# BEYOND CHUNKS AND BINARY GRAPHS: MULTIMODAL HYPERGRAPH REPRESENTATION FOR LONG DOCUMENT QUESTION ANSWERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long document question answering often requires grounding answers in textual, tabular, and visual evidence distributed across multiple pages, especially for visually rich documents such as scientific papers and technical reports. Existing Retrieval-Augmented Generation (RAG) approaches typically retrieve text chunks, multimodal chunks, full-page images, or binary graph structures, but these retrieval units provide limited support for organizing heterogeneous evidence into coherent retrieval units. We introduce DOCHYPER, a multimodal document hypergraph representation in which document entities are connected by hyperedges that encode n-ary evidence units within individual elements and among heterogeneous elements in the surrounding document context. Built from parsed document elements, DOCHYPER supports entity- and hyperedge-level retrieval with page- and layout-aware source grounding. Experiments on LongDocURL and MMLongBench-Doc show that our approach improves answer quality over chunk-, multimodal-chunk-, page-image-, and graph-based retrieval baselines. Ablations further analyze entity-level, hyperedge-level, and binary-graph retrieval, providing evidence for the value of organizing retrieved multimodal evidence as higher-order units with source-page grounding.

## 1 INTRODUCTION

Long document question answering increasingly requires retrieving evidence from visually rich documents such as scientific papers and technical reports and grounding answers in it Ma et al. (2024); Deng et al. (2025). Unlike short text passages, these documents often contain heterogeneous elements—textual passages, tables, captions, and figures, together with layout structures—distributed across multiple pages. Answering a question may require jointly using a textual statement, a table row, a figure region, and surrounding context, sometimes within one coherent evidence unit and sometimes across multiple retrieved pages. Therefore, the challenge is not only to retrieve relevant pages or passages, but also to preserve relationships among heterogeneous elements and to organize multiple page-grounded evidence units for answer generation.

Existing Retrieval-Augmented Generation (RAG) approaches commonly rely on retrieval units such as text chunks, multimodal chunks, coarse page images, or binary graph edges Lewis et al. (2020); Gao et al. (2023). Text-chunk-based methods split documents into text chunks, while multimodal-chunk-based methods augment chunk retrieval with visual crops, captions, or LVLM-generated descriptions. These chunk-based retrieval units organize evidence as separate chunks, which can fragment coherent evidence and make multi-page evidence aggregation difficult. Page-image retrieval methods such as M3DocRAG Cho et al. (2024) retrieve full-page images, preserving visual context but introducing substantial irrelevant content and leaving the model to rediscover fine-grained evidence relationships at generation time. Graph-based RAG approaches improve over flat chunks by modeling entities and binary relations, but representing a higher-order evidence unit as multiple binary edges can fragment the joint context among heterogeneous elements Peng et al. (2026); Edge et al. (2024). For example, a coherent evidence unit may involve a table, its caption, and a nearby explanatory paragraph.



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

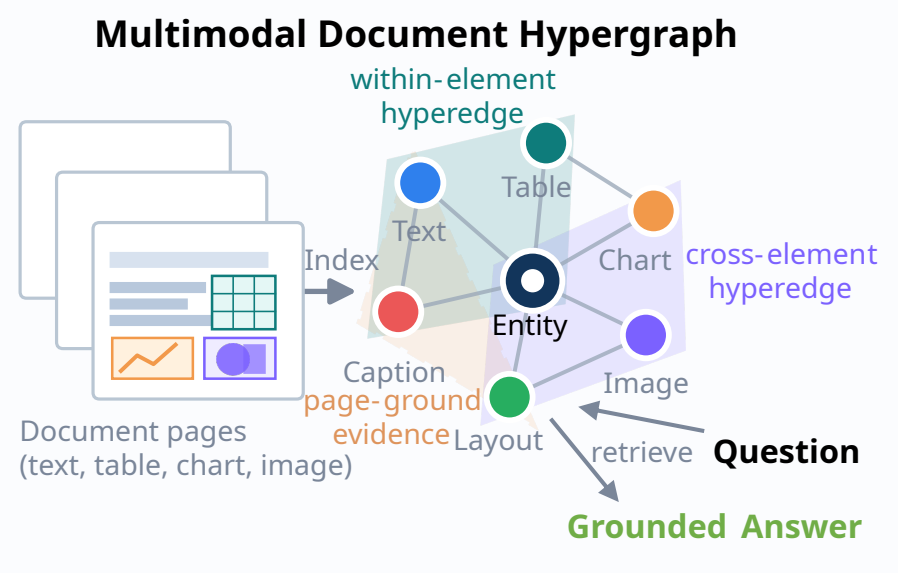



Figure 1: Illustration of DocHyper.

We argue that hypergraphs provide a useful representation for structured evidence units in long multimodal documents. A hypergraph generalizes a graph by allowing one hyperedge to connect more than two nodes, making it suitable for encoding n-ary evidence units among mutually related document entities Liu et al. (2020); Rosso et al. (2020). This is particularly relevant to multimodal documents, where meaningful relationships may arise within an individual element, such as entities co-occurring in a table or figure, or among heterogeneous elements in the surrounding document context, such as a figure, its caption, and surrounding textual analysis. While recent hypergraph-based RAG methods have explored n-ary relationships over text-centric knowledge graphs, they are not designed to organize multimodal document evidence with page- and layout-aware source contexts Luo et al. (2026); Feng et al. (2025). This leaves a gap in representing page-grounded higher-order relationships among heterogeneous document elements.

To bridge this gap, we propose DocHyper, a multimodal document hypergraph representation for long document question answering, as illustrated in Figure 1. DocHyper treats document entities as nodes and constructs hyperedges for text blocks, table semantics, image events, chart trends, and same-page cross-element evidence. Each entity and hyperedge retains modality, source-element, page, and layout metadata, enabling retrieved evidence to be traced back to its document context. For questions involving multiple pages, the retriever may return several page-grounded evidence units, which are jointly provided to the answer generator. We further design an automated construction pipeline that builds such indices from parsed document elements using specialized processors for text, tables, images, and charts. The index maintains an entity–hyperedge incidence bipartite graph, allowing retrieval from both fine-grained entities and higher-order evidence units before tracing selected evidence back to source pages for multimodal answer generation.

We evaluate our approach on LongDocURL and MMLongBench-Doc, comparing it with text-chunk RAG, multimodal-chunk RAG, binary graph RAG, page-image retrieval methods such as M3DocRAG, and agentic long document QA methods such as MDocAgent. As shown in Table 1, our hypergraph-based retrieval improves both page-level evidence recall and answer quality. On LongDocURL, our method achieves the best answer accuracy and F1 among compared methods while also improving page-level evidence recall. On MMLongBench-Doc, it outperforms the compared chunk-, multimodal-chunk-, graph-, and page-image retrieval baselines in answer quality. These results indicate that organizing higher-order evidence units with source-page grounding is effective across two long document QA benchmarks. We further conduct graph-structure ablations, evidence-organization analysis, and fine-grained evidence analysis to examine the roles of entity-level retrieval, hyperedge-level retrieval, higher-order evidence units, and the combination of multiple page-grounded evidence units for multi-page questions.

Our contributions are summarized as follows:

- We propose DOCHYPER, a multimodal document hypergraph representation for long document question answering, organizing textual, tabular, and visual evidence into within-element and cross-element higher-order evidence units while preserving page and layout metadata for source grounding.
- We design an automated and reproducible pipeline for constructing multimodal hypergraph indices from parsed document elements, organized as an entity–hyperedge incidence bipartite graph for entity- and hyperedge-level retrieval.
- We provide systematic experiments on LongDocURL and MMLongBench-Doc against chunk-, multimodal-chunk-, page-image-, and graph-based retrieval baselines, with ablations and analyses examining entity-level retrieval, hyperedge-level retrieval, binary-graph retrieval, evidence organization, and multi-page evidence combination.

## 2 RELATED WORK

### 2.1 LONG AND MULTIMODAL DOCUMENT QA

Document Intelligence increasingly relies on LLMs/LVLMs Meta (2024); Alibaba (2024) for answer generation and multimodal understanding. LLM-based systems improve reasoning over retrieved textual context, while LVLMs enable question answering over visually rich document pages. For long visually rich PDFs, however, relevant evidence may be sparse, distributed across pages. RAG methods select evidence before generation, with chunk-based and multimodal-chunk-based methods retrieving local text chunks or passages and augmenting them with image crops, captions, or LVLM-generated descriptions. Page-image retrieval directly retrieves rendered pages to preserve layout and visual context, and agentic pipelines further iterate over search, reading, verification, and answer refinement Cho et al. (2024); Yu et al. (2025); Han et al. (2025). These approaches improve evidence coverage in different ways, but their retrieved units are usually independent chunks, multimodal chunks, passages, or pages, leaving cross-element evidence organization to the downstream generator.

### 2.2 GRAPH AND HYPERGRAPH RETRIEVAL

Graph-based RAG represents external knowledge or document content as entities, relations, and communities, enabling structured retrieval beyond independent text chunks Edge et al. (2024); Wang et al. (2025); Liang et al. (2025); Zhang & Wu (2026). GraphRAG constructs entity knowledge graphs and community summaries for global queries, while HippoRAG Gutierrez et al. (2024) and LightRAG Guo et al. (2025) respectively employ graph propagation and dual-level retrieval to associate distributed information. However, these approaches primarily organize textual knowledge through binary relations, requiring multi-entity evidence to be recovered through relation composition or graph traversal.

Recent hypergraph-based RAG methods model higher-order associations directly. HyperGraphRAG Luo et al. (2026) represents n-ary relational facts as hyperedges, Hyper-RAG Feng et al. (2025) captures pairwise and beyond-pairwise correlations in domain knowledge, and HGRAG Wang et al. (2026) treats passages as hyperedges to integrate entity- and passage-level retrieval. Nevertheless, these methods remain primarily text-centric or domain-knowledge-oriented, with limited treatment of multimodal document elements, page provenance, and layout metadata. Our work instead studies hypergraph retrieval for visually rich documents, while keeping textual, tabular, and visual evidence traceable to its original document context.

## 3 METHOD

### 3.1 PROBLEM FORMULATION

We consider question answering over a long multimodal document $D$. Each document provides two complementary inputs: a sequence of parsed document elements $X_D = \{x_i\}_{i=1}^{N}$ and a set of page

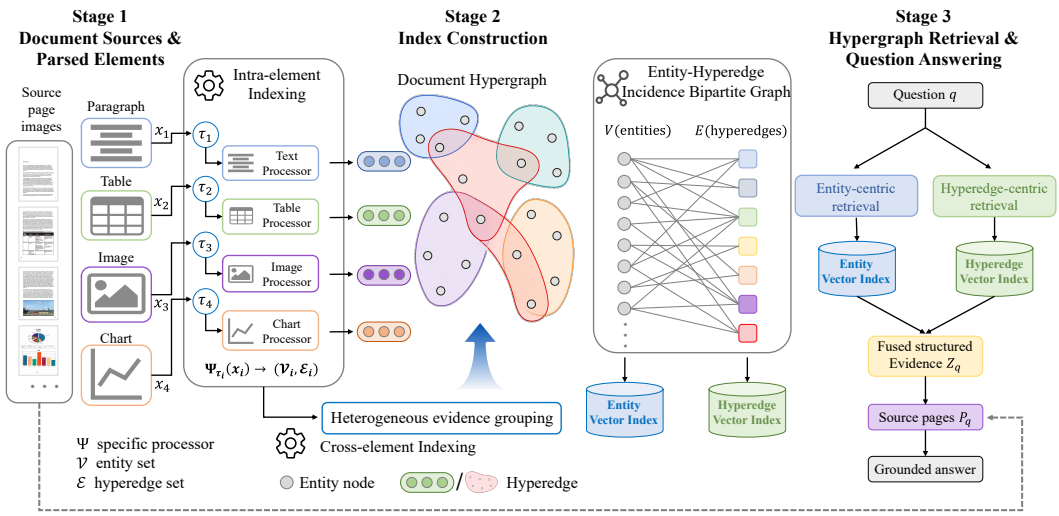



Figure 2: Overview of DocHyper. Parsed document elements $x_i$ are processed by type-specific processors $\Psi_{\tau_i}$ for intra-element indexing, producing entities and element-level hyperedges. Cross-element indexing performs same-page heterogeneous evidence grouping over representative entities from different source types.

images $I_D = \{I_p\}_{p=1}^{P}$. In our implementation, the document elements are obtained from MinerU parsing Wang et al. (2024). Each element $x_i$ contains content $c_i$, an element type $\tau_i$ (e.g., text, table, image, or chart), a page index $p_i$, layout metadata such as a bounding box $b_i$, and a source identifier. Given a question $q$, the goal is to produce an answer $a$ grounded in evidence from $D$.

DocHyper constructs a modality-aware hypergraph from the parsed multimodal elements. The hypergraph retrieval space is defined over document entities and retrievable evidence units derived from textual, tabular, and visual elements, while page images are retained as source-level visual evidence for answer generation rather than treated as graph nodes. Formally, for each document $D$, we build a document hypergraph $\mathcal{H}_D = (\mathcal{V}_D, \mathcal{E}_D)$, where $\mathcal{V}_D$ denotes document entity nodes and $\mathcal{E}_D$ denotes hyperedges connecting one or more related entities. At query time, a retriever selects a structured evidence set $Z_q \subseteq \mathcal{V}_D \cup \mathcal{E}_D$ and maps it to a source page set $P_q$. In the main setting, the corresponding page images $\{I_p \mid p \in P_q\}$ are used as the QA context for grounded answer generation.

### 3.2 Multimodal Document Hypergraph

We define a multimodal document hypergraph as a structured representation that organizes document evidence into entity nodes and attributed hyperedges. Formally, for a document $D$, we construct $\mathcal{H}_D = (\mathcal{V}_D, \mathcal{E}_D)$. Each node $v \in \mathcal{V}_D$ denotes a fine-grained, source-local document entity extracted from textual, tabular, or visual evidence. In our implementation, an entity identity is tied to its surface form, semantic category, source element, and page, so the graph records grounded entity occurrences rather than assuming global entity canonicalization across the document.

Each hyperedge is represented as $e = (S_e, r_e, d_e, m_e)$, where $S_e \subseteq \mathcal{V}_D$ is its incident entity set, $r_e$ is the hyperedge class, $d_e$ is a retrievable semantic representation of the evidence context, and $m_e$ contains source, page, and layout metadata. The representation $d_e$ is textual for text elements and textualized for tabular and visual elements, while the metadata $m_e$ preserves the path back to the original document source. Hyperedges with $|S_e| > 1$ encode n-ary evidence units; singleton hyperedges are retained when an element yields only one extracted entity, but they serve as evidence units rather than higher-order relationships.

Page and layout metadata are used for source grounding rather than treated as the same type of evidence as text, tables, or figures. This design separates three levels: entities provide fine-grained semantic anchors, hyperedges provide retrievable evidence contexts, and source pages provide visual grounding.

Our implementation defines 21 semantic entity types spanning textual, tabular, visual, chart-specific, and reference entities. On the hyperedge side, the current index instantiates five classes of retrievable evidence units. Text-block hyperedges use the textual content, or its shortened summary when needed, as the evidence context. Table-semantic hyperedges store a semantic description of a table together with its Markdown representation when available. Image-event hyperedges summarize scene descriptions, events, key visual information, and captions from images. Chart-trend hyperedges describe chart types, axes, series, and salient trends or observations. Cross-element hyperedges group representative entities from multiple source types on the same page, forming a page-level heterogeneous evidence context. The first four classes mainly capture within-element evidence, while cross-element hyperedges organize evidence from different document elements.

This representation differs from both chunk-based and binary-graph representations. A chunk stores local content without explicitly representing the membership between entities and evidence contexts. In DOCHYPER, the same entity can participate in multiple hyperedges, $v \in e_1$ and $v \in e_2$, which allows retrieval to move between a fine-grained entity and the different evidence contexts in which it appears. A binary graph decomposes a higher-order evidence unit into multiple binary links, which can fragment the joint context among related evidence. In contrast, a document hypergraph preserves the joint membership of multiple entities in a shared evidence context, enabling the retriever to access that context without decomposing it into independent binary links.

### 3.3 HYPERGRAPH INDEX CONSTRUCTION

As illustrated in Figure 2, the construction pipeline converts parsed document elements into a hypergraph retrieval index. During index construction, semantic extraction and summarization mainly use Qwen-Max Yang et al. (2025) for text-centric elements and Qwen-VL-Max Bai et al. (2025) for visual elements. For each element $x_i$, we dispatch it to an element-specific processor $\Psi_{\tau_i}$ according to its element type:

$$\Psi_{\tau_i}(x_i) \to (\mathcal{V}_i, \mathcal{E}_i), \tag{1}$$

where $\mathcal{V}_i$ contains the entities extracted from the element and $\mathcal{E}_i$ contains its element-level hyperedges. Text processors extract keywords, concepts, named entities, dates, and numerical values from a text block, and use the block itself or its shortened summary as a text-block hyperedge. Table processors extract table references, column headers, and row labels, use MLLM-recognized structure or parsed Markdown when available, and represent the whole table with a table-semantic hyperedge containing a natural-language description and its Markdown form. Image processors extract figure references and visual entities such as objects, persons, scenes, or text, and build image-event hyperedges from scene descriptions, event descriptions, key information, and captions. Chart processors extract chart references, axis labels and units, data series, legends, and chart types, and build chart-trend hyperedges that summarize trends, key observations, and salient data patterns. Each generated entity and hyperedge keeps source metadata, including the source element identifier, page index, and layout information when available.

After processing individual elements, cross-element indexing performs same-page heterogeneous evidence grouping. Entities are grouped by page index and source type. For each page containing entities from at least two source types, entities within each source type are sorted by extraction confidence. We then select up to $K_c$ representative entities by round-robin sampling across source types, so that no single source type dominates the cross-element hyperedge. The selected entities are connected with a cross-element hyperedge whose semantic representation summarizes the page-level heterogeneous context and whose metadata records the involved source types and entity count. This step adds explicit higher-order evidence units derived from same-page multimodal co-occurrence, without assuming a stronger cross-page semantic relationship.

The document-level hypergraph is the union of all element-level and cross-element outputs:

$$\mathcal{V}_D = \bigcup_i \mathcal{V}_i, \qquad \mathcal{E}_D = \left( \bigcup_i \mathcal{E}_i \right) \cup \mathcal{E}_D^{\text{cross}}. \tag{2}$$

Besides maintaining separate dictionaries for entities and hyperedges, the index builds an entity–hyperedge incidence bipartite graph $\mathcal{B}_D$: an entity node is connected to every hyperedge that contains it. This incidence graph is an indexing view of the hypergraph, not a replacement for the hypergraph representation. It enables efficient traversal from a matched entity to its associated evidence units and from a matched hyperedge to its connected entities. Finally, entity and hyperedge

embeddings are indexed separately, allowing the retriever to match queries against both fine-grained entity semantics and complete hyperedge evidence contexts.

### 3.4 HYPERGRAPH RETRIEVAL AND GROUNDED QA

At query time, DOCHYPER retrieves evidence from the hypergraph index through a dual-path strategy. We first extract two keyword sets from the question, $\phi_{\text{low}}(q)$ and $\phi_{\text{high}}(q)$. The low-level keywords target fine-grained entity matching, while the high-level keywords target broader evidence-unit matching.

The first path is entity-centric retrieval. Low-level keywords are embedded and matched against entity embeddings by cosine similarity. The top-$K_V$ entities are then expanded through the incidence graph to collect their incident hyperedges, and each expanded hyperedge inherits the highest score among its matched entities. This path recovers evidence anchored by specific terms, objects, table cells, or visual entities.

The second path is hyperedge-centric retrieval. High-level keywords are embedded and matched directly against hyperedge representations. The top-$K_E$ hyperedges are expanded to their incident entities, and each expanded entity inherits the highest score among the matched hyperedges that contain it. This path directly retrieves n-ary evidence units, such as table-level semantics, image or chart descriptions, and page-level heterogeneous contexts.

The two paths are fused by deduplicating entities and hyperedges by their identifiers and keeping the maximum available score for each item:

$$s(z) = \max_{t \in \mathcal{T}(z)} s_t(z), \quad z \in \mathcal{V}_D \cup \mathcal{E}_D, \tag{3}$$

where $\mathcal{T}(z)$ denotes the direct or expansion paths that retrieved $z$. The fused entities and hyperedges form the structured evidence set $Z_q$. We then trace each selected item to its source page through metadata. Pages are ranked by the highest score of evidence items mapped to them, and the corresponding page images are used as the main QA context. Textualized entity or hyperedge evidence can be additionally attached in the evidence-organization variants analyzed in the appendix.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

We evaluate DOCHYPER on LongDocURL Deng et al. (2025) and MMLongBench-Doc Ma et al. (2024), two long-document QA benchmarks over visually rich PDFs. For MMLongBench-Doc, we filter 244 unanswerable records from the 1,091 raw questions because these records almost always lack evidence-page annotations, while non-empty evidence pages are required for retrieval metrics, leaving 847 answerable questions.

**Baselines** TextRAG retrieves text chunks; MMRAG retrieves text chunks together with textualized descriptions of visual crops, following common chunk-based and multimodal-chunk-based RAG settings Lewis et al. (2020); Gao et al. (2023); M3DocRAG Faysse et al. (2025); Cho et al. (2024) retrieves page images directly; GraphRAG retrieves graph-structured entity/relation evidence; and MDocAgent Han et al. (2025) retrieves parsed text passages together with page images. These methods instantiate the chunk-based, multimodal-chunk-based, page-image-based, graph-based, and agentic passage-and-page settings discussed in the introduction. Additional baseline comparisons are maintained in Appendix C.1.

Unless otherwise specified, DOCHYPER uses entities and hyperedges as retrieval units, traces the selected units to source pages, and uses the corresponding page images as the QA context. Across all methods, MinerU is used whenever parsed document elements are required, and Qwen-Max or Qwen-VL-Max is used whenever index construction requires semantic extraction or summarization. We use Qwen2-VL-7B-Instruct Alibaba (2024) as the answer-generation backbone for all locally evaluated methods, and use Qwen-Turbo for short-answer extraction before metric computation.

**Metrics** We report source-page recall ($R_{\text{page}}$), average retrieved or traced pages ($Pg.$), normalized answer correctness (Acc), and soft answer-level $F_1$ in the main results. Acc and $F_1$ follow

Table 1: Main results on LongDocURL and MMLongBench-Doc. Retrieval Unit denotes matched retrieval units; QA Context Unit denotes the context provided to the answer generator. Abbreviations: PI: page images, TC: text chunks, VC: visual crops, E: entities, R: binary relations, HE: hyperedges, TP: text passages. $R_{\text{page}}$ denotes macro source-page recall; $Pg.$ denotes the average number of unique retrieved or traced source pages per question. $Acc$ measures average normalized answer correctness, and $F_1$ measures soft answer-level precision–recall balance.

| Method | Retrieval Unit | Retrieval Top-K | QA Context Unit | LongDocURL | | | MMLongBench-Doc | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $R_{\text{page}}$ / $Pg.$ | $Acc$ | $F_1$ | $R_{\text{page}}$ / $Pg.$ | $Acc$ | $F_1$ |
| **LVLMs** | | | | | | | | | |
| LVLM | - | - | PI | - | 38.7 | 39.1 | - | 26.7 | 27.1 |
| **RAG / Agentic Pipeline** | | | | | | | | | |
| TextRAG | TC | 5 | TC | 39.2/3.6 | 34.3 | 35.8 | 26.4/3.4 | 19.0 | 20.8 |
| | | 10 | | 46.2/6.2 | 35.5 | 36.7 | 32.5/5.6 | 20.4 | 22.0 |
| MMRAG | TC+VC | (5, 5) | TC+VC | 52.6/6.3 | 39.3 | 40.4 | 43.9/6.3 | 28.8 | 30.3 |
| | | (10, 5) | | 68.2/8.4 | 39.9 | 41.0 | 58.4/8.1 | <u>29.5</u> | <u>31.0</u> |
| | | (5, 5) | TC+PI | 52.6/6.3 | 45.4 | 46.2 | 43.9/6.3 | 28.0 | 28.8 |
| | | (10, 5) | | 68.2/8.4 | 46.0 | 46.7 | 58.4/8.1 | 26.5 | 27.5 |
| GraphRAG | E | 5 | E+R+TC | 35.7/4.7 | 33.9 | 35.1 | 28.7/4.0 | 17.6 | 19.6 |
| | | 10 | | 35.6/6.0 | 33.8 | 34.9 | 32.9/5.1 | 18.0 | 20.0 |
| | E+R | (5, 5) | | 39.2/5.3 | 34.8 | 35.7 | 32.4/4.5 | 18.2 | 19.9 |
| | | (10, 10) | | 39.6/6.2 | 34.7 | 35.9 | 35.0/5.3 | 19.1 | 21.1 |
| M3DocRAG | PI | 5 | PI | 40.6/5.0 | 35.5 | 36.0 | 35.6/5.0 | 23.9 | 24.2 |
| | | 10 | | 60.7/10.0 | 41.2 | 41.8 | 59.2/10.0 | 29.1 | 29.3 |
| MDocAgent | TP+PI | (5, 5) | TP+PI | 77.0/8.3 | 45.0 | 45.6 | 64.0/8.5 | 27.3 | 28.7 |
| | | (10, 10) | | 87.2/14.9 | <u>46.6</u> | <u>47.2</u> | 76.0/15.2 | 27.1 | 28.0 |
| **Ours** | | | | | | | | | |
| DocHyper | E+HE | (5, 5) | PI | 72.8/6.2 | **49.6** | **50.2** | 66.2/6.5 | **33.0** | **33.2** |

the evaluation definitions used in LongDocURL and MMLongBench-Doc. Fine-grained analyses further group questions by evidence source and page span when annotations are available.

### 4.2 Main Results

**Overall QA performance.** Table 1 shows that DocHyper achieves the best normalized answer correctness and soft answer-level $F_1$ on both benchmarks. The improvement is consistent across LongDocURL and MMLongBench-Doc, while using a moderate source-page budget of 6.2 and 6.5 pages per question on average.

**Graph-based retrieval comparison.** GraphRAG and DocHyper are both graph-based retrieval methods, but their behavior differs substantially. In our GraphRAG setting, we evaluate both entity-only retrieval and joint entity–relation retrieval, with retrieved entity, relation, and text-chunk evidence provided to the answer generator. Even with direct relation retrieval, GraphRAG remains organized around binary relations rather than intact higher-order evidence units, and its source-page recall and QA scores remain substantially below DocHyper. This suggests that long multimodal document QA benefits not only from retrieving graph-structured evidence, but also from preserving higher-order evidence contexts and tracing them to compact source-page inputs.

**Retrieval–answering trade-off.** The main results show that increasing macro source-page recall or context size alone does not guarantee better answer quality. MDocAgent obtains the highest source-page recall by retrieving substantially more pages, yet its final QA gains are limited; on MMLongBench-Doc, it even underperforms MMRAG despite using a much larger page budget. This suggests that broad page coverage can introduce additional visual context without sufficiently organizing the evidence needed by the answer generator. In contrast, DocHyper first selects entities and hyperedges and then traces them to source pages, producing a more compact evidence set for answer generation.

**Retrieval confidence intervals.** Figure 3 further separates evidence coverage from retrieval compactness on LongDocURL. DocHyper achieves the highest source-page precision among the plot-

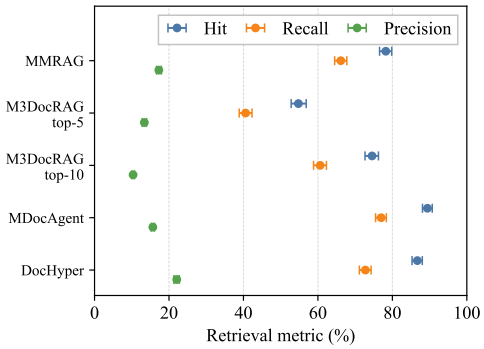



Figure 3: LongDocURL source-page retrieval performance with 95% bootstrap confidence intervals from 1,000 resamples.

Table 2: Ablation of graph structure and embedding backbone.

| Ablation | LongDocURL | | MMLongBench-Doc | |
|---|---|---|---|---|
| | Acc | $F_1$ | Acc | $F_1$ |
| DocHyper (default) | 49.6 | 50.2 | 33.0 | 33.2 |
| **Graph Structure** | | | | |
| Entities only | 44.6 | 45.1 | 27.9 | 28.2 |
| Hyperedges only | 43.2 | 43.7 | 29.6 | 30.0 |
| Entities + Pairwise edges | 44.8 | 45.3 | 27.8 | 28.2 |
| **Embedding Backbone** | | | | |
| BAAI/bge-m3 | 49.0 | 49.8 | 32.5 | 32.7 |
| Qwen3-Embedding-0.6B | 47.4 | 48.1 | 30.5 | 30.7 |
| Qwen3-Embedding-4B | 47.2 | 47.9 | 31.6 | 31.9 |

ted methods while maintaining competitive Hit and Recall. This supports the observation that structured evidence selection helps concentrate retrieved pages around question-relevant evidence.

### 4.3 Graph Structure Ablation

**Ablation settings.** The first row of Table 2 reports the default DocHyper setting, which uses all-MiniLM-L6-v2 Wang et al. (2020) embeddings and retrieves both entities and hyperedges. The graph-structure variants test this design by using only entities ("Entities only"), using only hyperedges ("Hyperedges only"), or replacing n-ary hyperedges with one-hop pairwise entity–entity edges ("Entities + Pairwise edges"). The embedding-backbone variants keep the default retrieval structure and replace only the embedding model.

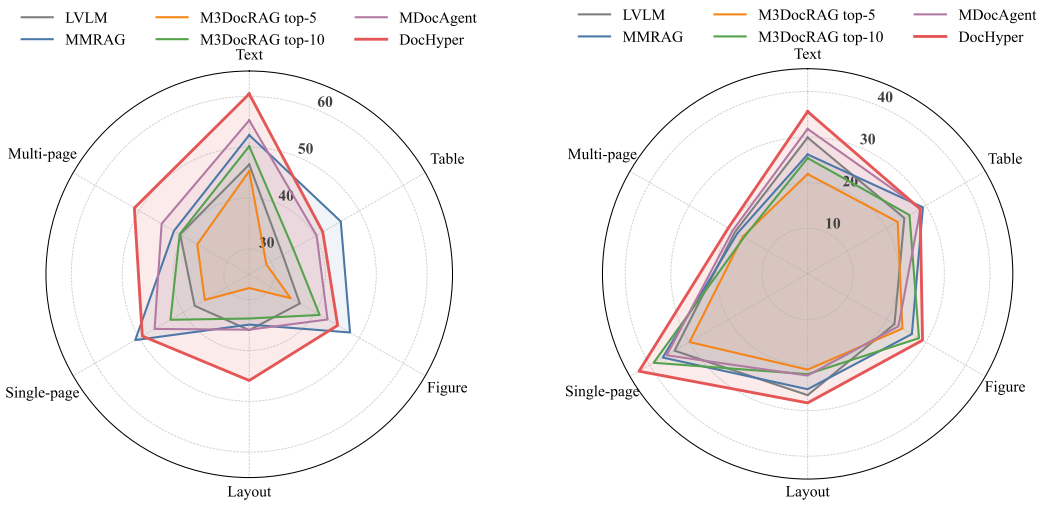
**Contribution of entity–hyperedge retrieval.** Table 2 shows that both sides of the incidence structure contribute to performance. Using "Entities only" leaves retrieval dominated by fine-grained entity matches without higher-order contexts, while using "Hyperedges only" weakens the anchors used to reach relevant evidence units. Both variants reduce answer quality on the two benchmarks, indicating that entities and hyperedges play complementary roles. Using "Entities + Pairwise edges" also degrades performance, suggesting that the benefit is not merely from adding graph connectivity, but from preserving higher-order evidence units that can be retrieved as intact units rather than fragmented pairs.

**Embedding backbone robustness.** We further replace the default all-MiniLM-L6-v2 embedding model with substantially larger embedding backbones, including BAAI/bge-m3 Chen et al. (2024), Qwen3-Embedding-0.6B, and Qwen3-Embedding-4B Zhang et al. (2025). These variants test whether larger embedding capacity alone can improve evidence retrieval under the same entity–hyperedge structure. None of the larger alternatives improves over the default setting, indicating that the main bottleneck is not simply embedding model scale; the organization of retrievable evidence units remains critical.

### 4.4 Fine-grained Evidence Analysis

**Fine-grained evidence categories.** Figure 4 shows that the aggregate gains in Table 1 are not uniform across evidence types. On LongDocURL, DocHyper is strongest on text, layout, and multi-page questions, while MMRAG, M3DocRAG, and MDocAgent remain competitive on more local table, figure, and single-page cases. On MMLongBench-Doc, DocHyper gives the strongest or most competitive profile on text, figure, layout, and page-span groups, but table-heavy questions remain close across TextRAG, MMRAG, MDocAgent, and DocHyper.

The layout axis deserves particular attention because it is closely tied to the quality of parsed document structure. MDocAgent retrieves text passages derived from MinerU-parsed document elements, and DocHyper constructs entities and hyperedges from the same parsed-document sub-




(a) LongDocURL. (b) MMLongBench-Doc.

Figure 4: Fine-grained QA $F_1$ radar charts on LongDocURL and MMLongBench-Doc. Each axis reports QA $F_1$ on the subset of questions associated with the corresponding evidence category or page-span group. The Figure axis corresponds to Figure evidence for LongDocURL and Figure $\cup$ Chart evidence for MMLongBench-Doc; Layout refers to generalized-text/layout evidence annotations; Single-page and Multi-page are determined by the number of annotated evidence pages.

strate. However, MDocAgent remains substantially weaker on layout/generalized-text questions, especially on LongDocURL. This suggests that high-quality document parsing is a useful foundation, but the additional benefit comes from reorganizing parsed elements into retrievable entity–hyperedge evidence units rather than treating them only as independent passages.

**Summary.** Across aggregate results, retrieval confidence intervals, ablations, and fine-grained analyses, the consistent pattern is that DocHyper benefits from retrieving structured evidence first and using it to ground a compact set of source pages. This avoids the two extremes of purely local chunk retrieval, which misses cross-element context, and broad page-image retrieval, which increases visual context size without explicitly organizing the evidence.

Due to space limits, additional method and experiment details are provided in the appendix, including prompts for entity extraction, hyperedge extraction, and question answering; final QA context formats used by baselines and DocHyper; index statistics and entity-type tables; index-construction cost analysis; and qualitative analysis results.

## 5 Conclusion

We present DocHyper, a multimodal document hypergraph representation for long-document question answering. By constructing entity and hyperedge evidence units from parsed document elements and tracing retrieved evidence back to source pages, DocHyper connects structured retrieval with grounded multimodal answer generation. Experiments on LongDocURL and MMLongBench-Doc show that organizing evidence as higher-order retrievable units improves answer quality over chunk-, multimodal-chunk-, page-image-, graph-based, and agentic passage-and-page baselines. The ablations and fine-grained analyses further highlight the complementary roles of entity-level retrieval, hyperedge-level retrieval, and source-page grounding, suggesting that long multimodal document QA benefits not only from retrieving relevant content, but also from preserving the evidence structure needed to use it.

### Acknowledgments

xxx.

### AI USE STATEMENT

In this work, we have not used generative AI tools for tasks requiring disclosure, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims or proofs, refining hypotheses, designing research methodology or experiments, implementing methods, translating text, cleaning or reformatting datasets, conducting qualitative or thematic data analysis, or interpreting results. The remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability. We reviewed and revised all AI-assisted text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

Alibaba. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024.

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

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216, 2024. doi: 10.48550/ARXIV.2402.03216. URL https://doi.org/10.48550/arXiv.2402.03216.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *CoRR*, abs/2411.04952, 2024. doi: 10.48550/ARXIV.2411.04952. URL https://doi.org/10.48550/arXiv.2411.04952.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 1135–1159. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.57. URL https://doi.org/10.18653/v1/2025.acl-long.57.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130, 2024. doi: 10.48550/ARXIV.2404.16130. URL https://doi.org/10.48550/arXiv.2404.16130.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Yifan Feng, Hao Hu, Xingliang Hou, Shiquan Liu, Shihui Ying, Shaoyi Du, Han Hu, and Yue Gao. Hyper-rag: Combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *arXiv preprint arXiv:2504.08758*, 2025.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023. doi: 10.48550/ARXIV.2312.10997. URL https://doi.org/10.48550/arXiv.2312.10997.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 10746–10761. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.568. URL https://doi.org/10.18653/v1/2025.findings-emnlp.568.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 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. *CoRR*, abs/2503.13964, 2025. doi: 10.48550/ARXIV.2503.13964. URL https://doi.org/10.48550/arXiv.2503.13964.

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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, Wen Zhang, and Huajun Chen. KAG: boosting llms in professional domains via knowledge augmented generation. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pp. 334–343. ACM, 2025. doi: 10.1145/3701716.3715240. URL https://doi.org/10.1145/3701716.3715240.

Yu Liu, Quanming Yao, and Yong Li. Generalizing tensor decomposition for n-ary relational knowledge bases. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (eds.), *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pp. 1104–1114. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380188. URL https://doi.org/10.1145/3366423.3380188.

Haoran Luo, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, et al. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *Advances in Neural Information Processing Systems*, 38:152206–152234, 2026.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffea0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html.

Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024. URL https://ai.meta.com/blog/meta-llama-3/.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *ACM Trans. Inf. Syst.*, 44(2): 35:1–35:52, 2026. doi: 10.1145/3777378. URL https://doi.org/10.1145/3777378.

Paolo Rosso, Dingqi Yang, and Philippe Cudré-Mauroux. Beyond triplets: Hyper-relational knowledge graph embedding for link prediction. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (eds.), *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pp. 1885–1896. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380257. URL https://doi.org/10.1145/3366423.3380257.

B Wang, C Xu, X Zhao, L Ouyang, F Wu, Z Zhao, R Xu, K Liu, Y Qu, F Shang, et al. An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

Changjian Wang, Weihong Deng, Weili Guan, Quan Lu, and Ning Jiang. Cross-granularity hypergraph retrieval-augmented generation for multi-hop question answering. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 33368–33376. AAAI Press, 2026. doi: 10.1609/AAAI.V40I39.40623. URL https://doi.org/10.1609/aaai.v40i39.40623.

Jinyu Wang, Jingjing Fu, Rui Wang, Lei Song, and Jiang Bian. PIKE-RAG: specialized knowledge and rationale augmented generation. *CoRR*, abs/2501.11551, 2025. doi: 10.48550/ARXIV.2501.11551. URL https://doi.org/10.48550/arXiv.2501.11551.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=zG459X3Xge.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yongyue Zhang and Yaxiong Wu. Mldocrag: Multimodal long-context document retrieval augmented generation. *CoRR*, abs/2602.10271, 2026. doi: 10.48550/ARXIV.2602.10271. URL https://doi.org/10.48550/arXiv.2602.10271.

# A METHOD DETAILS

## A.1 ENTITY CATEGORY

DOCHYPER extracts document entities from parsed text, tables, images, and charts. Table 3 lists the entity categories used in the current implementation. The implementation defines 21 semantic entity categories spanning textual, tabular, visual, chart-specific, and reference entities.

Table 3: Document entity categories used in DOCHYPER. Categories are grouped by their primary source type.

| No. | Entity category | Description |
|---|---|---|
| 1 | Text keyword | Question- or topic-relevant keywords extracted from text blocks |
| 2 | Text concept | Abstract concepts or technical terms in text blocks |
| 3 | Text person | Person names mentioned in text |
| 4 | Text organization | Organizations or institutions mentioned in text |
| 5 | Text location | Geographic or address-like locations in text |
| 6 | Text date | Dates or temporal expressions in text |
| 7 | Text number | Numerical values, quantities, or identifiers in text |
| 8 | Table header | Column headers extracted from tables |
| 9 | Table row label | Row labels or row identifiers in tables |
| 10 | Table cell | Cell-level table values when available |
| 11 | Table caption | Table captions or table titles |
| 12 | Image object | Objects or visual items detected in images |
| 13 | Image person | Persons appearing in images |
| 14 | Image scene | Scene-level visual entities |
| 15 | Image text | Text visible inside images |
| 16 | Chart axis | Axis labels, units, or ranges |
| 17 | Chart series | Data series or plotted variables |
| 18 | Chart legend | Legend entries in charts |
| 19 | Chart title | Chart titles when extracted |
| 20 | Figure reference | References to figures or images |
| 21 | Table reference | References to tables |

## A.2 INDEX STATISTICS

Table 4 reports corpus-level statistics of the constructed DOCHYPER indices on LongDocURL and MMLongBench-Doc. The statistics include doc and page counts, entity and hyperedge counts, average index size per doc and per page, and the distribution of hyperedges by type. Since the stored index records page indices for extracted entities and hyperedges, total pages are computed from the maximum observed page index in each doc.

## A.3 HYPERPARAMETER SETTINGS

Table 5 lists the main hyperparameters used for DOCHYPER index construction and retrieval/QA. Unless otherwise specified, these values are used for the main results.

## A.4 PROMPTS

This subsection reports the prompt templates used in our experiments. Table 6 summarizes the LLM/LVLM prompts and deterministic templates used by DOCHYPER. The prompts cover semantic entity extraction, element understanding, hyperedge-content construction, query understanding, answer generation, and short-answer normalization. Text-centric extraction and summarization use Qwen-Max Yang et al. (2025), visual element understanding uses Qwen-VL-Max Bai et al. (2025), and local answer generation uses Qwen2-VL-7B-Instruct Alibaba (2024). The final answer-generator context formats are reported separately in Figure 5.

Table 4: Index statistics of DOCHYPER on LongDocURL and MMLongBench-Doc. Counts are computed from the constructed multimodal hypergraph indices.

| Statistic | LongDocURL | MMLongBench-Doc |
|---|---|---|
| Docs | 396 | 135 |
| Total pages | 33,795 | 6,498 |
| Total entities | 1,301,881 | 229,225 |
| Total hyperedges | 238,612 | 51,823 |
| Avg. pages / doc | 85.3 | 48.1 |
| Avg. entities / doc | 3,287.6 | 1,698.0 |
| Avg. hyperedges / doc | 602.6 | 383.9 |
| Avg. entities / page | 38.5 | 35.3 |
| Avg. hyperedges / page | 7.1 | 8.0 |
| Text-block hyperedges | 192,435 | 42,779 |
| Table-semantic hyperedges | 12,149 | 1,646 |
| Image-event hyperedges | 16,122 | 3,145 |
| Chart-trend hyperedges | 4,074 | 1,005 |
| Cross-element hyperedges | 13,832 | 3,248 |

Table 5: Main hyperparameters used for DOCHYPER index construction and retrieval/QA. QA decoding parameters follow the local Qwen2-VL-7B-Instruct generation configuration unless explicitly specified.

| Hyperparameter | Value |
|---|---|
| **Index construction** | |
| Max entities per element | 20 |
| Max entities per cross-element hyperedge ($K_c$) | 15 |
| Max hyperedge content length | 2,000 chars |
| Text input truncation for entity extraction | 3,000 chars |
| **Retrieval & QA** | |
| Top-K entities ($K_V$) | 5 |
| Top-K hyperedges ($K_E$) | 5 |
| Max input tokens | 31,744 |
| Max output tokens | 1,024 |
| Sampling | true (model default) |
| Decoding temperature | 0.01 (model default) |
| Top-p | 0.001 (model default) |
| Top-k | 1 (model default) |

The following boxes instantiate the main prompts. For LVLM prompts, image inputs are explicitly marked to distinguish textual prompt content from visual payloads. Deterministic rows in Table 6 correspond to template-based construction steps rather than additional model calls.

Table 6: Overview of prompts and deterministic templates used in DOCHYPER, grouped by entity extraction, hyperedge construction, and retrieval/QA/evaluation stages. IDs starting with P denote LLM or LVLM prompts, while IDs starting with T denote deterministic construction templates.

| ID | Purpose | Model | Input | Output |
|---|---|---|---|---|
| **Entity extraction and element understanding** | | | | |
| P1 | Text entity extraction | Qwen-Max | Text element | Text entities |
| P2 | Table structure recognition | Qwen-VL-Max | Table image | Table headers, row labels, and Markdown |
| P3 | Image understanding | Qwen-VL-Max | Image element | Visual entities and image fields |
| P4 | Chart analysis | Qwen-VL-Max | Chart image | Chart entities and chart analysis fields |
| **Hyperedge construction** | | | | |
| P5 | Text summary | Qwen-Max | Long text element | Text-block hyperedge summary |
| P6 | Table semantic description | Qwen-Max | Table Markdown and caption | Table-semantic hyperedge content |
| T1 | Image-event content | Rule-based | Image understanding results | Image-event hyperedge content |
| P7 | Chart trend description | Qwen-Max | Chart analysis fields | Chart-trend hyperedge content |
| T2 | Cross-element content | Rule-based | Same-page representative entities | Cross-element hyperedge content |
| **Retrieval, answer generation, and evaluation** | | | | |
| P8 | Query keyword extraction | Qwen-Max | Question | Low-level and high-level keywords |
| P9 | Answer generation | Qwen2-VL-7B | Question and QA context | Final answer |
| P10 | Short-answer extraction | Qwen-Turbo | Generated response | Normalized short answer |

## Prompt for Entity Extraction

**Text Entity Extraction**

**[System]**
You are a professional information extraction assistant skilled at identifying key entities from text.

**[User]**
Extract key entities from the following text.

Text content:
`<content>`

Please extract the following types of entities: key concepts/terms, person names, organizations/institutions, locations, dates/times, and important numerical values.

Return in JSON format as follows:
{"entities": [{"name": "entity name", "type": "keyword/concept/person/organization/location/date/number", "description": "brief description"}]}

Notes: only extract important entities, avoid overly granular extraction, keep entity names concise, and return at most `<max_entities>` entities.

---

**Table Structure Recognition**

**[System]**
You are a professional table analysis assistant.

**[User]**
[Image Input: `<table_image>`]
Please analyze this table image and return structured information.

Please identify: column headers, row labels if any, and a brief description of the table.

Return in JSON format:
{"headers": ["column header 1", "column header 2"], "row_labels": ["row label 1", "row label 2"], "description": "table description", "markdown": "`| Col1 | Col2 |\n|---|---|\n| Data | Data |`"}

---

**Image Understanding**

**[System]**
You are a professional image analysis assistant skilled at identifying content and events in images.

**[User]**
[Image Input: <image_element>]
Please analyze this image in detail, identifying key visual elements and the overall scene.

Please provide: a list of main objects/persons detected, the overall scene description, and the main information or events conveyed by the image.

Return in JSON format:
{"objects": [{"name": "object name", "description": "object description", "type": "person/object/scene/text"}], "scene_description": "overall scene description", "event_description": "event or action description", "key_information": "key information conveyed by the image"}

---

**Chart Analysis**

**[System]**
You are a professional data visualization analysis assistant skilled at interpreting scientific charts.

**[User]**
[Image Input: <chart_image>]
Please analyze this scientific chart in detail.

Please identify and extract: x-axis label, unit, and range; y-axis label, unit, and range; chart type; data series and the trend of each series; chart title and legend; and key observations.

Return in JSON format:
{"chart_type": "chart type", "title": "chart title", "x_axis": {"label": "X-axis label", "unit": "unit", "range": "range description"}, "y_axis": {"label": "Y-axis label", "unit": "unit", "range": "range description"}, "data_series": [{"name": "series name", "color": "color", "trend": "trend description"}], "legend": ["legend item"], "key_observations": "key observations"}

## Prompt for Hyperedge Construction

**Text Summary**

**[System]**
You are a professional text summarization assistant.

**[User]**
Please generate a concise summary of the following text, preserving key information and relationships. Keep the summary under 200 words.

Text:
<long_text_content>

---

**Table Semantic Description**

**[System]**
You are a professional data analysis assistant skilled at describing table content.

**[User]**
Please generate a complete semantic description for the following table, describing its content, structure, and key data.

Table Markdown:
<table_markdown>

Table Caption: <table_caption>

Please generate a natural language description including: theme and purpose of the table, meaning of key columns/rows, important data points and trends, and any notable patterns.

Description:

---

**Chart Trend Description**

**[System]**
You are a professional data analysis expert skilled at describing chart trends and insights.

**[User]**
Please generate a detailed trend and analysis description for the following scientific chart.

Chart type: <chart_type>
X-axis: <x_axis>
Y-axis: <y_axis>
Data series: <data_series>
Chart title: <title>

Please generate a natural language description including: overall trend of the chart, patterns of change in each data series, key data points and turning points, and main conclusions and insights.

Description:

## Prompt for Query Keyword Extraction

**[System]**
You are a query understanding assistant for document retrieval.

**[User]**
Given the following question, extract two types of keywords:
1. Low-level keywords: concrete entity names, proper nouns, specific terms, numbers, dates, or identifiable objects mentioned or implied in the question.
2. High-level keywords: abstract themes, concepts, causal relationships, or summary descriptions that capture the intent of the question.

Question: <question>

Return ONLY a JSON object in this exact format, with no markdown and no extra text:
{"low_level_keywords": ["keyword1", "keyword2"], "high_level_keywords": ["concept1", "concept2"]}

Rules: low-level keywords should be short, high-level keywords should be descriptive phrases, and at least one keyword should be returned in each list.

## Prompt for QA Generation

**[System]**
You are a document QA assistant. Answer accurately based on the provided evidence.

**[User]**
[Page <page_no> Image] Below is the image of page <page_no>.
<page_image>

Optional evidence-organization variants:
[Page <page_no>]
<entity_or_hyperedge_text>

Answer the question based on the following document evidence.

Question: <question>

<format_instruction>
Answer:

Table 7: API cost statistics for index construction. Tokens / call are reported as average input / output tokens. Total cost is estimated for the full dataset based on the recorded cost per recorded doc, with 396 docs and 2,325 questions for LongDocURL and 135 docs and 847 questions for MMLongBench-Doc. "–" indicates no API calls of that type were made; "N/A" indicates the statistic was not recorded.

| Method | LLM calls / doc | LVLM calls / doc | LLM tokens / call | LVLM tokens / call | Cost / doc | Cost / query | Total cost |
|---|---|---|---|---|---|---|---|
| **LongDocURL** | | | | | | | |
| DocHyper | 464.6 | 174.8 | 251.1 / 244.7 | 744.7 / 249.0 | $0.26 | $0.0440 | $102.20 |
| GraphRAG | 243.5 | – | 1,853.1 / 964.2 | – | $0.51 | $0.0861 | $200.24 |
| MMRAG | – | 89.1 | – | 2,562.5 / 449.1 | $0.08 | $0.0132 | $30.60 |
| HyperGraphRAG | 144.1 | – | N/A | – | $0.71 | $0.1207 | $280.68 |
| **MMLongBench-Doc** | | | | | | | |
| DocHyper | 304.7 | 104.7 | 236.5 / 231.4 | 707.2 / 216.3 | $0.16 | $0.0248 | $21.00 |
| GraphRAG | 111.8 | – | 1,880.6 / 987.2 | – | $0.24 | $0.0378 | $32.02 |
| MMRAG | – | 44.6 | – | 2,562.5 / 449.1 | $0.04 | $0.0062 | $5.22 |
| HyperGraphRAG | 66.9 | – | N/A | – | $0.33 | $0.0528 | $44.74 |

**Prompt for Short-Answer Extraction**

**[System]**
Given the question and analysis, you are tasked to extract answers with required formats from the free-form analysis. The extracted answer should be one of Integer, Float, String, and List. If the analysis shows that the question cannot be answered from the given documents, type `Not answerable`. If the analysis only says it cannot read or understand the images or documents, type `Fail to answer`. Use `<concise_answer>` and `</concise_answer>` around the extracted answer, and use `<answer_format>` and `</answer_format>` around the answer format.

**[User]**
Question: `<question>`
Analysis: `<model_response>`

Extracted answer:
`<concise_answer><answer></concise_answer>`
Answer format:
`<answer_format><format></answer_format>`

### A.5 QA Context Comparison

Complementing the prompt templates summarized in Table 6, this subsection reports the final QA context formats passed to the answer generator. We distinguish the evidence units used for retrieval from the context units finally passed to the answer generator because the compared systems cover chunk-based RAG Lewis et al. (2020); Gao et al. (2023), page-image retrieval Faysse et al. (2025); Cho et al. (2024), binary-graph retrieval Peng et al. (2026); Edge et al. (2024), and agentic document QA Han et al. (2025). Figure 5 summarizes the final context schemas used by these baselines and by DocHyper.

## B Experiment Details

### B.1 Index Construction Cost

Table 7 summarizes the API cost of constructing indices for representative methods on LongDocURL and MMLongBench-Doc. We report average API calls per doc, average input/output tokens per call, estimated API cost per doc, estimated cost per query, and estimated total cost for the full dataset. These statistics focus on API usage during index construction and do not include local preprocessing, embedding computation, or answer-generation cost. For cost estimation, LLM calls use Qwen-Max pricing, with input, output, and cache-hit input tokens charged at $0.353, $1.412, and $0.071 per million tokens, respectively. LVLM calls use Qwen-VL-Max pricing, with input, output, and cache-hit input tokens charged at $0.235, $0.588, and $0.047 per million tokens, respectively.

# C Additional Results and Analysis

## C.1 Additional Baseline Comparisons

Table 8 mirrors the main comparison while omitting page-level retrieval metrics. It adds HippoRAG2 Gutiérrez et al. (2025), MLDocRAG Zhang & Wu (2026), and HyperGraphRAG Luo et al. (2026) to the baseline comparison.

**HippoRAG2.** In actual implementation, we split the Markdown text generated by MinerU into segments of 1,200 tokens, with an overlap of 100 tokens, and extract entity–relation triples using an OpenIE pipeline (Qwen-Max) to build a document-level graph where chunks and facts serve as dual node types. Personalized PageRank over the chunk–fact graph retrieves the top-10 chunks per question; we pass the top-5 or top-10 chunks as plain-text evidence to the answer generator (Qwen2-VL-7B-Instruct). Indexing across LongDocURL and MMLongBench-Doc used approximately 44k successful Qwen-Max calls at a total cost of $44.71.

**MLDocRAG.** In our execution, we start from MinerU's layout-parsed output: text elements are segmented with a 1,200-token sliding window and 100-token overlap, while non-text elements (images, tables, charts) are each kept as a single chunk assembled from captions and OCR. For each chunk we generate up to 10 synthetic QA pairs via Qwen-Max, and build a document-level MCQG graph with chunk and query dual node types connected by C-Q edges. At retrieval time, the user question matches query node embeddings (top-$\alpha$=50), expands through Q-Q edges, traces to source chunks via C-Q edges, and ranks them by max cosine similarity to select top-$\kappa$=5 chunks as plain-text evidence for Qwen2-VL-7B-Instruct answer generation. Due to the long indexing time, only MMLongBench-Doc results are available; LongDocURL results are pending.

**HyperGraphRAG.** In our execution, HyperGraphRAG segments MinerU-generated Markdown into fixed 1,200-token chunks with 100-token overlap (tiktoken gpt-4o encoding). For each chunk, Qwen-Max extracts (entity, type, description) triples and (hyper-relation, knowledge_fragment) hyperedges via an OpenIE prompt. Same-name entities across chunks are merged (LLM-summarized when the combined description is overly long), and identical hyperedges are deduplicated with weight accumulation, yielding a document-level hypergraph with entity and hyperedge dual node types connected by directed hyperedge-to-entity edges. At query time, an LLM extracts low-level (entity name) and high-level (hyper-relation) keywords from the question for a dual-path retrieval: the local path retrieves top-$K$ entities, expands along the graph to associated hyperedges, then traces back to source chunks; the global path retrieves top-$K$ hyperedges, expands to associated entities, and similarly traces back to chunks. Retrieved results (entities, hyperedges, source chunks) are fused into structured CSV evidence context, fed to Qwen2-VL-7B-Instruct for answer generation. We use $K = 5$ for the main evaluation. Indexing across LongDocURL and MMLongBench-Doc used approximately 66k successful Qwen-Max calls at a total cost of $325.42.

Table 8: Additional baseline results on LongDocURL and MMLongBench-Doc. Retrieval Unit denotes the evidence units retrieved by each method; QA Context Unit denotes the context provided to the answer generator. Abbreviations: PI: page images, TC: text chunks, VC: visual crops, Q: synthetic queries, E: entities, R: binary relations, HE: hyperedges, and TP: text passages. $Acc$ measures average normalized answer correctness, and $F_1$ measures soft answer-level precision–recall balance.

| Method | Retrieval Unit | Retrieval Top-K | QA Context Unit | LongDocURL | | MMLongBench-Doc | |
|---|---|---|---|---|---|---|---|
| | | | | *Acc* | $F_1$ | *Acc* | $F_1$ |
| **LVLMs** | | | | | | | |
| LVLM | - | - | PI | 38.7 | 39.1 | 26.7 | 27.1 |
| **RAG / Agentic Pipeline** | | | | | | | |
| TextRAG | TC | 5 | TC | 34.3 | 35.8 | 19.0 | 20.8 |
| | | 10 | | 35.5 | 36.7 | 20.4 | 22.0 |
| MMRAG | TC+VC | (5, 5) | TC+VC | 39.3 | 40.4 | 28.8 | 30.3 |
| | | (10, 5) | | 39.9 | 41.0 | <u>29.5</u> | <u>31.0</u> |
| | | (5, 5) | TC+PI | 45.4 | 46.2 | 28.0 | 28.8 |
| | | (10, 5) | | 46.0 | 46.7 | 26.5 | 27.5 |
| GraphRAG | E | 5 | E+R+TC | 33.9 | 35.1 | 17.6 | 19.6 |
| | | 10 | | 33.8 | 34.9 | 18.0 | 20.0 |
| | E+R | (5, 5) | | 34.8 | 35.7 | 18.2 | 19.9 |
| | | (10, 10) | | 34.7 | 35.9 | 19.1 | 21.1 |
| HippoRAG2 | TC+R | (5,5) | TC | 37.7 | 38.9 | 22.0 | 23.6 |
| | | (10,10) | | 40.0 | 40.8 | 23.2 | 24.2 |
| MLDocRAG | Q | 5 | TC+VC | - | - | 21.9 | 22.3 |
| HyperGraphRAG | E+HE | (5,5) | E+HE+TC | 33.4 | 35.3 | 17.2 | 18.2 |
| M3DocRAG | PI | 5 | PI | 35.5 | 36.0 | 23.9 | 24.2 |
| | | 10 | | 41.2 | 41.8 | 29.1 | 29.3 |
| MDocAgent | TP+PI | (5, 5) | TP+PI | 45.0 | 45.6 | 27.3 | 28.7 |
| | | (10, 10) | | <u>46.6</u> | <u>47.2</u> | 27.1 | 28.0 |
| **Ours** | | | | | | | |
| DocHyper | E+HE | (5, 5) | PI | **49.6** | **50.2** | **33.0** | **33.2** |

### C.2 Qualitative Case Studies

Tables 9 to 13 present five qualitative LongDocURL cases where DocHyper's advantage comes from retrieving structured hyperedge evidence rather than only increasing the number of retrieved pages. The selected cases cover table–figure co-occurrence, chart–text numeric grounding, figure–text parameter disambiguation, exact layout-title localization, and table cell grounding. Together, they show how text-block, table-semantic, image/chart, and same-page cross-element hyperedges preserve the evidence relationships needed by the answer generator.

Final QA Context Formats

**LVLM.** Direct page-image QA without retrieval.
Following is our question: <question><user_question></question>
Below is the <i>-th image (total <n> images). [Image Input: <page_image_i>]

**TextRAG.** Text-only QA over retrieved chunks.
Answer the question based on the following document evidence.
Evidence: ——Related Texts—— <retrieved_text_chunks>
Question: <question>
Answer:

**MMRAG.** Multimodal chunk QA with two context variants.
*TC+VC context: retrieved text chunks and textualized multimodal chunks are serialized, with retrieved crop images attached as image inputs.*
Evidence: ——Related Texts—— <retrieved_text_chunks>
——Related Images / Multimodal Evidence——
<retrieved_image_descriptions_or_multimodal_chunks>
Question: <question>
[Image Input: <visual_crop_image_1>, ..., <visual_crop_image_k>]
*TC+PI context: retrieved text chunks and visual crops jointly determine the source pages, while the user message contains the question and source-page images.*
Question: <question>
[Image Input: <page_image_1>, ..., <page_image_k>]

**GraphRAG.** Text-only QA over serialized graph evidence.
*Entity-only retrieval context: relations are expanded from retrieved entities.*
Question: <question>
Evidence: ——Entities—— <retrieved_entities>
——Relationships—— <relations_expanded_from_retrieved_entities>
——Sources—— <entity_associated_text_chunks>
*Entity–relation retrieval context: direct relation retrieval is fused with entity-expanded relations.*
Question: <question>
Evidence: ——Entities—— <retrieved_entities_and_relation_endpoint_entities>
——Relationships—— <expanded_relations_and_directly_retrieved_relations>
——Sources—— <entity_and_relation_associated_text_chunks>

**M3DocRAG.** Page-image QA over retrieved pages.
Question: <question>
[Image Input: <retrieved_page_image_1>, ..., <retrieved_page_image_k>]

**MDocAgent.** Agentic QA over retrieved text pages and page images.
*General agent text message:*
<retrieved_text_page_1> ... <retrieved_text_page_k>
Question: <question>
*General agent image message:*
[Image Input: <retrieved_page_image_1>, ..., <retrieved_page_image_k>]
Question: <question>
*Text / image agents: answer with modality-specific clues from the general agent.*
*Summary agent:*
General Agent: <general_response>; Text Agent: <text_response>; Image Agent: <image_response>

**DocHyper.** Page-aligned QA over source pages traced from retrieved entities and hyperedges.
[Page <page_no> Image] Below is the image of page <page_no>.
[Image Input: <page_image>]
Answer the question based on the following document evidence.
Question: <question>
Answer:

Figure 5: Final QA context formats used by baselines and DocHyper. The context format denotes what is passed to the answer generator after retrieval. Abbreviations follow Table 1.

Table 9: Qualitative case showing table–figure co-occurrence. DOCHYPER links the table title, table semantics, and same-page timing-diagram evidence through hyperedges. Page numbers are 1-based.

| Method | Item | Content |
|---|---|---|
| **Question ID:** extract_fig2tab_4066167_7_36_11. **Doc no.:** 4066167.<br>**Question:** List names of the figures at the page which contains a table named "Table 13. JTAG full voltage range electricals (continued)".<br>**Gold answer:** Figure 6. Test clock input timing; Figure 7. Boundary scan (JTAG) timing. **Source page:** 24. | | |
| **LVLM** | Score | 0.0 |
| | Pred answer | Figure 13.1; Figure 13.2; . . . |
| | Detailed response | The figures are Figure 13.1, Figure 13.2, . . . , which are unrelated figure-number strings rather than the requested figure titles. |
| | Observation | The direct LVLM setting does not first localize the named table and instead returns unrelated figure-number strings. |
| **MMRAG**<br>*TC+PI context* | Retrieved pages | 8, 9, 11, 16, 22, 23, 30, 33 |
| | Score | 0.0 |
| | Pred answer | Figure 4. TRACE_CLKOUT specifications; Figure 5. Trace data specifications |
| | Detailed response | The figures are Figure 4. TRACE_CLKOUT specifications and Figure 5. Trace data specifications. |
| | Observation | Multimodal chunk retrieval returns related timing-specification evidence but misses the source page containing the named continued table. |
| **M3DocRAG**<br>*top-5* | Retrieved pages | 32, 31, 22, 33, 15 |
| | Score | 0.0 |
| | Pred answer | Fail to answer |
| | Detailed response | Figure 6. JTAG_CLKOUT specifications |
| | Observation | Page-image retrieval selects visually related pages but does not preserve the table–figure co-occurrence needed by the question. |
| **M3DocRAG**<br>*top-10* | Retrieved pages | 32, 31, 22, 15, 9, 12, 27, 33, 29, 11 |
| | Score | 0.0 |
| | Pred answer | Figure 6 |
| | Detailed response | Figure 6. |
| | Observation | Increasing the page budget still fails to recover the full answer because the retrieved pages are not organized around the named table. |
| **MDocAgent** | Retrieved pages | Text: 23, 21, 22, 22, 27; Image: 27, 29, 30, 21, 17 |
| | Score | 0.0 |
| | Pred answer | Figure 4. TRACE_CLKOUT specifications; Figure 5. Trace data specifications |
| | Detailed response | The figures at the page containing the table are Figure 4. TRACE_CLKOUT specifications and Figure 5. Trace data specifications. |
| | Observation | The agentic pipeline finds plausible timing figures but selects figures from the wrong local context. |
| **DOCHYPER**<br>*PI context* | Retrieved pages | 7, 9, 13, 22, 24, 30, 33, 35, 36 |
| | Score | 1.0 |
| | Pred answer | Figure 6. Test clock input timing; Figure 7. Boundary scan (JTAG) timing |
| | Detailed response | ["Figure 6. Test clock input timing", "Figure 7. Boundary scan (JTAG) timing"] |
| | Observation | Entity and hyperedge retrieval links the table title, table semantics, timing-diagram entities, and same-page cross-element context, tracing the evidence to the correct source page. |

Table 10: Qualitative case showing chart–text numeric grounding. DOCHYPER benefits from hyperedges that connect chart entities, local text numbers, and the queried sector. Page numbers are 1-based.

| Method | Item | Content |
|---|---|---|
| **Question ID:** free_gpt4o_4115594_31_60_6. **Doc no.:** 4115594.<br>**Question:** How many chefs and cooks participated in the training survey from the culinary sector?<br>**Gold answer:** 3. **Source pages:** 50, 51. | | |
| **LVLM** | Score | 0.0 |
| | Pred answer | 10 |
| | Detailed response | The number of chefs and cooks who participated in the training survey from the culinary sector is 10. |
| | Observation | The direct LVLM setting confuses the chef/cook count with broader participant counts in nearby visual material. |
| **MMRAG**<br>*TC+PI context* | Retrieved pages | 31, 32, 37, 47, 49, 50, 51 |
| | Score | 0.0 |
| | Pred answer | 15 |
| | Detailed response | 15 chefs and cooks participated in the training survey from the culinary sector. |
| | Observation | Multimodal chunk retrieval reaches nearby pages but does not bind the chart labels to the correct local numbers. |
| **M3DocRAG**<br>*top-5* | Retrieved pages | 40, 37, 39, 49, 51 |
| | Score | 0.0 |
| | Pred answer | 15 |
| | Detailed response | 15 |
| | Observation | Page-image retrieval retrieves visually related pages but selects a total or neighboring count instead of the chef/cook slice. |
| **M3DocRAG**<br>*top-10* | Retrieved pages | 40, 39, 37, 50, 49, 51, 55, 48, 57, 52 |
| | Score | 0.0 |
| | Pred answer | 5 |
| | Detailed response | 5 |
| | Observation | Increasing the page budget still leaves the answer generator without an explicit chart–text grouping for the queried sector. |
| **MDocAgent** | Retrieved pages | Text: 30, 49, 30, 30, 37; Image: 39, 33, 50, 49, 36 |
| | Score | 0.0 |
| | Pred answer | 12 |
| | Detailed response | There were 12 participants in the training survey from the culinary sector. |
| | Observation | The agentic setting retrieves related evidence but answers from an aggregate participant count. |
| **DOCHYPER**<br>*PI context* | Retrieved pages | 31, 36, 37, 38, 50 |
| | Score | 1.0 |
| | Pred answer | 3 |
| | Detailed response | 3 |
| | Observation | A chart-trend hyperedge and a same-page cross-element hyperedge connect the chart labels Chef and Service staff with text numbers 3 and 1 under the culinary-sector context, allowing the model to select the correct count. |

Table 11: Qualitative case showing figure–text parameter grounding. DOCHYPER uses hyperedges to associate the figure reference, example context, and local numeric parameter. Page numbers are 1-based.

| Method | Item | Content |
|---|---|---|
| **Question ID:** free_gemini15_pro_4139004_25_54_3. **Doc no.:** 4139004.<br>**Question:** What is the conductivity of the coating described in Example 3 (Figures 8 and 9) in the Dirichlet case?<br>**Gold answer:** 0.3. **Source pages:** 28, 29. | | |
| **LVLM** | Score | 0.0 |
| | Pred answer | 0.1 |
| | Detailed response | The conductivity of the coating described in Example 3 (Figures 8 and 9) in the Dirichlet case is 0.1 ($S.m^{-1}$). |
| | Observation | The direct LVLM setting selects a nearby conductivity value from a different example. |
| **MMRAG**<br>*TC+PI context* | Retrieved pages | 25, 26, 27, 28, 32, 37, 38, 41, 42 |
| | Score | 0.0 |
| | Pred answer | 0.1 |
| | Detailed response | The conductivity of the coating described in Example 3 is 0.1 ($S.m^{-1}$). |
| | Observation | Chunk retrieval reaches the document region but does not disambiguate the Example 3 figure mention from other similar parameters. |
| **M3DocRAG**<br>*top-5* | Retrieved pages | 45, 46, 32, 38, 48 |
| | Score | 0.0 |
| | Pred answer | 0.005 |
| | Detailed response | 0.005 ($S.m^{-1}$). |
| | Observation | Visual page retrieval selects a wrong numerical parameter from another visually similar technical section. |
| **M3DocRAG**<br>*top-10* | Retrieved pages | 45, 31, 39, 48, 30, 53, 32, 46, 38, 51 |
| | Score | 0.0 |
| | Pred answer | 0.005 |
| | Detailed response | 0.005 ($S.m^{-1}$). |
| | Observation | The larger page set adds more distractor numeric values without preserving the figure–example association. |
| **MDocAgent** | Retrieved pages | Text: 37, 25, 39, 27, 34; Image: 37, 44, 34, 42, 36 |
| | Score | 0.0 |
| | Pred answer | 0.05 |
| | Detailed response | The conductivity of the coating described in Example 3 (Figures 8 and 9) in the Dirichlet case is 0.05 $S.m^{-1}$. |
| | Observation | The agentic pipeline also confuses nearby technical constants and examples. |
| **DOCHYPER**<br>*PI context* | Retrieved pages | 25, 28, 30, 32, 43 |
| | Score | 1.0 |
| | Pred answer | 0.3 |
| | Detailed response | 0.3 ($S.m^{-1}$) |
| | Observation | A cross-element hyperedge ties the local figure geometry to the text statement "conductivity, 0.3", preventing confusion with other conductivity values in neighboring examples. |

Table 12: Qualitative case showing exact layout-title localization. DOCHYPER retrieves title-level hyperedges instead of semantically related but incorrect subsections. Page numbers are 1-based.

| Method | Item | Content |
|---|---|---|
| **Question ID:** topic2title_4054021_35_37_3. **Doc no.:** 4054021.<br>**Question:** Which sections should be referred to for understanding the procedural rules and dispute adjudication agreements in the bidding process? Select titles from the doc that best answer the question, do not alter or analyze the titles themselves.<br>**Gold answer:** PROCEDURAL RULES; DISPUTE ADJUDICATION AGREEMENT. **Source pages:** 35, 37. | | |
| **LVLM** | Score | 0.0 |
| | Pred answer | Section 19.0 Claims, Disputes and Arbitration; Section 19.2; Section 19.3; . . . |
| | Detailed response | The sections include Section 19.0 Claims, Disputes and Arbitration and several detailed dispute-adjudication subsections. |
| | Observation | The direct LVLM setting returns semantically related subsections instead of the exact titles requested by the question. |
| **MMRAG**<br>*TC+PI context* | Retrieved pages | 24, 25, 27, 28, 29, 30, 31, 35, 36, 37 |
| | Score | 0.0 |
| | Pred answer | Section 4 - Contract Data; Section 19.0 Claims, Disputes and Arbitration; Section 19.4; . . . |
| | Detailed response | Section 4 - Contract Data, Section 19.0 Claims, Disputes and Arbitration, Section 19.4, Section 19.5, Section 19.6, Section 19.7. |
| | Observation | Although relevant pages are retrieved, chunk evidence encourages semantic expansion rather than exact layout-title extraction. |
| **M3DocRAG**<br>*top-5* | Retrieved pages | 42, 31, 38, 27, 35 |
| | Score | 0.0 |
| | Pred answer | Section 4 - Contract Data; 19.4; 19.7; 19.5 |
| | Detailed response | Section 4 - Contract Data, 19.4 [Obtaining Dispute Adjudication Board's Decision], 19.7 [Arbitration], 19.5 [Failure to Comply with Dispute Adjudication Board's Decision]. |
| | Observation | Page-image retrieval selects related contract pages but not the requested section titles. |
| **M3DocRAG**<br>*top-10* | Retrieved pages | 42, 27, 38, 31, 36, 35, 33, 41, 25, 40 |
| | Score | 0.0 |
| | Pred answer | Section 4 - Contract Data |
| | Detailed response | Section 4 - Contract Data |
| | Observation | More pages still fail to preserve title-level evidence units. |
| **MDocAgent** | Retrieved pages | Text: 36, 27, 30, 36, 37; Image: 38, 49, 34, 43, 35 |
| | Score | 0.202 |
| | Pred answer | Section 4 – Contract Data; Section 5 – Standard Forms; Section 6 – Specifications; Section 3 – Procedural Rules |
| | Detailed response | The sections are Section 4 – Contract Data, Section 5 – Standard Forms (Contract), Section 6 – Specifications, and Section 3 – Procedural Rules. |
| | Observation | The agent retrieves related passages but over-includes neighboring contract sections. |
| **DOCHYPER**<br>*PI context* | Retrieved pages | 30, 31, 33, 34, 35, 37 |
| | Score | 1.0 |
| | Pred answer | PROCEDURAL RULES; DISPUTE ADJUDICATION AGREEMENT |
| | Detailed response | ["PROCEDURAL RULES", "DISPUTE ADJUDICATION AGREEMENT"] |
| | Observation | Text-block hyperedges preserve exact heading units and keep the requested document titles separate from lower-level dispute clauses. |

Table 13: Qualitative case showing table cell grounding. DOCHYPER uses table-semantic hyperedges to preserve row–column units and avoid nearby numeric distractors. Page numbers are 1-based.

| Method | Item | Content |
|---|---|---|
| **Question ID:** free_gpt4o_4027862_74_102_7. **Doc no.:** 4027862.<br>**Question:** What was the CIF value (in B$ million) of imported ornamental plants in Brunei in 2020?<br>**Gold answer:** 0.19. **Source page:** 81. | | |
| **LVLM** | Score | 0.0 |
| | Pred answer | 0.13 |
| | Detailed response | The CIF value (in B$ million) of imported ornamental plants in Brunei in 2020 was 0.13–48.70 /plant. |
| | Observation | The direct LVLM setting confuses CIF value with average CIF price. |
| **MMRAG**<br>*TC+PI context* | Retrieved pages | 76, 78, 79, 82, 83, 85, 90, 91, 98 |
| | Score | 0.0 |
| | Pred answer | 811524 |
| | Detailed response | The CIF value of imported ornamental plants in Brunei in 2020 was 811,524 B$. |
| | Observation | Chunk retrieval extracts a related quantity but loses the row–column unit relationship needed for the value in millions. |
| **M3DocRAG**<br>*top-5* | Retrieved pages | 95, 84, 88, 102, 80 |
| | Score | 0.0 |
| | Pred answer | 10,696.54 |
| | Detailed response | 10,696.54. |
| | Observation | Page-image retrieval misses the specific ornamental-plants table row and selects another numerical entry. |
| **M3DocRAG**<br>*top-10* | Retrieved pages | 95, 84, 93, 81, 88, 86, 94, 78, 90, 80 |
| | Score | 0.0 |
| | Pred answer | 106.8 |
| | Detailed response | 106.80 |
| | Observation | The correct page enters the top-10 set, but the answer is still taken from a wrong cell because the table structure is not explicit. |
| **MDocAgent** | Retrieved pages | Text: 75, 82, 73, 82, 90; Image: 88, 92, 101, 96, 97 |
| | Score | 0.0 |
| | Pred answer | 16.8 |
| | Detailed response | The CIF value (in B$ million) of imported ornamental plants in Brunei in 2020 was 16.80 million. |
| | Observation | The agentic pipeline selects a nearby agricultural statistic rather than the required table cell. |
| **DOCHYPER**<br>*PI context* | Retrieved pages | 74, 75, 78, 80, 81, 83, 85, 91 |
| | Score | 1.0 |
| | Pred answer | 0.19 |
| | Detailed response | 0.19 |
| | Observation | A table-semantic hyperedge preserves the Total Import row and CIF Value (B$) column, while a cross-element hyperedge links the table to the Ornamental Plants Industry context on the source page. |