/OCR/footnote, or canonical entity names and aliases. Figures are therefore represented by text at this stage, not pixels.

C.2 RELATION-AWARE NETWORK ARCHITECTURE

The graph encoder has two message-passing layers with hidden dimension 128. Its propagation edge multiset E_{msg} contains the original edges and separately parameterized reverse relations, yielding 70 relation channels for the 13 node types. For a message $(u, r, v) \in E_{\text{msg}}$ from source u to target v through relation r, let c_{urv} be its confidence, d_{vr} the total incoming confidence for relation r, and n_v^+ the number of positive-confidence relation groups at v. The normalized weight $\bar{c}_{urv}$ is defined with $\epsilon = 10^{-12}$:

$$d_{vr} = \sum_{u:(u,r,v)\in E_{\text{msg}}} c_{urv}, \qquad n_v^+ = |\{r : d_{vr} > 0\}|, \qquad \bar{c}_{urv} = \frac{c_{urv}}{\max(d_{vr},\epsilon)\max(n_v^+, 1)}. \tag{19}$$

This normalizes within each relation before averaging across active relations, preventing high-count edge types from dominating by count alone.

The input projection, gated message update, and output residual are defined in equation 6, equation 7, and equation 8. The normalization above supplies their edge weights $\bar{c}_{urv}$. This relation-aware parameterization follows typed message passing (Schlichtkrull et al., 2018); relation gates are initialized to ones and do not implement attention softmax. The neighbor projection has no bias, whereas the input, self, and output projections have biases. The GNN has 283,392 trainable parameters, excluding the frozen text encoder.

C.3 SELF-SUPERVISED TRAINING

At each epoch, each node feature row is independently masked with probability 0.2, with at least one masked node ensured. Let $\mathcal{M} \subseteq V$ be the masked set, $\widetilde{X}$ the feature matrix with masked rows set to zero, and $\widehat{\mathbf{z}}_v$ the normalized network output for node v. Write $\widetilde{\mathbf{x}}_v$ for its row associated with node v. Node types and edges remain visible. The residual in equation 8 uses $\widetilde{\mathbf{x}}_v$, so masked targets cannot bypass reconstruction. With the frozen uncorrupted target $\mathbf{x}_v$, masked reconstruction loss $\mathcal{L}_{\text{mask}}$ and visible-node anchoring loss $\mathcal{L}_{\text{anchor}}$ are

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|}\sum_{v\in\mathcal{M}}\left(1 - \widehat{\mathbf{z}}_v^\top \mathbf{x}_v\right),$$
$$\mathcal{L}_{\text{anchor}} = \frac{1}{|V\setminus\mathcal{M}|}\sum_{v\notin\mathcal{M}}\left(1 - \widehat{\mathbf{z}}_v^\top \mathbf{x}_v\right). \tag{20}$$

The anchor loss is zero if no visible node remains. Feature reconstruction supplies corpus-level self-supervision, as in masked graph representation learning (Hou et al., 2022).

For a randomly sampled masked-node subset $\mathcal{B}$ of at most 256 nodes, let $\mathcal{B}_v^+ = \{u \in \mathcal{B} : \mathbf{x}_v^\top \mathbf{x}_u > 0.99999\} \cup \{v\}$ be node v's positive set. Using temperature $T = 0.1$, the multi-positive contrastive loss $\mathcal{L}_{\text{con}}$ is

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{B}|} \sum_{v \in \mathcal{B}} \log \frac{\sum_{u \in \mathcal{B}_v^+} \exp(\widehat{\mathbf{z}}_v^\top \mathbf{x}_u / T)}{\sum_{u \in \mathcal{B}} \exp(\widehat{\mathbf{z}}_v^\top \mathbf{x}_u / T)}. \tag{21}$$

Nearly identical features are thus positives rather than false negatives. The total training objective $\mathcal{L}$ combines these three losses with unit reconstruction/anchor weights and contrastive weight 0.1:

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{anchor}} + 0.1\mathcal{L}_{\text{con}}. \tag{22}$$

We train on the full corpus graph with no query labels or answer-model gradients. Table 9 gives the optimization settings. After 40 epochs, the final model exports node vectors using the complete, unmasked features; there is no retrieval-based checkpoint selection.

Table 9: Configuration of the 146-paper GNN index. Encoder batch size does not denote the GNN optimization batch size.

Setting	Value
Hidden dimension / GNN layers	128 / 2
Output dimension / residual scale	768 / $\eta = 0.25$
Epochs / seed	40 / 42
Optimizer / learning rate	AdamW / 10^{-3}
Weight decay / gradient-norm clip	10^{-4} / 1.0
Node mask rate	0.2
Contrastive batch / temperature	At most 256 masked nodes / 0.1
Anchor / contrastive loss weights	1.0 / 0.1
Frozen encoder batch / input length	16 / 256 tokens
GNN update scope	Full graph each epoch

C.4 Retrieval with Graph Representations

In equation 9, content nodes support their owning paper; entities support papers linked through positive-confidence content edges. Let $\mathcal{T}(q)$ be the active node-type channels, $\rho_t(p, q)$ paper p's competition rank under channel score $a_t(p, q)$, and λ_t its nonnegative channel weight. With smoothing constant $k = 60$, the graph paper score is

$$S_{\text{G}}(p \mid q) = \sum_{t \in \mathcal{T}(q)} \frac{\lambda_t}{k + \rho_t(p, q)}. \tag{23}$$

Missing channel entries contribute zero. Non-paper channels discard nonpositive node cosines and retain paper ranks at most 32, including ties; the paper channel retains all papers. Competition ranks assign equal scores equal ranks, such as $1, 1, 3$. Query entity-type cues activate available entity channels; otherwise all entity types are active. Table 10 gives the weights. Local section/chunk/figure lists use node cosines within their parent scopes.

Table 10: Inner paper-recall weights λ_t. The entity weight is divided equally among active entity types.

Channel	No effective entity cue	Effective entity cue
Paper	1.0	1.0
Section	1.0	0.5
Chunk	1.0	0.5
Figure	0.5	0.25
All active entity types combined	1.0	3.0

The graph lists enter the outer fusion in equation 2. The inner channel weights λ_t, outer branch weight $\alpha = 0.3$, and representation residual $\eta = 0.25$ operate on different quantities and are not interchangeable. Appendix B.4 specifies the prefetch and source-constraint behavior.

D Evaluation Metrics

D.1 Paper Identification and Factual Answering

For query q, let $\widehat{\mathcal{P}}_q$ be the predicted paper set and $\mathcal{P}_q$ the gold set. Paper-set F1 is

$$\mathrm{F1}_q^{\mathrm{paper}} = \frac{2|\widehat{\mathcal{P}}_q \cap \mathcal{P}_q|}{|\widehat{\mathcal{P}}_q| + |\mathcal{P}_q|}. \tag{24}$$

The reported score is the query-level macro-average. This metric evaluates paper identification in the answer. The answer-only resolver matches normalized and compact title variants to stored paper IDs. When a title variant matches multiple papers, their distinct IDs are retained in corpus-index order. Repeated matches to the same ID are merged. Explicit title matches take priority. Otherwise, the resolver may infer a set from affirmed query constraints or entities mentioned in the answer. Precision and recall use $\max(1, |\widehat{\mathcal{P}}_q|)$ and $\max(1, |\mathcal{P}_q|)$ as denominators, respectively, and F1 is zero when both scores are zero.

Let $\mathcal{G}_q$ contain the reference facts, $\mathcal{T}_q$ the distinct reference facts correctly expressed in the answer, and U_q the number of factual assertions contradicted by or unsupported by the source evidence. Then

$$P_q = \frac{|\mathcal{T}_q|}{|\mathcal{T}_q| + U_q}, \qquad R_q = \frac{|\mathcal{T}_q|}{|\mathcal{G}_q|}, \qquad F_q = \frac{2P_qR_q}{P_q + R_q}. \tag{25}$$

Each reference fact is credited at most once. Precision, recall, and F1 are macro-averaged separately over the relevant factual queries; reported macro-F1 is not recomputed from the displayed macro-precision and macro-recall.

A reference fact is credited only when the answer correctly expresses it and satisfies the applicable paper-linking requirement. A substantive claim without a reference-fact match is checked against the source paper or target visual. A source-supported additional claim contributes to neither $\mathcal{T}_q$ nor U_q. A claim contradicted by or unsupported by the source evidence contributes to U_q and cannot establish reference-fact coverage. A direct-list fact is credited when the answer names the corresponding gold paper, since the query supplies the entity condition. Hard-composition facts additionally require a correct positive entity-relation assertion and either a correct saved paper-selection link or an explicit gold-paper mention.

D.2 Aggregation and Evaluation Populations

Text Paper-set F1 uses 116 queries; text Fact P/R/F1 use 24 queries and 72 propositions. Multimodal Paper-set F1 uses 30 queries; multimodal Fact P/R/F1 use 15 reasoning queries and 45 facts. Entity ablations use the type-specific subsets in Appendix A. All quality scores are percentages.

E Implementation and Cost Accounting

E.1 Retention Settings and Input Cost

The text sweep includes eight finite controller settings,

$$b \in \{100, 140, 180, 240, 320, 480, 640, 800\},$$

and an uncapped setting. Table 1 shows six representative configurations; Appendix F retains the full sweep, including $b = 140, 180, 800$. The controller setting b is an estimated-token retention target for the text-budget formulation in equation 1; for example, $b = 100$ produces 174.42 average retained tokens. The corpus-expansion experiment uses $b = 240$ and a 6,500-character cap. For the serialized bundle s, pruning tests the following character and estimated-token targets:

$$|s| \le 6500, \qquad \widehat{\tau}_t(s) = \lceil 0.9\, n_{\mathrm{re}}(s) \rceil \le b, \tag{26}$$

where $|s|$ is the character count and $n_{\mathrm{re}}(s)$ counts individual CJK characters, maximal runs of ASCII letters, digits, or underscores, and individual remaining non-whitespace characters. The bundle is

reserialized after each removal and both criteria are checked. Pruning stops when the targets are met or no further fields can be removed under the paper-preservation rules. The reported Input count measures the resulting retained context. The setting referred to as "uncapped" disables only the estimated-token condition; the 6,500-character cap remains active.

In the text comparison, Input is the context-token count recorded for each query after retrieval and any compression, stored as `context_token_count` in the evaluation records. Tables report its mean over the evaluated queries. Pre. counts online preprocessing tokens, and Total = Pre. + Input. Context and the multimodal Text column denote the same text-context quantity. In multimodal and paired-ablation tables, preprocessing token counts are zero for all configurations, so the Pre. column is omitted and Total = Text + Visual. Generated output tokens are excluded from these input-cost measures.

For all methods, Input is computed by tokenizing the final serialized context with `cl100k_base`. The estimator $\widehat{\tau}_t$ is used for retention control. The counted context includes retained source identifiers, provenance fields, section and chunk text, figure captions, and OCR-derived text, together with the query when included in the serialization. Outer task and system instructions, the output schema, chat-template framing, and generated output are excluded. Pre. counts tokens processed across online compression passes using each preprocessing model's tokenizer, including instructions supplied to those passes.

For RAG-SCHOLAR's multimodal runs and the crop ablation, Visual counts merged image-patch tokens. The attached images are resized by the Qwen2-VL image processor loaded from the answer checkpoint through `AutoProcessor`. Given image grids (T_j, H_j, W_j) and spatial merge size μ, the count is $\sum_j T_j H_j W_j / \mu^2$, where T_j is the temporal grid size and H_j, W_j are the spatial grid sizes.

E.2 TIMING AND IMPLEMENTATION SETTINGS

Prep. is the mean wall-clock time from the start of query-time retrieval or compression until the text context and any attached images are ready for answer generation. It includes retrieval, bundle construction, evidence selection and serialization, and online compression or crop preparation. Corpus-expansion measurements use deterministic replay through retrieval and prompt construction, stopping before generation.

Queries are processed sequentially after a single setup phase, reusing loaded resources and in-process caches. Every query, including the first, contributes to the mean. Initial setup is timed separately, while initialization triggered within a query remains in its timer. Offline PDF parsing, entity extraction, graph construction, feature indexing, and storage are excluded.

Experiments use two NVIDIA GeForce RTX 3090 GPUs (24 GB VRAM each), two Intel Xeon Silver 4314 CPUs (32 physical cores and 64 threads in total), and CUDA 12.4. Qwen2-VL inference uses FP16 on the GPU, while CLIP window scoring runs on CPU.

Default chunk lengths are 350, 700, and 1,100 characters for the minimum, target, and maximum, respectively. Each figure is linked to at most two nearby chunks. Lexical relevance uses character 2–5-gram TF–IDF cosine similarity. Multimodal runs rank chunks within the selected papers.

Text-retention and multimodal runs use Qwen2-VL-7B-Instruct with its chat template, greedy decoding (temperature 0), and a 700-token output limit. Text-retention runs request at least four sentences, while multimodal runs request two to four sentences.

Table 11 presents the shared user-message template and output schema. Braced fields denote query inputs or task-dependent instructions. The fields `selection_instruction`, `answer_instruction`, `image_instruction`, and `answer_schema_hint` specify the paper-selection rule, answer style, image-use guidance, and expected answer description, respectively.

Table 11: Shared answer-generation prompt template and output schema.

Task: Answer the user query using only the provided paper context bundle. You must choose from the candidate papers listed below and must not invent any new paper IDs.

Table 11: Shared answer-generation prompt template and output schema (continued).

User query: {query}
Candidate papers: {candidate_block}
Attached image manifest: {attached_image_manifest}
Context bundle: {context_bundle}
Instructions:
1. Only use paper IDs that appear in the candidate paper list.
2. {selection_instruction}
3. If no candidate paper truly answers the query, return an empty list.
4. Treat `query_match_evidence`, `matched_entities`, and `cross_paper_neighbors` as graph-derived evidence signals and prioritize them when they directly match the query.
5. {answer_instruction}
6. The natural-language `answer` is the main user-facing response. It should not mention evaluator logic, `support_type`, `fact_type`, or verification rules.
7. Use `structured_answer` only as a compact audit trail for the claims you already made in the answer; do not introduce new unsupported details there.
8. In `structured_answer.paper_answers`, include only the papers you predicted.
9. Each `supporting_facts` item should cover one important query-relevant claim, not a generic restatement of the whole answer.
10. Prefer 1-3 high-confidence supporting facts per paper. For multi-constraint queries, cover distinct constraints such as institution, method, dataset, model, reference, or task instead of repeating the same one.
11. Every `support.support_value` should be short, high-precision, and directly traceable to the bundle. Prefer exact short phrases copied from the bundle over paraphrases.
12. Prefer support types in this order when possible: `matched_entity`, `query_match_evidence`, `section_summary`, `figure_caption`, `chunk_text`, then broader fields like `authors`, `institutions`, `paper_title`, or `cross_paper_neighbor`.
13. Use `chunk_text` only when needed, and copy a short contiguous span from the bundle rather than writing a new sentence. Avoid long quotations.
14. Keep each `support.support_value` compact, ideally a short phrase rather than a full sentence, to preserve JSON stability.
15. If you cannot point to a verifiable support row for a claim, omit that claim from `structured_answer` and keep the natural-language answer cautious.
16. Allowed `fact_type` values: `institution`, `author`, `reference`, `method`, `dataset`, `model`, `task`, `topic`, `visual`, `other`.
17. Allowed `support_type` values: `paper_title`, `authors`, `institutions`, `query_match_evidence`, `matched_entity`, `section`, `section_summary`, `chunk_text`, `figure_caption`, `cross_paper_neighbor`.
18. If a detail is uncertain, omit it instead of inventing it.
19. {image_instruction}
20. Return JSON only.
Return this JSON schema:

```
{"predicted_paper_ids": ["paper_id_1", "paper_id_2"], "predicted_paper_titles": [
"paper title 1", "paper title 2"], "answer": "{answer_schema_hint}", "evidence":
[{"paper_id": "paper_id_1", "reason": "why it matches"}], "structured_answer": {
"paper_answers": [{"paper_id": "paper_id_1", "paper_title": "paper title 1",
"match_summary": "1-2 sentence explanation of why this paper matches the query",
"supporting_facts": [{"fact_type": "method", "claim": "short claim about the
paper", "support": [{"support_type": "matched_entity", "support_value": "exact
method or entity phrase from the bundle"}, {"support_type": "chunk_text",
"support_value": "short quoted span from a chunk or summary"}]}]}],
"overall_conclusion": "short overall conclusion across the selected papers",
"abstained": false}}
```

E.3 Baseline and Ablation Configurations

Comparisons cover three families: compact-evidence methods, CompAct and ReComp (Yoon et al., 2024; Xu et al., 2024); page-level visual-retrieval methods, ColPali, M3DocRAG, VisRAG, and MDocAgent (Faysse et al., 2025; Cho et al., 2025; Yu et al., 2025; Han et al., 2025); and long-context compression methods, LLMLingua-2, LongLLMLingua, Selective Context, and LLMLingua (Pan et al., 2024; Jiang et al., 2024; Li et al., 2023; Jiang et al., 2023). ColPali denotes the visual-retrieval component within its answering pipeline. Venue tags follow the source publications. ICCVW denotes ICCV Workshops: M3DocRAG appears in the published paper *M3DocVQA: Multi-modal Multi-page Multi-document Understanding* (Cho et al., 2025). ACL Findings is distinguished from the ACL main conference; MDocAgent is labeled by its arXiv preprint (Han et al., 2025).

All text-suite baselines use Qwen2-VL-7B-Instruct at temperature 0. Their retrieval or compression stages operate over the same 146-paper collection. Answer prompts supply the query and source-labeled context, requesting answers based on that evidence and exact titles for any identified papers.

Compression pipelines provide retained textual evidence, while page-retrieval pipelines provide extracted page text for text queries.

CompAct uses CompAct-7b to process the top 30 BM25-ranked section blocks in groups of five, with at most six compression iterations and a 500-token generation limit per iteration. ReComp uses its trained NQ extractive compressor to rank sentences from the top eight retrieved sections, considering up to 30 sentences per section and retaining the top two. LLMLingua-2 uses the XLM-RoBERTa-large MeetingBank checkpoint, while LLMLingua and LongLLMLingua use Qwen2.5-1.5B-Instruct for compression. All three use a 2,500-token target. LongLLMLingua additionally applies query-conditioned context ordering and dynamic compression. Selective Context scores English phrases with GPT-2 and retains scored units under a 2,500-token budget.

The ColPali pipeline uses ColQwen2-v1.0. M3DocRAG uses ColPali-v1.2 with the ColPaliGemma-3B backbone, and VisRAG uses VisRAG-Ret. MDocAgent's text-suite adaptation uses its ColBERTv2.0 retrieval branch with the shared answer model. Each page-based baseline ranks 48 candidates from 2,744 pages. The retained page-text budgets are 1,400 tokens for ranked queries and 2,200 for exhaustive queries. The multimodal ColPali, M3DocRAG, and VisRAG runs additionally attach selected page images.

For each of the six entity types (author, institution, method, dataset, model, and reference), the ablation removes all nodes of that type and their incident edges. It also removes cross-paper relations whose recorded evidence contains a deleted entity or is attributed to the removed type, then rebuilds paper-pair summaries from the remaining relations. Paper, section, chunk, and figure records retain their original text and embedded metadata. Entity names can therefore remain accessible through textual evidence, while the deleted nodes no longer supply entity matches or cross-paper support. Retrieval scores are computed from the resulting graph, and the same paired query sets and answer-scoring procedures are used for the full and ablated variants. Results are reported separately for direct and related queries.

The hierarchical/flat comparison uses the same parsed chunks, stored metadata, context serializer, and retention settings ($b = 240$ and a 6,500-character cap). Flat retrieval ranks chunks across the collection by TF–IDF, aggregates their scores to select source papers and sections, and assembles the selected evidence in the same bundle format. Both modes use the same TF–IDF implementation. Metadata remains available during context assembly in both modes. The paired answering runs use the same Qwen2-VL-7B-Instruct checkpoint, temperature 0, a 700-token output limit, and concise `structured_v1` prompts. Corpus-expansion cost measurements replay retrieval and context preparation without answer generation.

The crop ablation compares two views of the same retrieved figure. The crop variant supplies a single enclosing crop of the selected windows, while the uncropped variant supplies the complete extracted figure asset at its native pixel resolution. Both views pass through the same Qwen2-VL image processor, which applies its image resizing and patch-merging procedure. The runs use the same answer checkpoint, prompt template, decoding settings, and one-image limit. Across all 30 queries, candidate papers, source figures, and text-context token counts match query by query.

F Complete Text Results

Table 12 reports Paper F1 and factual precision, recall, and F1 for the nine retention settings and ten text baselines. Table 13 gives the corresponding retention-setting costs. For the six retention settings and ten baselines in Table 1, Table 15 compares mean Input on the 24 factual queries with costs on the full 116-query suite. Each quality metric is macro-averaged over its evaluation subset, as defined in Appendix D. Input denotes the mean recorded context-token count, with preprocessing accounted for separately.

The $b = 320$ setting achieves 69.42 Fact F1 with 250.97 context tokens on average, compared with 71.67 Fact F1 and 565.21 tokens for uncapped retrieval. The highest Paper F1, 50.91, occurs at $b = 480$. The sweep thus characterizes how retention affects paper identification, factual answering, and input cost.

Table 12: Complete text-quality records. Paper F1 is evaluated on 116 queries, and factual scores on 24 queries with 72 facts. Scores are percentages.

Method / setting	Paper F1	Fact P	Fact R	Fact F1
RAG-SCHOLAR($b = 100$)	45.76	70.14	62.50	65.56
RAG-SCHOLAR($b = 140$)	45.47	71.53	63.89	66.94
RAG-SCHOLAR($b = 180$)	46.57	70.14	63.89	66.27
RAG-SCHOLAR($b = 240$)	48.79	69.10	65.33	66.00
RAG-SCHOLAR($b = 320$)	50.36	73.26	68.06	69.42
RAG-SCHOLAR($b = 480$)	50.91	71.53	61.11	64.44
RAG-SCHOLAR($b = 640$)	50.54	63.89	55.56	58.47
RAG-SCHOLAR($b = 800$)	49.21	68.75	65.28	65.89
RAG-SCHOLAR(uncapped)	49.37	74.11	70.83	71.67
CompAct *(EMNLP'24)*	39.86	62.15	50.00	53.31
ReComp *(ICLR'24)*	40.39	26.39	29.17	27.00
ColPali *(ICLR'25)*	31.84	39.31	58.33	45.07
M3DocRAG *(ICCVW'25)*	31.21	37.06	54.17	42.30
VisRAG *(ICLR'25)*	29.97	29.44	29.17	28.05
MDocAgent *(arXiv'25)*	36.67	32.91	37.50	33.19
LLMLingua-2 *(ACL Findings'24)*	36.78	19.44	11.11	13.19
LongLLMLingua *(ACL'24)*	30.75	30.30	31.94	28.60
Selective Context *(EMNLP'23)*	27.15	23.61	15.28	17.78
LLMLingua *(EMNLP'23)*	19.78	10.30	13.89	11.25

Table 13: Complete retention-setting costs, including $b = 140, 180, 800$. Pre. counts online preprocessing tokens; Input reports the mean context-token count recorded during evaluation; Total = Pre. + Input. Baseline costs are reported in Table 1. Generated output is excluded.

Method / setting	Pre.	Input	Total	Prep. (s)
RAG-SCHOLAR($b = 100$)	0	174.42	174.42	11.39
RAG-SCHOLAR($b = 140$)	0	182.86	182.86	11.38
RAG-SCHOLAR($b = 180$)	0	196.23	196.23	11.37
RAG-SCHOLAR($b = 240$)	0	218.20	218.20	11.38
RAG-SCHOLAR($b = 320$)	0	250.97	250.97	11.37
RAG-SCHOLAR($b = 480$)	0	321.36	321.36	11.35
RAG-SCHOLAR($b = 640$)	0	387.05	387.05	11.36
RAG-SCHOLAR($b = 800$)	0	447.53	447.53	11.40
RAG-SCHOLAR(uncapped)	0	565.21	565.21	11.80

G Complete Multimodal Results

Table 14 decomposes the factual scores in Table 2. RAG-Scholar achieves the highest Paper F1 and Fact F1 among the compared systems, together with the lowest text and visual input costs. Its factual precision is 70.00, compared with 61.11 for M3DocRAG, while recall is 53.33 versus 55.56. Evaluation subsets and the factual-scoring rule are defined in Appendix D.

Table 14: Full multimodal quality. Paper F1 uses 30 queries; factual metrics use the 15 figure-reasoning queries with 45 core facts.

Method	Paper F1	Fact P	Fact R	Fact F1
RAG-Scholar	100.00	70.00	53.33	60.00
ColPali *(ICLR'25)*	56.33	46.67	46.67	46.29
M3DocRAG *(ICCVW'25)*	62.44	61.11	55.56	56.92
VisRAG *(ICLR'25)*	44.67	26.67	24.44	25.33

Table 15: Mean token costs on the full text suite and its factual subset. The 24-query column uses the same questions as Fact F1. Pre. and Total are averaged over all 116 queries, with Total = Pre. + Input in that group.

	Full suite (116)			Factual subset (24)
Method / setting	Pre.	Input	Total	Input
RAG-Scholar($b = 100$)	0	174.42	174.42	125.42
RAG-Scholar($b = 240$)	0	218.20	218.20	209.29
RAG-Scholar($b = 320$)	0	250.97	250.97	273.58
RAG-Scholar($b = 480$)	0	321.36	321.36	457.42
RAG-Scholar($b = 640$)	0	387.05	387.05	622.63
RAG-Scholar(uncapped)	0	565.21	565.21	1082.08
CompAct	11026.36	154.96	11181.32	176.83
ReComp	0	244.28	244.28	248.38
ColPali	0	1854.68	1854.68	1399.54
M3DocRAG	0	1854.42	1854.42	1399.25
VisRAG	0	1854.46	1854.46	1399.46
MDocAgent	0	1838.84	1838.84	1399.33
LLMLingua-2	42729.27	2687.00	45416.27	2687.00
LongLLMLingua	24421.79	2304.91	26726.70	2335.17
Selective Context	14106.06	2500.00	16606.06	2500.00
LLMLingua	23596.27	2234.35	25830.62	2255.92

H Complete Corpus Expansion

The collections contain 146, 200, 250, 298, and 597 papers and share the same 116 text queries. Gold paper sets and reference facts are constructed separately for each collection to cover its full graph, including the added papers. Both retrieval modes are evaluated against the same gold annotations at each size, using the inference and scoring rules in Appendix D. Paper F1 uses all 116 queries: predicted papers in the corresponding gold set count as true positives, and other predicted papers count as false positives. Fact P/R/F1 use the fixed 24-query subset with the reference facts for the corresponding collection. Costs are measured by deterministic pre-generation replay at $b = 240$ with a 6,500-character cap. Context tokens include the serialized query and evidence; outer instructions, the output schema, chat framing, and generated answers are excluded.

The expanded collections are nested supersets of the 146-paper core, adding 54, 104, 152, and 451 papers, respectively. The expansion pool is drawn from arXiv searches for retrieval-augmented generation, large-language-model retrieval, graph-based question answering, and long-context language

models. Records retain versioned arXiv identifiers and are deduplicated against the original collection by source identifier, DOI, or normalized title. Candidates with downloadable PDFs are ordered by descending publication year and then title; papers with parsed content are appended in this order. Each selected paper retains its graph identifier across collection sizes.

At each size, hierarchical and flat retrieval use the same indexed paper, section, and chunk records. Flat retrieval applies the global chunk-ranking procedure in Appendix E.3, with a 120-chunk candidate pool and the same five-paper, two-section, and two-chunk retention limits. Both modes use Qwen2-VL-7B-Instruct with greedy decoding, concise answers, and a 320-token output limit for this sweep. Timing follows Appendix E: queries run sequentially after setup, loaded resources and in-process caches are reused, and every query contributes to the mean. Offline index construction and answer generation are excluded from preparation time.

Table 16 reports the quality measurements across the five corpus sizes. Hierarchical retrieval achieves higher Paper F1 and Fact F1 at every size. At 597 papers, it reaches 42.61 Paper F1 and 71.49 Fact F1, compared with 30.24 and 38.51 for flat retrieval. Preparation takes 39.38 seconds for hierarchical retrieval and 123.62 seconds for flat retrieval at this size. Figure 4 summarizes the quality and preparation-time trends.

Table 16: Corpus-expansion quality on the same 116 text queries, using gold annotations covering the full graph at each size. Fact P/R/F1 use the fixed 24-query subset with the corresponding collection's reference facts. All scores are percentages.

Papers	Retrieval	Paper F1	Fact P	Fact R	Fact F1
146	Hierarchical	48.8	69.1	65.3	66.0
146	Flat	33.2	63.2	56.9	58.8
200	Hierarchical	43.1	68.4	65.3	65.7
200	Flat	31.5	69.1	59.7	62.5
250	Hierarchical	43.3	67.4	65.3	65.1
250	Flat	31.0	64.9	55.6	58.3
298	Hierarchical	42.8	71.5	69.4	69.3
298	Flat	29.9	64.2	55.6	58.0
597	Hierarchical	42.6	76.0	69.4	71.5
597	Flat	30.2	41.7	37.5	38.5

I Complete Component Ablations

I.1 Entity-Node Ablations

Tables 17 and 18 report the full-index and removal variants for each of the six tested entity types. The removal procedure is defined in Appendix E.3. Direct queries target the named type, while related queries also include compositional constraints involving it; their type-specific memberships are given in Appendix A. Author and Reference each have the same query set in their direct and related scopes, which yields matching results across the two tables for each type.

On related queries, removing institution, method, dataset, or model nodes reduces Fact F1 from 83.52 to 35.00, 83.31 to 34.82, 90.80 to 47.39, and 86.73 to 42.38, respectively. In the same comparisons, the full index retains 458.9–558.3 context tokens on average, whereas each removal exceeds 1,070 tokens. Author nodes improve Fact Recall from 56.25 to 75.00, with Fact F1 increasing from 62.50 to 64.58 and context decreasing from 888.6 to 652.4 tokens. These results connect entity-guided retrieval to both factual coverage and compact evidence selection.

I.2 Visual Cropping

Table 19 decomposes the factual scores for the paired crop and no-crop variants in Table 5. Both variants retain the same text context, averaging 1,090.70 tokens, and the same source figure for each query. Cropping reduces mean visual input from 523.87 to 217.80 tokens while Paper F1 remains 100.00. Fact Precision changes from 75.56 to 70.00 and Fact Recall from 51.11 to 53.33, giving a

Table 17: Entity-node ablation on direct queries. Full retains the complete index; Removed excludes the named entity type under the protocol in Appendix E.3. Quality scores are percentages; Context is the mean recorded text-context token count. Factual scoring follows Appendix D.

Entity	Variant	Paper F1	Fact P	Fact R	Fact F1	Context
Author	Full	70.54	60.42	75.00	64.58	652.4
Author	Removed	46.73	75.00	56.25	62.50	888.6
Institution	Full	87.70	72.78	100.00	82.54	86.8
Institution	Removed	0.00	8.33	8.33	8.33	1056.0
Method	Full	45.37	75.78	97.88	84.68	116.8
Method	Removed	4.17	10.78	6.88	5.14	1048.8
Dataset	Full	51.20	78.39	100.00	86.82	102.6
Dataset	Removed	17.13	41.67	17.36	24.17	1112.9
Model	Full	43.65	75.48	93.75	81.22	127.2
Model	Removed	5.68	16.67	4.86	7.50	1109.0
Reference	Full	27.72	76.24	94.96	83.62	708.2
Reference	Removed	4.06	8.33	0.83	1.52	1105.9

Table 18: Entity-node ablation on related queries. Full retains the complete index; Removed excludes the named entity type under the protocol in Appendix E.3. Quality scores are percentages; Context is the mean recorded text-context token count. Factual scoring follows Appendix D.

Entity	Variant	Paper F1	Fact P	Fact R	Fact F1	Context
Author	Full	70.54	60.42	75.00	64.58	652.4
Author	Removed	46.73	75.00	56.25	62.50	888.6
Institution	Full	78.45	78.67	93.33	83.52	458.9
Institution	Removed	25.83	35.00	35.00	35.00	1074.8
Method	Full	47.08	77.69	91.65	83.31	485.1
Method	Removed	26.67	38.39	35.92	34.82	1073.6
Dataset	Full	62.22	87.28	97.10	90.80	558.3
Dataset	Removed	40.10	56.52	43.84	47.39	1089.6
Model	Full	61.45	84.40	93.25	86.73	525.9
Model	Removed	39.75	47.62	40.87	42.38	1080.2
Reference	Full	27.72	76.24	94.96	83.62	708.2
Reference	Removed	4.06	8.33	0.83	1.52	1105.9

0.22-point difference in Fact F1. The paired results show that cropping substantially reduces visual input while maintaining nearly the same aggregate factual F1 under the scoring rule in Appendix D.

Table 19: Visual-cropping ablation on the 15 figure-reasoning queries with 45 reference facts. Fact P, R, and F1 are macro-averaged separately and reported as percentages. Paper F1 and input costs appear in Table 5.

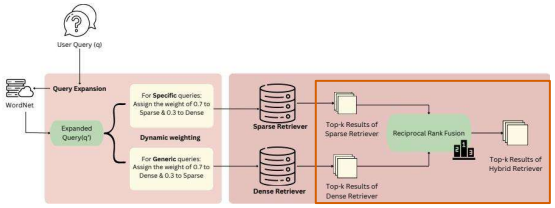
Variant	Fact P	Fact R	Fact F1
Visual selection: multimodal suite			
Crop	70.00	53.33	60.00
No crop	75.56	51.11	60.22

Figure 5 shows three visual-cropping examples from the multimodal suite. The crops occupy 25.5%, 42.8%, and 70.6% of the original figure areas, respectively. Visual token counts decrease from 384 to 98, 810 to 342, and 588 to 432, corresponding to reductions of 74.5%, 57.8%, and 26.5%.

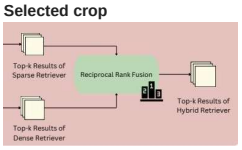
Hybrid Retriever Visual tokens reduced by 74.5%

Query: Show me the hybrid retriever pipeline figure for hallucination mitigation in large language models

Original figure



Selected crop



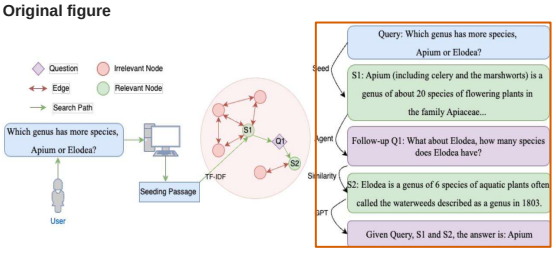
Visual tokens: 384
901 × 337 px
Area: 100.0%

Visual tokens: 98
385 × 201 px
Area kept: 25.5%

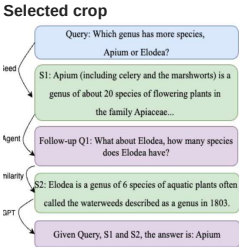
CuriousLLM Visual tokens reduced by 57.8%

Query: In the CuriousLLM workflow example, how does the generated follow-up question help seemingly unrelated passages become part of the path to the final answer?

Original figure



Selected crop



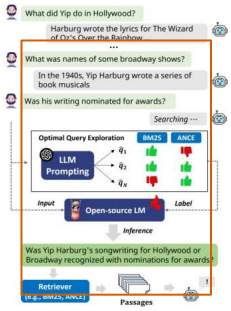
Visual tokens: 810
1265 × 518 px
Area: 100.0%

Visual tokens: 342
542 × 518 px
Area kept: 42.8%

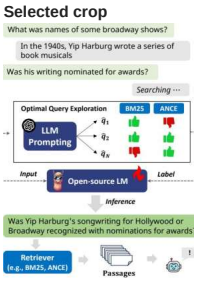
RETPO Visual tokens reduced by 26.5%

Query: Show me the RETPO overview figure for aligning LLM rewrites with retriever preferences in conversation

Original figure



Selected crop



Visual tokens: 588
587 × 793 px
Area: 100.0%

Visual tokens: 432
495 × 664 px
Area kept: 70.6%

Figure 5: Visual-cropping examples using source diagrams from (a) Hybrid Retrieval, (b) CuriousLLM, and (c) RETPO. Each query appears above its image pair. Left: the original figure with the enclosing crop outlined in orange. Right: the selected crop, displayed at the same size as the outlined region. Visual token counts after Qwen2-VL preprocessing and image dimensions appear below each image. "Area kept" denotes the fraction of the original figure's pixel area.