HiEvi-RAG: Hierarchical Evidence-Driven Reasoning for Long Document Understanding

Junyu Xiong[1] **Yonghui Wang**[1] **Rongjian Gu**[1] **Chenyu Liu**[2]
Bing Yin[2] **Wengang Zhou**[1*] **Houqiang Li**[1]
[1]University of Science and Technology of China
[2]iFLYTEK Research
xiongjyu@mail.ustc.edu.cn, zhwg@ustc.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) streamlines long-document understanding by leveraging retrieval mechanisms to restrict input images to a highly curated subset. However, existing multimodal RAG pipelines primarily face two critical challenges: first, standard semantic similarity retrievers frequently fetch topically overlapping yet answer-void distractor pages that mislead downstream generation; second, rigid single-pass pipelines heavily depend on initial retrieval success, where any omission of core evidence inevitably causes cascading errors. To address these challenges, we introduce **HiEvi-RAG**, a hierarchical, evidence-driven multimodal RAG framework for closed-domain document understanding. HiEvi-RAG systematically factorizes complex queries into a cooperative four-stage pipeline: (1) hierarchical question decomposition to break multi-hop root queries into atomic child questions; (2) coarse visual page retrieval leveraging a multimodal retriever to fetch candidate pages based on semantic similarity; (3) fine-grained page verification via EviAgent, a specialized multi-page verifier trained with GRPO to execute cross-page reasoning over multi-image blocks; and (4) memory-guided iterative generation that leverages accumulated sub-question context to execute multi-round, dynamic reasoning over the prioritized sequence. Extensive evaluations across four benchmarks demonstrate the robust efficacy and synergy of our framework, which significantly outperforms existing open-source baselines and exceeds the strongest reported baseline by an average of 8.05% in accuracy.

*Corresponding author.

1 Introduction

Closed-domain document understanding has become an indispensable core capability for practical AI systems deployed across various sectors, including science, finance, enterprise knowledge management, and technical support (Hui et al., 2024; Suri et al., 2025; Deng et al., 2025; Dong et al., 2024; Xia et al., 2024). In these scenarios, users typically pose queries over a proprietary document collection, where the answers are sparsely distributed across hundreds of pages containing diverse and rich information such as plain text, tables, charts, and illustrations (Wang et al., 2025a; Chia et al., 2025). However, due to the prohibitive number of images in these closed-domain documents, feeding the entire collection into a model in an end-to-end manner easily exceeds the context window limits of existing LLMs.

To address this bottleneck, a growing body of work on Retrieval-Augmented Generation (RAG) has emerged, which leverages retrieval mechanisms to restrict the number of input images to a highly curated subset prior to answer generation (Lewis et al., 2020; Cho et al., 2024; Suri et al., 2025). Initially, to bypass traditional OCR limitations and capture multimodal information, ColPali (Faysse et al., 2025) extends late interaction to rendered page images, while M3DocRAG (Cho et al., 2024) establishes this method as the core interface for multi-page QA. Subsequently, research shifted toward filtering distractors and optimizing ranking precision: MDocAgent (Han et al., 2025) and SimpleDoc (Jain et al., 2025) refine results via module collaboration or iterative cascading; RagVL (Chen et al., 2025) and MM-R5 (Xu et al., 2025) employ instruction-tuning or reinforcement learning to train strong VLM rerankers; and

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 7167–7181
October 24-29, 2026 ©2026 Association for Computational Linguistics

MoLoRAG (Wu et al., 2025) leverages page topology graphs for logic-aware traversal. Concurrently, DocAgent (Sun et al., 2025) and ALDEN (Yang et al., 2026) utilize memory feedback or active exploration protocols to gather evidence dynamically, while DMAP (Fu et al., 2026) constructs human-aligned structural maps for global document comprehension. Despite these advancements, two challenges remain: first, relying on semantic similarity misleads models with topically related yet answer-void pages; second, single-pass pipelines rigidly depend on initial retrieval, where any omission of core evidence makes correct generation impossible.

To address these challenges, we propose **HIEVI-RAG**, a four-step framework for long-document understanding. First, **hierarchical question decomposition** breaks down the complex, multi-hop root question into multiple atomic sub-questions, reducing retrieval difficulty while assisting the root question's resolution from different logical levels. Second, **coarse-grained page retrieval** utilizes a multimodal retriever (e.g., ops-ColQwen3) to fetch the top-k pages based on semantic similarity for each question, including both root and sub-questions. Third, for **fine-grained evidence verification**, inspired by DocR1 (Xiong et al., 2025), we employ GRPO (Shao et al., 2024) to train **EVIAGENT**, an evidence-aware multi-page reasoning model capable of cross-page reasoning and grounding over multi-image inputs to pinpoint question-relevant evidence. In this stage, for sub-questions, we strictly retain only the evidence pages identified by EVIAGENT; for the root question, EVIAGENT's verified evidence pages are prioritized at the front, while non-evidence pages follow in order of their initial semantic similarity, thereby achieving precise reranking. Fourth, during **memory-guided iterative generation**, an empty memory is initialized. The sub-questions and their corresponding evidence pages are fed into the model, with the sub-questions, generated reasoning traces, and intermediate answers subsequently accumulated in the memory. For the root question, the model takes the root query, the accumulated memory, and a batch of images as input. If the current memory and image batch are insufficient to resolve the question, the system appends the latest reasoning trace to the memory and fetches the next batch of images for iterative reasoning.

To verify the effectiveness of our approach, we conduct extensive experiments and analyses across four distinct datasets. The results demonstrate that our method significantly outperforms existing open-source baselines, exceeding the previous state-of-the-art by an average of 8.05%. Furthermore, we perform comprehensive ablation studies to validate the individual efficacy of each proposed module.

Our contributions can be summarized as follows:

- We propose HIEVI-RAG, a four-step multimodal RAG framework for long-document QA that unifies hierarchical question decomposition, coarse-to-fine retrieval, and memory-guided iterative generation.
- We introduce a fine-grained verification mechanism via EVIAGENT, which performs cross-page reasoning over multi-image inputs to precisely ground answer-supporting evidence.
- We formulate a memory-guided iterative generation process that utilizes sub-question reasoning traces to assist the root query while dynamically updating an explicit memory across multiple rounds to eliminate cascading retrieval errors.

2 Related Work

2.1 Document Visual Question Answering

Advancements in DocVQA have bifurcated into OCR-based and OCR-free paradigms. OCR-based methods convert pages into structured textual and spatial tokens to feed downstream reasoning frameworks (Tang et al., 2023; Wang et al., 2024a; Zhu et al., 2025), yet their performance remains inherently bounded by compounding parsing errors and the serialization loss of visual context. Conversely, OCR-free approaches directly train VLMs on raw document images, evolving from architectural scaling for resolution management (Ye et al., 2023; Feng et al., 2023; Hu et al., 2025) to evidence-aware reasoning via reinforcement learning (Xiong et al., 2025; Zheng et al., 2026). While these methods substantially enhance single- or multi-page reader execution, effectively navigating and distilling massive long-document repositories remains a largely orthogonal challenge.

2.2 Retrieval-Augmented Generation

Multimodal RAG methodologies have transitioned from coarse-grained text-retrieval extensions to visual-centric, layout-aware document discovery. Early pipelines leverage visual retrievers and VLM-based rerankers to screen candidate pages inde-

pendently based on superficial similarity scoring (Faysse et al., 2025; Cho et al., 2024; Chen et al., 2025; Xu et al., 2025). More recent efforts incorporate iterative multi-agent navigation, memory feedback, or topological page graphs to capture structural inter-page logic (Wu et al., 2025; Sun et al., 2025; Yang et al., 2026; Fu et al., 2026). Despite these scaling improvements, existing pipelines still evaluate retrieved pages as isolated query-image pairs, leaving them highly susceptible to topically overlapping but answer-void distractors. To bridge these gaps, HIEVI-RAG casts validation as a joint, cross-page verification problem and bypasses cascading failure modes via memory-guided iterative reasoning. Detailed related works are deferred to Appendix A.

3 Method

3.1 Overview and Task Formulation

Let a document be represented as an ordered sequence of page images $D = \{p_1, p_2, \ldots, p_N\}$ and q denote a user question in a closed-domain setting. The final answer is assumed to be grounded in D, although the supporting evidence may be sparsely distributed across multiple non-contiguous pages. When the document length N scales to hundreds of pages, direct end-to-end VLM ingestion becomes computationally prohibitive due to context window constraints. To bypass this bottleneck, HIEVI-RAG decomposes long-document QA into four cooperative stages: (1) hierarchical question decomposition, (2) coarse visual page retrieval, (3) evidence-aware page verification, and (4) memory-guided iterative generation. A core design rationale of our framework is the multi-granularity operation across stages: coarse retrieval is strictly recall-oriented, EVIAGENT conducts fine-grained verification over grouped page candidates via cross-page reasoning, and the final generator iteratively consumes the verified evidence sequence backed by an explicit memory. The overall pipeline is illustrated in Figure 1.

3.2 Hierarchical Question Decomposition

Given a complex root question q, this stage constructs a shallow reasoning hierarchy, as illustrated in Figure 1(1). Specifically, a lightweight LLM extracts key entities, structural attributes, constraints, and logical relations from q to instantiate a set of atomic child questions:

$$\mathcal{Q}(q) = \{q_1, q_2, \ldots, q_m\}, \qquad m \leq 5. \quad (1)$$

Each child question q_i is formulated as a distinct natural query covering a non-redundant subset of the global information need, while strictly avoiding near-paraphrases of the full prompt.

Crucially, partitioning the global query into localized targets reduces the inherent difficulty of multi-hop reasoning. This upfront decomposition prevents retrieval omissions typical of standalone root queries, ensuring core evidence is captured before downstream generation. The corresponding prompt template and example are detailed in Appendix B.1 and F.

3.3 Coarse Visual Page Retrieval

Following decomposition, we employ `Ops-ColQwen3-4B` (OpenSearch-AI, 2026), a ColPali-style late-interaction multimodal retriever (Faysse et al., 2025; Khattab and Zaharia, 2020), to construct a candidate page pool from rendered document images. As illustrated in Figure 1(2), this retrieval stage is split into offline index construction and online query matching.

Offline Index Construction. Each document page image $p_i \in D$ is encoded offline into a multi-vector representation to preserve fine-grained visual features:

$$\mathbf{v}_{p_i} = \mathcal{E}_{\text{page}}(p_i), \qquad i = 1, 2, \ldots, N. \quad (2)$$

The derived page embeddings are subsequently stored to build the global document index.

Online Query Matching. Upon receiving a root question q, the stage online encodes both q and its derived atomic child questions $q_i \in \mathcal{Q}(q)$ to obtain their respective text embeddings:

$$\mathbf{v}_q = \mathcal{E}_{\text{query}}(q), \quad \mathbf{v}_{q_i} = \mathcal{E}_{\text{query}}(q_i). \quad (3)$$

These query embeddings are then matched against the offline page index using the late-interaction operator to compute similarity scores across all document pages. Based on these scores, we rank and retain the Top-K ($K = 20$) candidate pages for the root question and each child question, denoted as T_q and T_{q_i} respectively.

3.4 Evidence-Aware Page Verification

To filter out topically related but answer-void distractors from the coarse retrieval pool, we introduce EVIAGENT, a specialized model trained to perform binary evidence page verification via cross-page reasoning. As illustrated in Figure 1(3), this stage

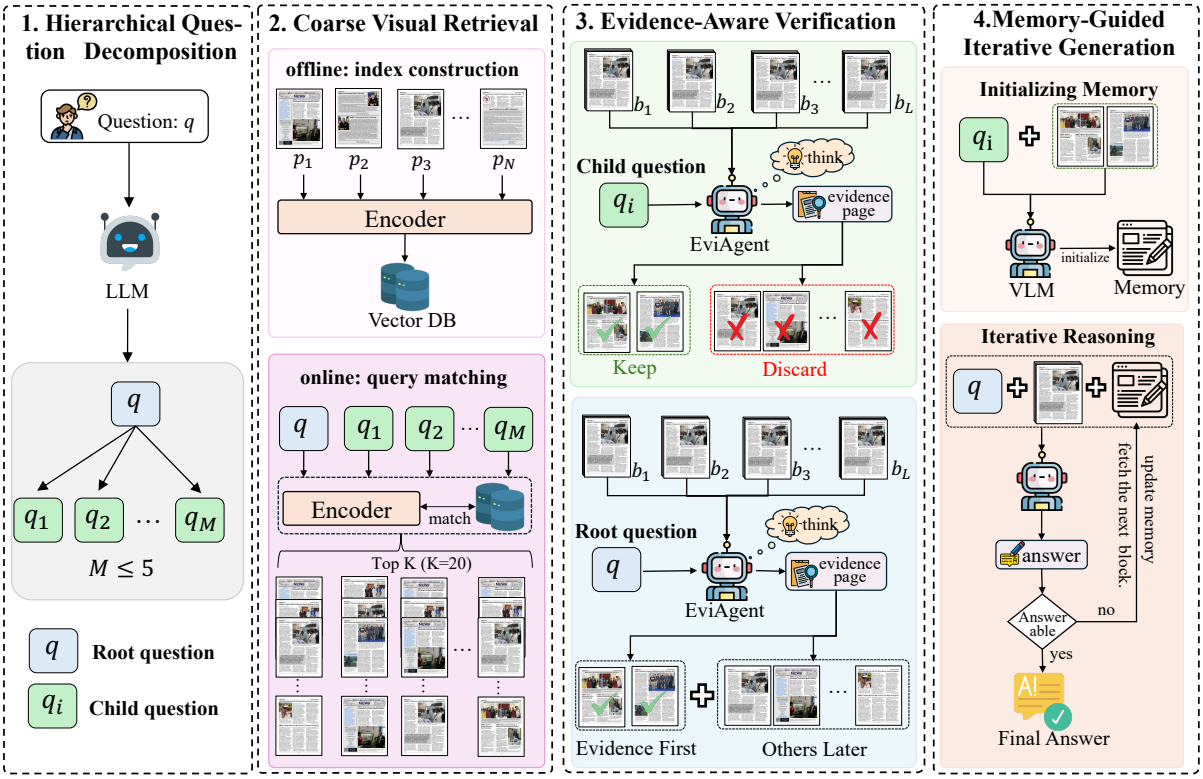

Figure 1: Overview of HIEVI-RAG. The system first builds a shallow hierarchy from the original question to atomic child questions, retrieves Top-K candidate pages for both root and child questions with a visual retriever, verifies and reranks grouped-k candidates with EVIAGENT, and performs memory-guided iterative reasoning over the verified evidence order.

takes the coarse candidate lists T_q and T_{q_i} as input, partitions them into non-overlapping blocks of size k, and leverages the trained agent to output structured decisions.

Training EVIAGENT with Evidence-Aware GRPO. Inspired by DocR1 (Xiong et al., 2025), we train EVIAGENT using an Evidence-Aware Group Relative Policy Optimization (EviGRPO) pipeline to specialize a VLM for multi-page document verification. The policy is optimized across six open-source multi-page datasets (Van Landeghem et al., 2023; Tito et al., 2022; Zhao et al., 2022; Schimanski et al., 2026; Yu et al., 2025a; Tanaka et al., 2023) with statistics detailed in Appendix E.

During the training phase, the VLM is constrained to generate a structured content containing three mandatory fields: $o = (\tau, \mathbf{y}, a)$. The corresponding prompt template is detailed in Appendix B.2.

- **Reasoning Trace** (τ): Contained within the <think> tag, τ encapsulates intra-page visual inspection and cross-page reasoning.
- **Evidence Decisions** ($\mathbf{y} \in \{T, F\}^k$): Enclosed within the <evidence_page> tag, $\mathbf{y}$ enforces a strict one-to-one binary sequence matching the exact order of input images to prevent evaluation omissions.
- **Final Answer** (a): Wrapped within the <answer> tag, a delivers the terminal text response derived from the verified evidence.

The policy is optimized via a weighted joint reward function:

$$R = \lambda_{\text{fmt}} r_{\text{fmt}} + \lambda_{\text{evi}} r_{\text{evi}} + \lambda_{\text{acc}} r_{\text{acc}}, \quad (4)$$

where r_{fmt} enforces schema adherence, and r_{acc} evaluates final answer correctness. The evidence reward r_{evi} computes the F_1 score against ground-truth page labels to accurately measure verification performance and prevent reward hacking.

Fine-Grained Page Validation. During online inference, each retrieved candidate list is split into $L = \lceil K/k \rceil$ blocks of size k: $b_t^k = \{p_{(t-1)k+1}, \ldots, p_{\min(tk,K)}\}$, where $t \in \{1, \ldots, L\}$ is the block index. Each block is verified jointly by the trained EVIAGENT. To prevent

information fragmentation across boundaries, all pages verified as true evidence are aggregated into a final consolidated block for a global cross-block synthesis. The refined signals are then routed to two distinct downstream policies:

- **Discarding for Child Questions**: For each child question q_i, we extract only the verified evidence pages:

$$E(q_i) = \{p \in T_{q_i} : \mathbf{y}_p = T\}. \quad (5)$$

  Non-evidence pages are discarded to restrict downstream generation strictly to localized, verified facts.

- **Reranking for the Root Question**: For the root question q, let $E(q)$ denote the verified evidence page set. We construct a prioritized sequence $\Pi(q)$ while preserving the original retriever ranking order within each partition:

$$\Pi(q) = E(q) \oplus (T_q \setminus E(q)), \quad (6)$$

  where $\oplus$ denotes sequence concatenation. Both $E(q)$ and the remaining subset $T_q \setminus E(q)$ strictly inherit the original ranking order determined by the coarse visual retriever. Non-evidence pages are retained as a backup suffix to secure recall for downstream iterative reasoning.

3.5 Memory-Guided Iterative Generation

This final stage leverages the resolved child questions and the prioritized sequence $\Pi(q)$ to generate the terminal answer via a two-step, memory-enhanced pipeline. As illustrated in Figure 1(4), the process consists of initializing an explicit memory layer with sub-question contexts and iteratively reasoning over the root question.

Initializing Memory with Sub-Questions. Before addressing the root query, the framework sequentially resolves each child question $q_i \in \mathcal{Q}(q)$ over its verified evidence block $E(q_i)$ using EVIAGENT:

$$(\tau_i, \mathbf{y}_i, a_i) = \mathcal{A}\big(q_i, E(q_i)\big), \quad (7)$$

where $\mathcal{A}$ denotes the generative policy of EVIAGENT. Initializing the memory as $M_0 = \emptyset$, the state updates cumulatively by appending the complete triplets:

$$M_i = M_{i-1} \cup \{(q_i, \tau_i, a_i)\}. \quad (8)$$

This mechanism distills sub-question verification into informative textual facts, seeding the working memory with both mid-level execution traces and terminal local answers.

Iterative Reasoning for the Root Question. Upon consolidating the sub-question memory M_m (where $m = |\mathcal{Q}(q)|$), HIEVI-RAG evaluates the root question q by scanning the prioritized sequence $\Pi(q)$ via a sliding window of size k. At execution round t, EVIAGENT ingests the root query, the persistent memory state, and the t-th evidence block $b_t^k \subset \Pi(q)$:

$$(\tau_t, \mathbf{y}_t, a_t) = \mathcal{A}\big(q, b_t^k, M_{m+t-1}\big). \quad (9)$$

The execution terminates immediately, returning a_t as the global response if $a_t \neq \alpha_\emptyset$, where $\alpha_\emptyset$ denotes the designated abstention token (i.e., `NOT_ANSWERABLE`). Otherwise, the current reasoning trace τ_t is appended to the memory to guide the next window:

$$M_{m+t} = M_{m+t-1} \cup \{(q, \tau_t)\}. \quad (10)$$

This iterative accumulation continues until a valid answer is produced or $\Pi(q)$ is exhausted. By maintaining text-based memory across rounds, this recurrent design forces the model to synthesize historical context globally, preventing critical information omission inherent in local window constraints.

4 Experiments

Benchmark	# QA Samples	Min. Pages	Max. Pages	Avg. Pages
PaperTab	393	2	152	10.7
FetaTab	1016	2	216	16.3
MMLongBench	1,082	9	468	47.8
LongDocURL	2,325	51	149	89.0

Table 1: Benchmark statistics for the four-dataset evaluation protocol. Average page counts are rounded to one decimal place.

4.1 Experimental Setup

Benchmarks. We evaluate HIEVI-RAG on four benchmarks: PaperTab and FetaTab (Hui et al., 2024), which target question answering over scientific and Wikipedia-style tables; and MMLongBench (Wang et al., 2025a) and LongDocURL (Deng et al., 2025), which feature extended multimodal documents stressing cross-page retrieval, visual grounding, and multi-hop reasoning. Structural statistics are provided in Table 1.

Method	Param.	PaperTab (Acc.)	FetaTab (Acc.)	MMLongBench (Acc.)	LongDocURL (Acc.)
M3DocRAG (Cho et al., 2024)	7B	28.5	63.8	36.2	49.0
MDocAgent (Han et al., 2025)	7B	30.0	66.3	38.5	46.9
MoLoRAG+ (Wu et al., 2025)	7B	31.0	69.2	41.0	51.9
ALDEN (Yang et al., 2026)	7B	24.5	62.3	39.2	55.1
URaG (Shi et al., 2026)	7B	–	–	33.8	52.2
Doc-$V^\star$ (Zheng et al., 2026)	7B	–	–	42.1	56.3
HIEVI-RAG	8B	**44.0**	**72.9**	**48.2**	**65.7**

Table 2: Main experimental results across the four benchmarks. The best results are highlighted in **bold**, and the second-best results are underlined. The symbol "–" indicates that the baseline result is not reported. ALDEN uses the ColQwen+ColBERT pipeline, and Doc-$V^\star$ refers to the GRPO-optimized model.

Variant	PaperTab	FetaTab	MMLongBench	LongDocURL	Avg.
HIEVI-RAG	**44.0**	**72.9**	**48.2**	**65.7**	**57.7**
w/o Evidence-Aware Verification	39.7 (-4.3)	69.6 (-3.3)	44.5 (-3.7)	59.1 (-6.6)	53.2 (-4.5)
w/o Hierarchical Decomposition	41.3 (-2.7)	70.5 (-2.4)	45.8 (-2.4)	62.4 (-3.3)	55.0 (-2.7)
w/o Iterative Reasoning	41.9 (-2.1)	72.0 (-0.9)	46.1 (-2.1)	63.2 (-2.5)	55.8 (-1.9)

Table 3: Ablation analysis of individual architectural modules. For *w/o Evidence-Aware Verification*, HIEVI-RAG's Step 3 is bypassed; sub-questions in Step 4 initialize memory directly using the unverified Top-K coarse retrieval, and the root question iteratively reasons over the un-reranked sequence. For *w/o Hierarchical Decomposition*, HIEVI-RAG's Step 1 and the memory initialization phase in Step 4 are omitted, while the iterative reasoning loop remains active. For *w/o Iterative Reasoning*, the sliding window in HIEVI-RAG's Step 4 is deactivated, force-feeding the Top-5 candidate pages into the generator in a single-pass manner.

Implementation Details. For hierarchical question decomposition, a lightweight `Qwen3.5-4B` model is deployed by default to generate at most five child questions. The baseline coarse retriever is configured as `ops-ColQwen3-4B` (OpenSearch-AI, 2026). Both root and child queries retrieve the Top-K ($K = 20$) candidate pages, which are subsequently verified using a grouped window size of $k = 7$. We instantiate EVIAGENT using `Qwen3-VL-8B-Instruct` (Qwen Team, 2025) as the foundational backbone. For the joint reward function, the balancing scaling weights are explicitly set as $(\lambda_{\text{fmt}}, \lambda_{\text{evi}}, \lambda_{\text{acc}}) = (0.1, 0.5, 0.4)$, prioritizing the evidence verification performance. Training is executed on an $8\times$ NVIDIA A100 GPU cluster. All reported HIEVI-RAG results are averaged over two runs with different random seeds. PDFs are rendered at 150 DPI; `Ops-ColQwen3-4B` uses 1,280-dimensional embeddings, while the VLM processes at most 512 visual tokens per image.

Metrics. For MMLongBench and LongDocURL, we follow their original evaluation protocols, employing standard accuracy and rule-based string matching tailored to various answer types. For PaperTab and FetaTab, we adopt an LLM-as-a-judge paradigm leveraging Qwen3.5-27B as the evaluator to compute binary accuracy (0 or 1) by assessing whether the generated response semantically matches the ground truth. Additionally, retrieval-related performance is quantified via Recall@K, NDCG@K, and MRR@K, with detailed mathematical formulations deferred to Appendix C. To further assess the robustness of LLM-based evaluation, we conduct an independent cross-evaluator validation using Gemini 2.5 Pro, with detailed results reported in Appendix D.

4.2 Main Results

Table 2 presents a performance comparison across the four benchmarks. HIEVI-RAG consistently outperforms all competitive baselines, validating the synergistic effects of its core architectural components.

On PaperTab and FetaTab, HIEVI-RAG achieves **44.0%** and **72.9%** accuracy, yielding absolute improvements of +13.0% and +3.7% over the strongest baseline (MoLoRAG+), respectively. Tabular documents typically require exact cell alignment; standard retrievers frequently introduce

Method	Generator	Retriever	PaperTab (Acc.)	FetaTab (Acc.)	MMLongBench (Acc.)	LongDocURL (Acc.)
MoLoRAG+ (Wu et al., 2025)	Qwen2.5-VL-7B-Instruct	ColQwen2.5	31.0	69.2	41.0	51.9
HIEVI-RAG	Qwen2.5-VL-7B-Instruct	ColQwen2.5	34.2	69.1	42.6	55.4
	Qwen2.5-VL-7B-Instruct	Ops-ColQwen3-4B	38.3	71.4	45.0	60.2
	Qwen3-VL-8B-Instruct	ColQwen2.5	36.0	70.8	43.8	57.1
	Qwen3-VL-8B-Instruct	Ops-ColQwen3-4B	**44.0**	**72.9**	**48.2**	**65.7**

Table 4: Sensitivity analysis over generation backbones and visual retrievers. MoLoRAG+ is included as the external baseline. The last row corresponds to the main HIEVI-RAG configuration in Table 2. The remaining HIEVI-RAG rows are controlled variants that change only the listed generator or retriever while keeping the same evidence-verification and iterative-generation pipeline.

false positives due to repetitive schema structures. HIEVI-RAG mitigates this bottleneck by executing fine-grained verification via EVIAGENT, effectively isolating answer-void distractors.

Similarly, on MMLongBench and LongDocURL, HIEVI-RAG secures **48.2%** and **65.7%** accuracy, outperforming Doc-$V^\star$ by +6.1% and +9.4%, respectively. These tasks necessitate cross-page retrieval and multi-hop reasoning, where single-pass pipelines easily omit critical information. The consistent performance gains demonstrate the efficacy of our memory-guided iterative generation.

4.3 Ablation Studies

Contribution of Individual Modules. To isolate the empirical impact of each architectural phase in HIEVI-RAG, we evaluate three ablation variants across all benchmarks (Table 3). Disabling any single stage yields a consistent performance degradation, validating the synergy of our core components.

Specifically, *w/o Evidence-Aware Verification* drops the average score by -4.5% absolutely. Without Step 3 verification, answer-void distractors occupy the front of the sequence, misleading the generator into premature early stopping and missing genuine evidence. Furthermore, initializing memory over unverified Top-K inputs introduces catastrophic cascading noise. Meanwhile, *w/o Hierarchical Decomposition* compromises the average score by -2.7%. This demonstrates that executing complex queries as standalone prompts without HIEVI-RAG's Step 1 induces severe retrieval omissions, whereas atomic decomposition guarantees the upfront recall of dispersed evidence. Lastly, *w/o Iterative Reasoning* consistently degrades results by -1.9% on average. This variant deactivates the sliding window in HIEVI-RAG's Step 4 and force-feeds the Top-5 pages in a single pass. The result proves that while verification serves as an effective filter, iterative reasoning acts as an indispensable safety net against information omission under local window constraints.

Backbone and Retriever Sensitivity. To evaluate the robustness of HIEVI-RAG under varying underlying architectures, we perform a grid sensitivity analysis by cross-combining different generative backbones and visual retrievers (Table 4).

Specifically, when deploying the identical backbone and retriever configuration as MoLoRAG+ (Qwen2.5-VL-7B-Instruct + ColQwen2.5), HIEVI-RAG outperforms this baseline in almost all benchmarks. It delivers clear absolute improvements of +3.2% on PaperTab and +3.5% on LongDocURL, demonstrating that our primary advantages stem strictly from architectural innovations rather than the scaling of foundational models. Furthermore, scaling up either the retriever to Ops-ColQwen3-4B or the generator to Qwen3-VL-8B-Instruct provides independent performance increments across benchmarks. The combination of both upgraded components yields the best performance, proving that our framework effectively leverages stronger retrieval and generation capabilities to maximize final accuracy.

Effectiveness of Fine-Grained Page Validation. To evaluate the impact of the verification stage on retrieval quality, we compare the retrieval metrics of the baseline against the reranked outputs generated by our verifier both before and after training under various Top-K settings on MMLongBench and LongDocURL (Table 5).

Specifically, deploying an untuned *Qwen3-VL-8B-Instruct* yields marginal or even detrimental performance shifts. While it provides negligible fluctuations on MMLongBench (e.g., a modest $+0.2$ point increase in NDCG@1 at Top-1), it

Top-K	Method	MMLongBench			LongDocURL		
		Recall	NDCG	MRR	Recall	NDCG	MRR
1	Ops-ColQwen3-4B (baseline)	54.2	70.4	70.4	53.5	73.4	73.4
	+ Qwen3-VL-8B-Instruct	54.2 (+0.0)	70.6 (+0.2)	70.6 (+0.2)	52.9 (−0.6)	72.1 (−1.3)	72.1 (−1.3)
	+ EviAgent	59.4 (+5.2)	78.3 (+7.9)	78.3 (+7.9)	57.3 (+3.8)	78.8 (+5.4)	78.8 (+5.4)
3	Ops-ColQwen3-4B (baseline)	75.2	73.8	77.2	74.8	73.7	81.1
	+ Qwen3-VL-8B-Instruct	75.6 (+0.4)	74.4 (+0.6)	77.6 (+0.4)	73.7 (−1.1)	72.5 (−1.2)	79.8 (−1.3)
	+ EviAgent	80.1 (+4.9)	80.6 (+6.8)	83.7 (+6.5)	77.5 (+2.7)	77.3 (+3.6)	84.9 (+3.8)
5	Ops-ColQwen3-4B (baseline)	81.4	76.2	78.2	81.0	76.4	82.0
	+ Qwen3-VL-8B-Instruct	81.8 (+0.4)	76.6 (+0.4)	78.5 (+0.3)	80.2 (−0.8)	75.4 (−1.0)	80.7 (−1.3)
	+ EviAgent	85.2 (+3.8)	82.2 (+6.0)	84.4 (+6.2)	83.2 (+2.2)	79.7 (+3.3)	85.6 (+3.6)

Table 5: Retrieval performance on MMLongBench and LongDocURL under different Top-K settings. Values in parentheses denote the absolute difference compared with the Ops-ColQwen3-4B baseline under the same Top-K setting.

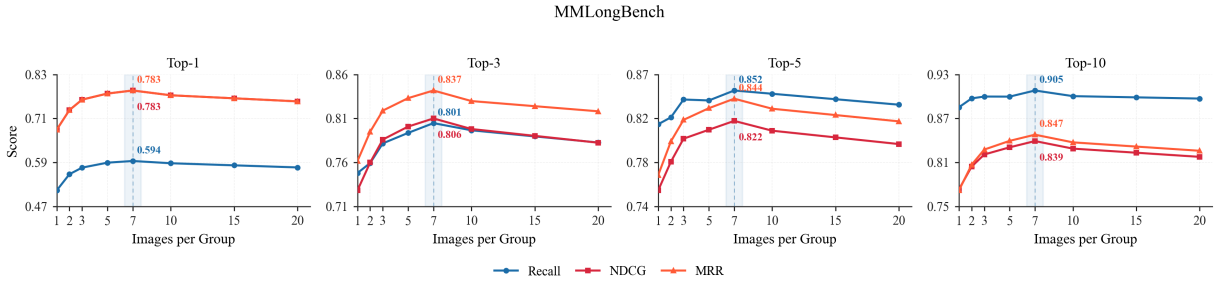

Figure 2: Effect of block size k on MMLongBench retrieval quality. A moderate block size provides the best balance between cross-page comparison and visual-context overload.

consistently degrades retrieval precision on LongDocURL, with NDCG slipping by −1.3 points at Top-1 and −1.0 points at Top-5. This demonstrates that off-the-shelf VLMs lack the necessary alignment to distinguish authentic evidence from complex multimodal distractors. In sharp contrast, our trained EviAgent consistently and substantially outperforms the *baseline*. It elevates the NDCG@1 from 70.4% to 78.3% on MMLongBench, and from 73.4% to 78.8% on LongDocURL. Furthermore, at Top-5, EviAgent pushes the final Recall to 85.2% and 83.2% on the respective benchmarks. These comprehensive gains across Recall, NDCG, and MRR solidly confirm EviAgent's superior capacity in identifying authentic evidence and filtering out answer-void noise.

Sensitivity to Block Size. To evaluate the impact of the block size k defined in HiEvi-RAG's Step 3, we analyze the retrieval quality under varying scales across MMLongBench (Figure 2). The empirical results indicate that both excessively large and excessively small values of k degrade retrieval quality.

Specifically, a moderate block size (e.g., $k = 7$) achieves the optimal balance and yields the highest retrieval metrics. Setting k too small isolates candidate pages. This constrains the model's capacity for cross-page reasoning and comparison. Conversely, expanding k excessively degrades performance. This drop is driven by visual context overload, where the excessive input scale dilutes model attention. Therefore, maintaining a tightly bounded block size k is essential to maximize page validation accuracy.

5 Conclusion

In this work, we presented HiEvi-RAG, a hierarchical, evidence-driven multimodal RAG framework for closed-domain long-document understanding, whose core innovation lies in its cooperative four-stage pipeline. Specifically, the framework first employs hierarchical question decomposition to partition complex multi-hop queries into atomic child questions, drastically reducing initial retrieval difficulty. Second, coarse visual retrieval leverages rendered page indexing to guarantee upfront evidence recall. Third, fine-grained page verification utilizes the EviAgent to perform cross-page reasoning over multi-image blocks, precisely suppressing topically related yet answer-void distractors. Fourth, memory-guided iterative generation accumulates sub-question traces into

an explicit text-based memory, dynamically executing sliding-window reasoning to eliminate cascading information omissions. Extensive experiments across four benchmarks demonstrate that HIEVI-RAG significantly outperforms existing open-source baselines. Furthermore, comprehensive ablation studies analyses solidly validate the individual efficacy and systemic synergy of each cooperative phase within our pipeline.

Limitations

Despite the state-of-the-art performance achieved by HIEVI-RAG, several inherent limitations warrant further investigation.

First, our framework introduces non-trivial computational overhead during the inference stage. Specifically, the hierarchical question decomposition, the multi-block page verification via the EVIAGENT, and the subsequent memory-guided generation inherently mandate multiple sequential LLM/VLM invocations. This compound multi-stage architecture inevitably increases cumulative response latency and inference costs compared to standard single-pass pipelines.

Second, the memory-guided generation mechanism relies heavily on the quality of textual intermediate reasoning traces; if the initial sub-questions yield severely hallucinated or biased summaries, such errors may cascade into the persistent memory layer and occasionally misguide the terminal root-question synthesis.

Third, our framework was optimized and evaluated predominantly on born-digital electronic documents with pristine rendering. Its zero-shot generalization capabilities across real-world handwritten inputs, heavily degraded physical scans with text artifacts, or scene-text images captured in natural environments require broader empirical validation in future work.

Ethical Considerations

HIEVI-RAG is designed for closed-domain long-document understanding and may be applied to documents containing sensitive or proprietary information. All experiments in this paper are conducted on publicly available benchmarks and open-source datasets. We do not collect new user data or perform human-subject annotation. Nevertheless, deployment in real-world enterprise or financial scenarios should ensure proper access control, data privacy protection, and auditing of generated answers. Because the system may still produce incorrect or unsupported answers when retrieval or evidence verification fails, we recommend using the framework as an assistive tool rather than an autonomous decision-making system in high-stakes domains.

References

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. 2024. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 709–718.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Xuhui Jiang, and Jian Guo. 2025. VLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8140–8158, Suzhou, China. Association for Computational Linguistics.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2025. M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9233–9250, Suzhou, China. Association for Computational Linguistics.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3DocRAG: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint 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. 2025. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1135–1159, Vienna, Austria. Association for Computational Linguistics.

Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Junyu Xiong, Mengyu Zhou, Yun Lin, José Cambronero, Yeye He, Shi Han, and Dongmei Zhang. 2024. Encoding spreadsheets for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20728–20748.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2023. DocPedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*.

ShunLiang Fu, Yanxin Zhang, Yixin Xiang, Xiaoyu Du, and Jinhui Tang. 2026. DMAP: Human-aligned structural document map for multimodal document understanding. In *Proceedings of the ACM Web Conference 2026*.

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

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120.

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

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. UDA: A benchmark suite for retrieval augmented generation in real-world document analysis. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*.

Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Shengyu Dai, Zhenwen Shao, Qingyun Wu, and Huazheng Wang. 2025. SimpleDoc: Multi-modal document understanding with dual-cue page retrieval and iterative refinement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28410–28427, Suzhou, China. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2026. Textmonkey: An ocr-free large multimodal model for understanding document. *IEEE transactions on pattern analysis and machine intelligence*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. LayoutLLM: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2023. KOSMOS-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Ahmad Mohammadshirazi, Pinaki Prasad Guha Neogi, Ser-Nam Lim, and Rajiv Ramnath. 2024. DLaVA: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness. *arXiv preprint arXiv:2412.00151*.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. 2025. DocVLM: Make your VLM an efficient reader. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29005–29015.

OpenSearch-AI. 2026. Ops-Colqwen3: State-of-the-Art Multimodal Embedding Model for Visual Document Retrieval. `https://huggingface.co/OpenSearch-AI/Ops-Colqwen3-4B`.

Qwen Team. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Tobias Schimanski, Imene Kolli, Yu Fan, Ario Saeid Vaghefi, Jingwei Ni, Elliott Ash, and Markus Leippold. 2026. pdfQA: Diverse, challenging, and realistic question answering over PDFs. *arXiv preprint arXiv:2601.02285*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yongxin Shi, Jiapeng Wang, Zeyu Shan, Dezhi Peng, Zening Lin, and Lianwen Jin. 2026. Urag: Unified retrieval and generation in multimodal llms for efficient long document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 25357–25365.

Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. 2025. DocAgent: An agentic framework for multi-modal long-context document understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17701–17716, Suzhou, China. Association for Computational Linguistics.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. VisDoM: Multi-document QA with

visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6088–6109, Albuquerque, New Mexico. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. SlideVQA: A dataset for document visual question answering on multiple images. *arXiv preprint arXiv:2301.04883*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19254–19264. IEEE.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. Hierarchical multimodal transformers for multi-page DocVQA. *arXiv preprint arXiv:2212.05935*.

Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. 2023. Document understanding dataset and evaluation (DUDE). *arXiv preprint arXiv:2305.08455*.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. DocLLM: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Yonghui Wang, Shi-Yong Chen, Zhenxing Zhou, Siyi Li, Haoran Li, Wengang Zhou, and Houqiang Li. 2024b. Root: Vlm based system for indoor scene understanding and beyond. *arXiv preprint arXiv:2411.15714*.

Yonghui Wang, Shaokai Liu, Li Li, Wengang Zhou, and Houqiang Li. 2024c. Swinshadow: Shifted window for ambiguous adjacent shadow detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11):1–20.

Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. 2023a. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*.

Yonghui Wang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. 2022. Udoc-gan: Unpaired document illumination correction with background light prior. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5074–5082.

Yonghui Wang, Wengang Zhou, Yunyao Mao, and Houqiang Li. 2023b. Detect any shadow: Segment anything for video shadow detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3782–3794.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. 2025a. MMLongBench: Benchmarking long-context vision-language models effectively and thoroughly. *arXiv preprint arXiv:2505.10610*.

Zining Wang, Tongkun Guan, Pei Fu, Chen Duan, Qianyi Jiang, Zhentao Guo, Shan Guo, Junfeng Luo, Wei Shen, and Xiaokang Yang. 2025b. Marten: Visual question answering with mask generation for multi-modal document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. 2025. MoLoRAG: Bootstrapping document understanding via multi-modal logic-aware retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14024–14045, Suzhou, China. Association for Computational Linguistics.

Shiyu Xia, Junyu Xiong, Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Mengyu Zhou, Yeye He, Shi Han, and Dongmei Zhang. 2024. Vision language models for spreadsheet understanding: Challenges and opportunities. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 116–128.

Junyu Xiong, Yonghui Wang, Weichao Zhao, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. 2025. DocR1: Evidence page-guided GRPO for multi-page document understanding. *arXiv preprint arXiv:2508.07313*.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. 2025. MM-R5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*.

Tianyu Yang, Terry Ruas, Yijun Tian, Jan Philip Wahle, Daniel Kurzawe, and Bela Gipp. 2026. Alden: Reinforcement learning for active navigation and evidence gathering in long documents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13371–13392.

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

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

Wenhan Yu, Zhaoxi Zhang, Wang Chen, Guanqiang Qi, Weikang Li, Lei Sha, Deguo Xia, and Jizhou Huang. 2025a. SciEGQA: A dataset for scientific evidence-grounded question answering and reasoning. *arXiv preprint arXiv:2511.15090*.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Cheng Yang, Jiangning Zhang, Shuicheng Yan, and Xiaobin Hu. 2025b. Visual document understanding and reasoning: A multi-agent collaboration framework with agent-wise adaptive test-time scaling. *arXiv preprint arXiv:2508.03404*.

Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 2024. TaPERA: Enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12824–12840, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yuanlei Zheng, Pei Fu, Hang Li, Ziyang Wang, Yuyi Zhang, Wenyu Ruan, Xiaojin Zhang, Zhongyu Wei, Zhenbo Luo, Jian Luan, Wei Chen, and Xiang Bai. 2026. Doc-V*: Coarse-to-fine interactive visual reasoning for multi-page document VQA. *arXiv preprint arXiv:2604.13731*.

Zhaoqing Zhu, Chuwei Luo, Zirui Shao, Feiyu Gao, Hangdi Xing, Qi Zheng, and Ji Zhang. 2025. A simple yet effective layout token in large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

A Detailed Related Work

A.1 Document Visual Question Answering

Recent advancements in DocVQA primarily diverge into OCR-based and OCR-free paradigms (Wang et al., 2022, 2023b, 2024a,b,c, 2023a). OCR-based methods serialize document pages into structured textual representations to serve as inputs for downstream reasoning. Representative architectures leverage multimodal pretraining or specialized attention mechanisms to fuse textual, visual, and layout features, as exemplified by UDOP (Tang et al., 2023), DocFormerv2 (Appalaraju et al., 2024), and DocLLM (Wang et al., 2024a). To enhance efficiency and structured generation, LayoutLLM (Luo et al., 2024) frames layout-sensitive reasoning as an instruction-tuning problem, whereas LayTokenLLM (Zhu et al., 2025) employs compact layout tokens to represent spatial bounding boxes without expanding lengthy coordinate sequences. For structured data, TaPERA (Zhao et al., 2024) incorporates content planning and execution-based reasoning to improve faithfulness in long-form table QA. Although highly effective when structural parsing is reliable, the performance of these methods remains intrinsically upper-bounded by cascading parsing errors and the inevitable loss of visually-situated evidence during text serialization.

OCR-free methods bypass external parsing pipelines by training VLMs to comprehend document images end-to-end. Early foundational architectures prioritized enhancing reading capabilities under constrained token budgets; within this cohort, UReader (Ye et al., 2023) introduces auxiliary reading and key-point generation tasks, KOSMOS-2.5 (Lv et al., 2023) pretrains on text-intensive images for structured text generation, DocPedia (Feng et al., 2023) processes visual inputs in the frequency domain, TextMonkey (Liu et al., 2026) incorporates high-resolution modeling with token filtering, and the mPLUG-DocOwl series (Ye et al.; Hu et al., 2024, 2025) employs feature compression to scale from single-page to multi-page document understanding. More recently, frameworks have evolved toward specialized reasoning and optimization protocols. Specifically, GOT (Wei et al., 2024) frames vision-based reading as a unified OCR-free generator, whereas DLaVA (Mohammadshirazi et al., 2024) and Marten (Wang et al., 2025b) implement visual answer localization and mask prediction to enhance model interpretability. To boost efficiency, DocVLM (Nacson et al., 2025) uses compact textual queries to lower visual token costs, while MACT (Yu et al., 2025b) orchestrates collaborative multi-agent reasoning at test time. To shift from passive reading to active pol-

icy optimization, DocR1 (Xiong et al., 2025) and Doc-$V^\star$ (Zheng et al., 2026) leverage evidence-guided GRPO to explicitly train visual reasoning trajectories for multi-page environments.

A.2 Retrieval-Augmented Generation

RAG methodologies have evolved from traditional text passage retrieval to multi-modal and structure-aware document discovery. Following the text-centric foundations laid by passage retrieval (Lewis et al., 2020) and token-level late interaction (Khattab and Zaharia, 2020), visual RAG systems have increasingly focused on page-level document interfaces. ColPali (Faysse et al., 2025) extends late interaction to rendered page images, while M3DocRAG (Cho et al., 2024) establishes this approach as a core interface for multi-page QA. To decouple retrieval from reading, MDocAgent (Han et al., 2025) and SimpleDoc (Jain et al., 2025) utilize multi-module collaboration and iterative cascading, while RagVL (Chen et al., 2025) and MM-R5 (Xu et al., 2025) employ instruction-tuning or reinforcement learning to train specialized visual rerankers. Concurrently, efforts have been dedicated to structured or active navigation to handle complex layouts: MoLoRAG (Wu et al., 2025) leverages page topology graphs for logic-aware traversal, DocAgent (Sun et al., 2025) and ALDEN (Yang et al., 2026) introduce memory feedback and active exploration protocols, and DMAP (Fu et al., 2026) constructs human-aligned structural maps to guide global understanding.

B Prompt Templates

B.1 Prompt For Hierarchical Question Decomposition

The system prompt for hierarchical question decomposition is detailed in Figure 3. Given the root question as input, the model is strictly constrained to output a maximum of five atomic child questions, each formatted strictly as an interrogative sentence.

B.2 Prompt For EviAgent

The system prompt used for both training and inference of EviAgent is provided in Figure 4. Given a question and a set of input pages, EviAgent is required to produce structured reasoning, make a binary evidence decision for each page, and generate the final answer.

C Retrieval Metrics

Let G_q denote the gold evidence-page set for question q, and let $R_q^K = (r_1, \ldots, r_K)$ denote the ranked top-K retrieved pages. We report three standard retrieval metrics.

Recall@K. Recall@K measures the fraction of gold evidence pages recovered in the top-K results:

$$\mathrm{Recall@K}(q) = \frac{|G_q \cap R_q^K|}{|G_q|}. \tag{11}$$

Higher Recall@K indicates better evidence coverage.

NDCG@K. Normalized Discounted Cumulative Gain emphasizes retrieving gold pages early in the ranked list. With binary relevance labels $\mathrm{rel}_i = \mathbb{I}(r_i \in G_q)$, we compute

$$\mathrm{DCG@K}(q) = \sum_{i=1}^{K} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)}, \tag{12}$$

and normalize it by the ideal ranking:

$$\mathrm{NDCG@K}(q) = \frac{\mathrm{DCG@K}(q)}{\mathrm{IDCG@K}(q)}. \tag{13}$$

This metric rewards both correctness and ranking quality.

MRR@K. Mean Reciprocal Rank focuses on how early the first relevant page appears:

$$\mathrm{RR@K}(q) = \begin{cases} \frac{1}{\min\{i \mid r_i \in G_q\}}, & \text{if } G_q \cap R_q^K \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

MRR@K is the average reciprocal rank over all questions.

D Evaluator Robustness

To assess the robustness of our evaluation method, we independently re-evaluate the model predictions using Gemini 2.5 Pro. As shown in Table 6, the resulting scores closely align with those obtained under our original evaluation method, indicating that the reported performance gains are not sensitive to the choice of evaluator.

E Training Datasets

Table 7 details the statistics of the multi-page training data utilized for EviAgent. To ensure a balanced and comprehensive optimization, we uniformly sample 1,000 instances from each of the

System Prompt:
You are a retrieval query decomposition assistant.
Your task is to decompose a user question into multiple concise search queries that help retrieve documents containing the key information needed to answer the question.

Guidelines:
1. First identify all key elements in the question, including entities (people, organizations, locations), topics or concepts, numbers or statistics, dates or time constraints, and conditions or qualifiers.
2. After identifying these key elements, construct retrieval queries by using one or two of the identified elements to form concise and searchable queries.
3. Each query should focus on retrieving documents that mention those specific elements or their relationships.
4. Preserve all quantitative constraints, temporal conditions, and logical relationships from the original question.
5. Queries should be concise and phrased in a way that could realistically appear in documents.
6. Avoid redundant queries that express the same concept.
7. Keep the total number of generated queries within {MAX_QUERIES}.
8. Return JSON only in the exact format: {"queries": ["...", "..."]}
Do not include explanations, comments, or markdown.

Example:
Input: "What are the views of 20% of the people in California regarding autonomous vehicles in 2023?"
Output: {"queries": ["California residents views on autonomous vehicles", "Where 20% of California residents are mentioned?", "Public opinion on autonomous driving in 2023"]}

Figure 3: System prompt for hierarchical question decomposition.

Evaluation	PaperTab	FetaTab	MMLongBench-Doc	LongDocURL
Original Evaluation	44.0	72.9	48.2	65.7
Gemini 2.5 Pro	44.6	73.5	49.1	66.1

Table 6: Cross-evaluator validation using Gemini 2.5 Pro. Results are reported in accuracy (%).

		Input Page			Evidence Page		
Dataset	# Samples	Min	Max	Avg	Min	Max	Avg
DUDE	1000	2	20	6.93	1	8	1.16
MP-DocVQA	1000	2	20	6.94	1	1	1.00
MultiHiertt	1000	3	7	4.06	1	3	1.48
pdfQA	1000	4	10	9.95	1	10	1.42
SciEGQA	1000	2	20	11.42	1	2	1.50
SlideVQA	1000	20	20	20.00	1	3	1.40

Table 7: Training statistics for EVIAGENT.

six constitutive datasets, with all selected samples inherently featuring multi-page visual contexts:

- **DUDE** (Van Landeghem et al., 2023) and **MP-DocVQA** (Tito et al., 2022) supply rich structural information from real-world complex electronic layouts and scanned industry documents.
- **MultiHiertt** (Zhao et al., 2022) and **pdfQA** (Schimanski et al., 2026) inject demanding financial reports and scientific literature requiring cross-page reasoning over hierarchical tables and textual narratives.
- **SciEGQA** (Yu et al., 2025a) provides dense academic papers with tightly coupled illustrations, charts, and mathematical proofs.
- **SlideVQA** (Tanaka et al., 2023) introduces sequential presentation slides characterized by sparse, highly stylized cross-page visual elements.

We additionally verify that there is no sample overlap between the EVIAGENT training data and the evaluation sets used in our four benchmarks.

F Case Study

Figure 5 illustrates a concrete execution trace of our decomposition layer. The complex, multi-hop root query requires concurrent visual counting and cross-year mathematical synthesis. Directly retrieving pages using this intricate query typically misleads standard semantic retrievers. By partitioning the root query into three localized, atomic child questions, our framework effectively simplifies the retrieval objectives, ensuring all disjoint tabular and visual evidence pages are fully recalled without cascading omissions.

System Prompt:

You are a specialized multimodal model for fine-grained multi-image visual understanding and evidence image localization. You are an expert at carefully inspecting multiple images, understanding detailed visual and textual information within each image, and identifying which images contain evidence relevant to a given question.

You will be given some images along with a question. Carefully examine every image. For each image, perform detailed visual analysis, including text, layout, objects, diagrams, tables, and any other potentially informative elements. Clearly describe what information appears in each image and whether it may contribute to answering the question.

After analyzing all images, perform cross-image reasoning to determine which images contain relevant evidence related to the question. Some images may not directly answer the question but may still provide necessary supporting information. However, once the existing evidence is sufficient to answer the question directly and completely, do not select any additional images.

After identifying the evidence images, you must use only the information from these evidence images to further reason and derive the final answer to the question. Do not use information from non-evidence images. If the answer can be determined, output a concise answer using a short phrase or a few words only. If the answer cannot be determined based on the evidence images, output "Not Answerable".

Your output must contain exactly three parts. First, provide your reasoning process wrapped in <think></think>. Inside <think>, your reasoning MUST be structured into three clearly separated parts:

(1) Image-level analysis: analyze each image one by one.
(2) Cross-image reasoning: explain which images are relevant and why.
(3) Answer reasoning: based only on the selected evidence images, derive the final answer step by step.

Second, provide the evidence-image judgment wrapped in <evidence_page></evidence_page> as a comma-separated sequence of T or F in image order. The number of entries must exactly match the number of input images.

Third, provide the final answer wrapped in <answer></answer> following the rules above.
Do not output anything outside these three sections.

Figure 4: System prompt used for EviAgent training and inference.

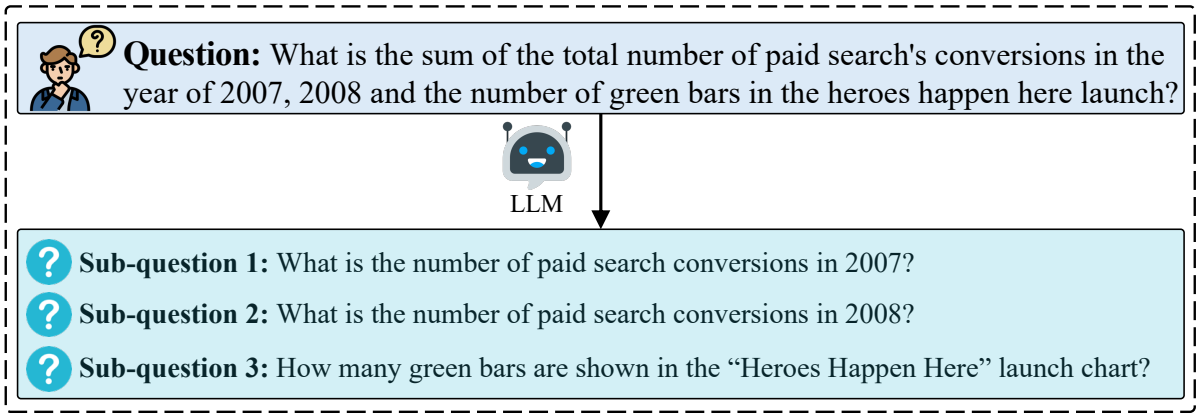

Figure 5: Case study of hierarchical question decomposition in HiEvi-RAG.