

E2E-VisRAG: End-to-End Visual Retrieval-Augmented Generation with Differentiable Retrieval

Zehua Wang^{1*}, Zhaojin Zhang^{1*}, Xiaolong Weng¹, Yanzhi Chen¹

Ying Xiong^{1†}, Buzhou Tang^{1,2,3†}

¹Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China

²Peng Cheng Laboratory, Shenzhen, China

³Guangdong Provincial Key Laboratory of Shenzhen University, Shenzhen, China
wangzehua@stu.hit.edu.cn, tangbuzhou@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) grounds large language models (LLMs) in external evidence, but existing RAG systems remain poorly matched to visually rich documents. OCR-based methods discard layout and non-textual visual cues, and although vision-language models (VLMs) can better understand native visual information, existing visual retrieval approaches often compress entire pages into global embedding vectors, obscuring fine-grained spatial information such as table cells and local annotations. Meanwhile, discrete Top- K selection in cascaded architectures blocks gradient flow between retrieval and generation, preventing joint optimization. In this work, we propose **E2E-VisRAG**, an end-to-end visual RAG framework that operates directly on native visual patches. Building upon patch-level visual encoding, we introduce a differentiable Top- K selection mechanism to enable joint training of retrieval and generation, and employ a gated residual fusion module to inject retrieval-enhanced visual features while suppressing noise. This design abandons the traditional paradigms of relying on OCR, global embeddings, and retrieval-generation separation, maximizing the preservation of original document visual semantics. Experiments on Vidore and MMLongBench-Doc show that, under matched 3B generator capacity, E2E-VisRAG outperforms non-end-to-end visual retrieval baselines, improving both Recall@1 and Accuracy@1 over OCR-based, global-embedding, and patch-level retrieval methods.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as an efficient paradigm for large language models to access external knowledge (Lewis

*Equal contribution, co-first author.

†Corresponding author.

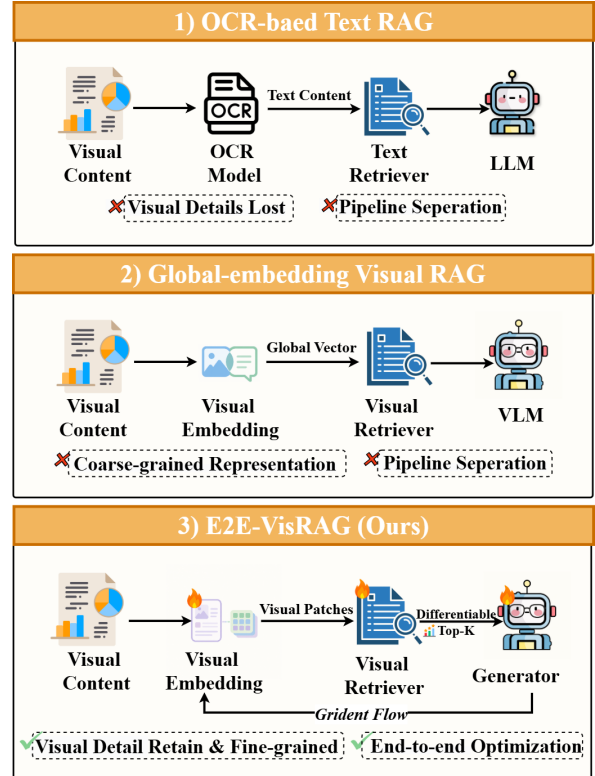


Figure 1: Comparison between existing visual RAG pipelines and our E2E-VisRAG.

et al., 2020; Guu et al., 2020). Meanwhile, in real-world scenarios, a substantial amount of high-value knowledge carriers do not exist solely in pure text form; they often heavily rely on visual structures such as layout and spatial arrangement, as well as non-textual elements like charts and images to convey core semantics. Consequently, Visual Retrieval-Augmented Generation (Visual RAG) has gradually attracted attention, which not only retrieves relevant page contents but also preserves visual information as evidence.

Most existing visual RAG systems inherit one of two design paradigms. The first is an OCR-based text-centric architecture, which parses document images into plain text before applying con-

ventional text RAG pipelines. The second is a visual retrieval model, which retrieves native visual features and delegates generation to a vision-language model (VLM). Despite improving multimodal knowledge processing, these approaches suffer from three structural limitations:

(1) Modality conversion. OCR-based pipelines serialize document images into text before retrieval. This conversion discards spatial layout, misses non-textual structures such as charts and diagrams, and the OCR parsing itself introduces additional computational overhead and error propagation risk. **(2) Coarse visual matching.** Global-embedding retrievers (Radford et al., 2021; Zhai et al., 2023) map each image into a single embedding via average pooling. Such representations are effective for image-level semantic matching, but they make it difficult to recover evidence localized in a table cell, a plot region, or a handwritten annotation. **(3) Retrieval-generation disconnect.** Because Top- K selection is discrete, gradients from the generation objective cannot update the retriever (Bengio et al., 2013; Lewis et al., 2020; Cao et al., 2025). As a result, the retriever is trained to rank documents under a proxy relevance objective rather than under the final answer-generation objective.

To address these challenges, we propose **E2E-VisRAG**, an end-to-end visual RAG framework¹. It processes raw images directly without OCR, eliminating information loss and computational overhead. Inspired by ColPali (Faysse et al., 2025), we adopt patch-level visual encoding with a token-level MaxSim late interaction for fine-grained local matching. A differentiable Top- K selection mechanism bridges the retrieval-generation gap by allowing generation gradients to backpropagate into retrieval, enabling true end-to-end optimization. Finally, a Gated Residual Fusion module injects retrieval-enhanced features into generation via a learned per-patch gate that amplifies evidence while suppressing noise, unlike fusion-in-decoder methods that prepend retrieved text (Izacard and Grave, 2021).

We evaluate E2E-VisRAG on two multimodal long-document benchmarks, Vidore (Faysse et al., 2025) and MMLongBench-Doc (Ma et al., 2024). Experimental results show that the framework outperforms OCR-based baselines, global-embedding visual retrieval, and patch-level visual retrieval

without end-to-end optimization, in terms of both retrieval recall and answer accuracy. Notably, the joint training mechanism enables generation supervision signals to directly optimize the retrieval representation space, significantly improving the synergy between retrieval and generation.

In summary, our main contributions are:

- We process raw visual information directly and achieve fine-grained retrieval via patch-level encoding with token-level MaxSim late interaction, eliminating OCR-induced information loss and global embedding’s coarse-grained limitation.
- We propose E2E-VisRAG, which bridges the retrieval-generation disconnect through a differentiable Top- K selection mechanism and employs a gated residual fusion module for noise-suppressed feature injection.
- We validate E2E-VisRAG on Vidore and MMLongBench-Doc, where it significantly outperforms OCR-based, global-embedding, and non-end-to-end patch-level baselines.

2 Related Work

2.1 Text-Based Retrieval-Augmented Generation

Early RAG systems retrieve text passages using sparse retrieval (BM25 (Robertson and Zaragoza, 2009)) or dense retrieval (DPR (Karpukhin et al., 2020)) and prepend them to the input query. Lewis et al. (2020) showed that jointly training retriever and generator improves open-domain QA; subsequent work explored iterative retrieval (Jiang et al., 2023), late-interaction scoring (Khattab and Zaharia, 2020), and token compression for long passages (Xu et al., 2024). These methods require all content to be expressed as token sequences, making them ill-suited for visually rich documents where layout and non-textual structures carry essential semantics. Moreover, the gradient disconnect between retriever and generator caused by discrete Top- K selection remains an open problem even in text-only pipelines (Shi et al., 2024). Recent text-based systems have begun to address multi-hop reasoning through retrieval-level RL (Wang et al., 2026a) and symbolic step verification (Wang et al., 2026b), yet visual RAG has not kept pace: it still struggles with fine-grained single-hop retrieval and end-to-end optimization.

¹Code available at <https://github.com/Vmyosotis/E2E-VisRAG>

2.2 Multimodal and Visual Retrieval-Augmented Generation

OCR-based methods. A dominant line of work parses documents with OCR before applying dense retrieval (Mathew et al., 2021; Tito et al., 2021). Zhang et al. (2025) show that OCR errors cascade through the pipeline, and text serialization inherently discards spatial layout and non-textual visual structures. Augmenting OCR with image captions partially recovers visual content, but captions are lossy summaries that routinely omit fine-grained details such as axis values and table cell entries.

Global-embedding visual retrieval. CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) project images and queries into a shared embedding space via average pooling. Compressing entire pages into single vectors precludes retrieval based on local visual evidence, and the frozen encoder provides no gradient path from the generation loss to the retrieval parameters.

Patch-level visual retrieval. ColPali (Faysse et al., 2025) encodes document pages as visual patch token sequences and scores them via MaxSim late interaction, achieving strong OCR-free recall. However, the retriever is trained independently of the generator, leaving a retrieval-generation disconnect and providing no mechanism to suppress retrieval noise during generation. Our framework extends ColPali by closing this loop via a differentiable Top- K operator and a Gated Residual Fusion module that enables stable joint optimization.

3 Method

As shown in Figure 2, the E2E-VisRAG framework consists of three stages: Visual Encoding, Differentiable Retrieval, and Gated Generation. The visual encoding stage encodes document images and text queries into dense visual patch features. The differentiable retrieval stage computes relevance via MaxSim (Khattab and Zaharia, 2020) and selects the most relevant patches through a differentiable Top- K selection mechanism. The gated generation stage fuses the raw and retrieval-optimized features of each retrieved document, then injects them into the generation pathway in a controlled residual manner, ultimately decoding the answer.

3.1 Visual Encoding

Given a text query q and a candidate document collection $\mathcal{D} = \{d_1, \dots, d_N\}$, where each document

d_i corresponds to a page image I_i , this stage first encodes the image into a sequence of visual patch tokens $\{v_1, \dots, v_S\}$ via a frozen vision encoder (ViT (Dosovitskiy et al., 2021)), then maps them to the language embedding space through a multimodal projector, yielding the original multimodal input embeddings:

$$H_{\text{orig},i} = \mathcal{F}_{\text{proj}}(\text{ViT}(I_i)) \in \mathbb{R}^{S \times D}, \quad (1)$$

where S is the number of visual tokens and D is the hidden dimension. Subsequently, the frozen backbone, equipped with trainable Retrieval LoRA adapters (Hu et al., 2022), contextualizes $H_{\text{orig},i}$ to produce retrieval-oriented document representations:

$$H_{\text{raw},i} = \mathcal{F}_{\text{ret}}(H_{\text{orig},i}) \in \mathbb{R}^{S \times D}, \quad (2)$$

where $\mathcal{F}_{\text{ret}}$ denotes the contextualization function induced by the frozen backbone parameters together with the Retrieval LoRA adapters. Each row of $H_{\text{raw},i}$ corresponds to a retrieval-optimized patch representation $h_j \in \mathbb{R}^D$ ($j = 1, \dots, S$), contextualized from the original patch token v_j in $H_{\text{orig},i}$. The query text q , after tokenization and embedding, is similarly processed through the Retrieval LoRA adapters to obtain retrieval-oriented query tokens:

$$Q = \{q_1, \dots, q_m\} = \mathcal{F}_{\text{ret}}(\text{Embed}(q)), \quad (3)$$

where $q_k \in \mathbb{R}^D$ ($k = 1, \dots, m$) and m is the number of query tokens.

3.2 Differentiable Retrieval

This stage computes relevance scores between query q and each candidate document $d_i \in \mathcal{D}$, and performs differentiable Top- K selection.

Relevance computation. We adopt the MaxSim late interaction mechanism. For each query token q_k , we compute its maximum inner-product similarity with all visual patch representations h_j of document d_i , then sum over query tokens:

$$s(q, d_i) = \sum_{k=1}^m \max_{j=1}^S (q_k^\top h_j), \quad (4)$$

where h_j denotes the j -th token representation in $H_{\text{raw},i}$, corresponding to the visual patch token v_j in $H_{\text{orig},i}$. This yields a score vector $s = (s_1, \dots, s_N)$ over the N candidate documents in $\mathcal{D}$, where $s_i = s(q, d_i)$.

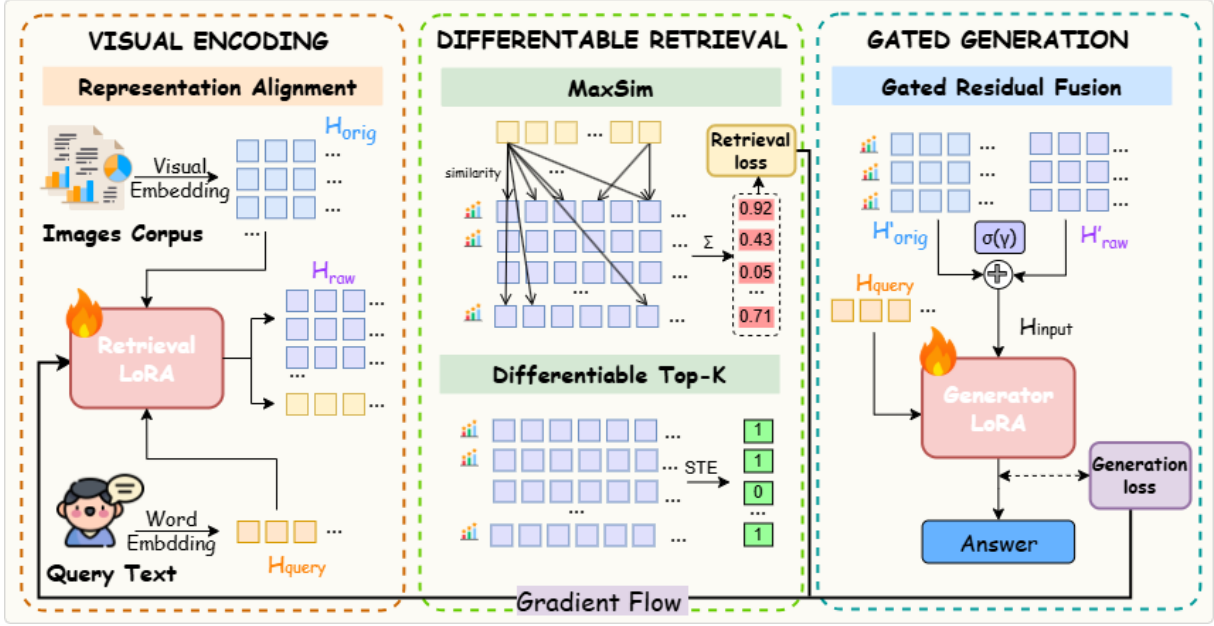


Figure 2: Overview of E2E-VisRAG. Given a query and document images, the framework encodes visual and textual inputs (Stage 1), differentially retrieves the top- K relevant documents via MaxSim scoring and STE-based selection (Stage 2), and generates the answer by injecting gated residual-fused document representations into the generation pathway (Stage 3).

Differentiable Top- K selection. To prevent discrete Top- K operations from blocking gradients, we employ sequential soft selection combined with the Straight-Through Estimator (STE) (Bengio et al., 2013). We first apply temperature scaling to obtain $\mathbf{z} = (z_1, \dots, z_N)$, where $z_i = s_i / \tau_{\text{sel}}$ and τ_{sel} is a temperature hyperparameter. We then perform K rounds of iterative selection: at round j , previously selected documents are masked out, and selection probabilities over the remaining documents are computed as:

$$\mathbf{p}^{(j)} = \text{softmax}(\mathbf{z} + \mathbf{M}^{(j-1)}) \in \mathbb{R}^N, \quad (5)$$

where $\mathbf{M}^{(j-1)} \in \mathbb{R}^N$ is a mask vector with $-\infty$ at selected positions and 0 elsewhere.

During the forward pass, we perform hard selection by picking the document with maximum probability:

$$\mathbf{K}_{\text{hard}}^{(j)} = \text{one_hot}\left(\arg \max_{i=1}^N p_i^{(j)}\right) \in \mathbb{R}^N. \quad (6)$$

During backpropagation, gradients flow through soft probabilities via the STE weight:

$$\mathbf{W}^{(j)} = \mathbf{K}_{\text{hard}}^{(j)} + (\mathbf{p}^{(j)} - \text{sg}(\mathbf{p}^{(j)})) \in \mathbb{R}^N, \quad (7)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator.

The retrieved representation at round j is computed as a weighted aggregation over all candidate documents:

$$H_{\text{ret}}^{(j)} = \sum_{i=1}^N W_i^{(j)} H_{\text{raw},i}, \quad (8)$$

where $H_{\text{ret}}^{(j)}$ corresponds to the retrieval-optimized representation of the document selected at round j . After K rounds, we obtain a set of K retrieved document representations $\{H_{\text{ret}}^{(1)}, \dots, H_{\text{ret}}^{(K)}\}$ for subsequent generation.

3.3 Gated Generation

Let $\mathcal{R} = \{d_{r_1}, \dots, d_{r_K}\}$ denote the set of K documents selected by the differentiable Top- K mechanism in §3.2. For each retrieved document $d_{r_k} \in \mathcal{R}$, we fuse its raw visual feature H_{orig,r_k} (direct ViT output) and retrieval-optimized feature H_{raw,r_k} (Retrieval LoRA output), then inject the fused representations into the generation pathway.

Gated Residual Fusion. For the i -th retrieved document, we first apply layer normalization to its raw feature $H_{\text{orig},i}$ and retrieval-optimized feature $H_{\text{raw},i}$:

$$\tilde{H}_{\text{orig},i} = \text{LN}_{\text{orig}}(H_{\text{orig},i}), \quad (9)$$

$$\tilde{H}_{\text{raw},i} = \text{LN}_{\text{raw}}(H_{\text{raw},i}). \quad (10)$$

The residual increment is computed as:

$$\Delta H_i = \tilde{H}_{\text{raw},i} - \tilde{H}_{\text{orig},i}. \quad (11)$$

After projection transformation:

$$\Delta H'_i = W_p \Delta H_i, \quad (12)$$

where $W_p \in \mathbb{R}^{D \times D}$ is a learnable projection matrix.

The final fused representation is:

$$H_{\text{input},i} = H_{\text{orig},i} + \sigma(\gamma) \cdot \Delta H'_i, \quad (13)$$

where $\gamma \in \mathbb{R}^D$ is a learnable gate vector and $\sigma(\cdot)$ denotes the sigmoid function, initialized to $\sigma(\gamma) \approx 0.2$. The gate controls the injection strength of retrieval-optimized information, allowing the model to preserve raw visual information while gradually learning to utilize retrieval-enhanced features.

Generation Decoding. The system prompt and query text are tokenized and embedded to obtain E_{sys} and E_{query} , respectively. We concatenate these with the K fused document representations, then decode through the Generation LoRA adapter $\mathcal{F}_{\text{gen}}$ to produce the answer:

$$H_{\text{final}} = \text{Concat}(E_{\text{sys}}, H_{\text{input}}, E_{\text{query}}), \quad (14)$$

$$\text{Answer} = \mathcal{F}_{\text{gen}}(H_{\text{final}}). \quad (15)$$

3.4 Training Pipeline

Training proceeds in two stages. In Stage 1, the vision encoder, multimodal projector, and language backbone are frozen; only the Retrieval LoRA adapters are trained with the contrastive objective $\mathcal{L}_{\text{retrieval}}$ to establish a meaningful patch-level representation space. The retrieval loss adopts an InfoNCE-style (Oord et al., 2018) contrastive form:

$$\mathcal{L}_{\text{retrieval}} = -\log \frac{\exp(s(q, d^+)/\tau_{\text{cl}})}{\sum_{d' \in \mathcal{B}} \exp(s(q, d')/\tau_{\text{cl}})}, \quad (16)$$

where d^+ denotes the positive document, $\mathcal{B}$ is the set of all documents in the current batch (including d^+ and in-batch negatives), τ_{cl} is the temperature parameter for contrastive learning, and $s(\cdot, \cdot)$ is the MaxSim score computed in Eq. (4).

In Stage 2, the vision encoder, multimodal projector, and language backbone remain frozen, while

all trainable parameters—Retrieval LoRA, Generation LoRA, and Gated Residual Fusion—are jointly optimized under the combined objective:

$$\mathcal{L} = \lambda_r \mathcal{L}_{\text{retrieval}} + \lambda_g \mathcal{L}_{\text{generation}}, \quad (17)$$

where λ_r and λ_g are scalar weighting coefficients controlling the contribution of retrieval and generation tasks. The generation loss is the standard autoregressive language modeling loss:

$$\mathcal{L}_{\text{generation}} = -\sum_t \log P(y_t | y_{<t}, H_{\text{final}}). \quad (18)$$

Generation gradients backpropagate through the differentiable Top- K selection mechanism into the retrieval pathway, achieving integrated optimization of retrieval and generation.

4 Experiments

We conduct experiments to answer the following research questions (RQ):

RQ1. Effectiveness: Does E2E-VisRAG outperform existing baselines on both retrieval and generation?

RQ2. Visual retrieval paradigm: (1) Can patch-level native visual retrieval preserve fine-grained visual information more effectively than OCR-based and global-embedding approaches? (2) Can end-to-end joint training further optimize patch-level retrieval beyond offline pre-training?

RQ3. Ablation study: How do end-to-end joint optimization, differentiable Top- K , and Gated Residual Fusion each contribute?

RQ4. Visual structure awareness: Does end-to-end visual processing enable spatial reasoning tasks (e.g., counting list items from layout) that are intractable under OCR serialization or frozen VLM inference?

4.1 Datasets and Evaluation

Datasets. We evaluate E2E-VisRAG on two representative multimodal long-document benchmarks. **Vidore** (Faysse et al., 2025) comprises five visual document types—infographics (Infographic-VQA (Mathew et al., 2022)), academic papers (arxiv-qa (Li et al., 2024)), general documents (DocVQA (Mathew et al., 2021)), tables (TAT-DQA (Zhu et al., 2022)), and PDF-style documents (PDF)—enabling evaluation across diverse visual layouts. **MMLongBench-Doc** (Ma et al., 2024) contains 1,091 in-document queries across 135

Method	Inf.		Arx.		DVQA		TAT		PDF		Overall		MMDoc-G		Vidore-R		MMDoc-R	
	A@1	A@5	A@1	A@5	A@1	A@5	A@1	A@5	A@1	A@5	A@1	A@5	A@1	A@5	R@1	R@5	R@1	R@5
OCR+BM25 _{3B}	15.0	11.7	47.0	52.9	13.7	18.0	20.3	21.6	33.0	38.2	25.3	29.2	11.2	14.4	31.0	44.6	21.2	44.0
OCR+BM25 _{7B}	14.4	15.0	27.5	29.4	13.7	17.4	21.9	25.4	34.4	36.4	24.0	26.6	9.5	15.6	31.0	44.6	21.2	44.0
OCR+BGE _{3B}	30.6	27.8	54.9	54.9	26.8	30.4	20.6	21.3	51.1	50.3	38.4	39.1	15.4	21.1	60.2	71.4	35.2	62.8
OCR+BGE _{7B}	33.3	27.8	49.0	54.9	25.7	30.6	26.4	27.9	52.0	52.7	38.6	40.9	14.9	21.4	60.2	71.4	35.2	62.8
OCR+Cap+BM25 _{3B}	33.3	36.1	60.8	58.8	20.5	21.1	17.5	15.9	40.7	48.2	32.8	35.6	13.7	18.2	47.4	62.6	31.2	57.5
OCR+Cap+BM25 _{7B}	27.8	36.1	54.9	54.9	19.3	19.9	20.6	26.0	49.4	46.9	35.1	35.6	13.5	18.3	47.4	62.6	31.2	57.5
OCR+Cap+BGE _{3B}	55.6	41.7	66.7	64.7	31.7	36.0	20.3	24.1	51.5	57.5	43.0	46.0	19.8	23.4	62.6	75.8	40.0	66.6
OCR+Cap+BGE _{7B}	52.8	47.2	60.8	60.8	30.4	35.4	23.8	27.0	58.4	58.0	44.9	46.3	19.7	24.2	62.6	75.8	40.1	66.6
SigLIP _{3B}	50.0	55.6	70.6	<u>80.4</u>	7.5	14.9	22.5	29.8	23.5	31.3	22.5	29.8	15.3	23.1	25.8	39.2	24.3	53.4
SigLIP _{7B}	66.7	<u>66.7</u>	58.8	66.7	11.2	17.4	3.2	11.8	24.6	33.1	24.1	31.2	16.4	27.9	25.8	39.2	24.3	53.4
ColPali _{3B}	<u>69.4</u>	61.1	80.4	74.5	59.0	60.9	25.4	<u>29.5</u>	60.9	62.4	58.4	58.9	24.8	30.1	<u>78.8</u>	<u>90.6</u>	<u>45.3</u>	<u>73.7</u>
ColPali _{7B}	77.8	69.4	70.2	73.7	<u>61.6</u>	<u>66.2</u>	<u>28.9</u>	<u>29.5</u>	66.8	63.4	<u>61.5</u>	<u>61.5</u>	28.3	36.1	<u>78.8</u>	<u>90.6</u>	<u>45.3</u>	<u>73.7</u>
Ours _{3B}	63.9	61.1	<u>78.4</u>	84.3	62.1	66.5	34.3	33.7	<u>65.3</u>	<u>63.0</u>	61.6	62.4	<u>26.6</u>	<u>30.8</u>	83.6	94.4	46.1	74.1

Table 1: Main retrieval and generation results on Vidore and MMLongBench-Doc. Generation quality is evaluated using Acc@1 (A@1) and Acc@5 (A@5), while retrieval performance is measured by Recall@1 (R@1) and Recall@5 (R@5). Methods with subscript “_{3B}” use Qwen2.5-VL-3B-Instruct as the generator, and those with subscript “_{7B}” use Qwen2.5-VL-7B-Instruct. Our proposed method is built upon the 3B model. **Bold**: best; underline: second best. All generation results are averaged over 5 runs.

long documents; 253 questions deemed unanswerable due to item errors are excluded, given the uncertainty of the 3B-parameter generator on such items. Training uses the full ColPali training set, with a 500-sample validation set.

Evaluation Metrics. Retrieval is measured by Recall@1 and Recall@5. Generation quality is assessed by Acc@1 and Acc@5: the Top-1/Top-5 retrieved images are fed to the generator, and GPT-4o-mini (OpenAI, 2024) serves as an LLM-as-Judge (Zheng et al., 2023) for binary correctness evaluation.

4.2 Baselines

We compare against three RAG families:

(1) **OCR-based text retrieval.** Document images are parsed by OCR, then indexed for sparse (BM25) or dense (BGE (Xiao et al., 2023), bge-large-en-v1.5) retrieval. *OCR+Caption* variants augment OCR text with VLM-generated captions before retrieval.

(2) **Global-embedding visual retrieval (SigLIP).** Queries and images are mapped to a shared embedding space via average pooling for image-level similarity.

(3) **Patch-level visual retrieval (ColPali).** Late-interaction scores over visual patch tokens achieve strong OCR-free recall, but its retriever is not co-trained with the generator.

All baseline methods use Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025) as generators and E2E-VisRAG is built upon the 3B model.

4.3 Results and Analysis

4.3.1 Effectiveness (RQ1)

Table 1 shows that E2E-VisRAG achieves the best retrieval performance on both benchmarks and the best generation performance on Vidore, while its generation on MMLongBench-Doc surpasses all 3B and OCR-based baselines (though it falls slightly behind ColPali_{7B}). Compared with ColPali_{3B} under matched generator capacity, E2E-VisRAG improves on both retrieval and generation across both datasets, demonstrating that end-to-end joint optimization yields better retrieval-generation synergy regardless of document domain. OCR-based methods remain substantially weaker than visual retrievers on all metrics; SigLIP’s poor retrieval performance further confirms that global embeddings are insufficient for fine-grained document retrieval.

Table 2 analyzes performance under different page scopes. All methods degrade noticeably on multi-page questions, yet E2E-VisRAG achieves the strongest results on both single-page and multi-page settings, with particularly significant gains on the multi-page subset. This indicates that end-to-end optimization not only improves retrieval precision, but also enhances the interaction between

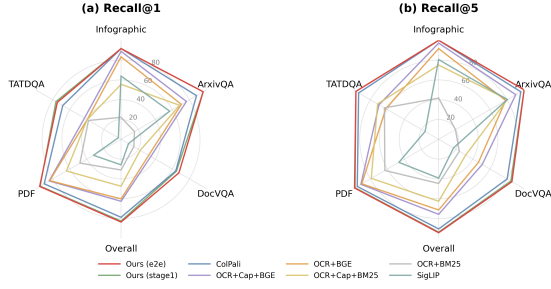


Figure 3: (a) Recall@1 and (b) Recall@5 across document types. Patch-level retrieval dominates OCR-based and global-embedding baselines; end-to-end optimization yields further gains.

retrieved evidence and downstream generation, benefiting long-document reasoning.

Method	Single-Page		Multi-Page	
	A@1	A@5	A@1	A@5
OCR+BGE _{3B}	22.43	28.08	6.17	11.89
OCR+Cap+BGE _{3B}	29.33	32.89	7.22	10.72
SigLIP _{3B}	21.63	31.80	6.89	11.44
ColPali _{3B}	35.02	39.54	11.28	17.44
Ours_{3B}	38.20	40.21	19.62	18.28

Table 2: Performance comparison on MMLongBench-Doc under different evidence-page scopes. Generation quality is evaluated by Acc@1 and Acc@5. All generation results are averaged over 5 runs.

4.3.2 Visual Retrieval Paradigm (RQ2)

Figure 3 compares retrieval performance across diverse document types.

First, **patch-level visual retrieval consistently outperforms both OCR-based and global-embedding paradigms**. ColPali achieves an average Recall@1 of 78.8%, surpassing all OCR variants (31.0–62.6%) and SigLIP (25.8%). Our Stage 1 retriever further improves to 82.6% (+3.8%), demonstrating the benefit of task-specific adaptation.

Second, **end-to-end joint optimization yields additional gains**. E2E-VisRAG reaches 83.6% Recall@1 (+1.0% over Stage 1), with the most pronounced improvement on DocVQA (+3.1%). The gains are consistent across both R@1 and R@5 (Figure 3a–b), indicating that joint training enhances retrieval quality without sacrificing recall breadth.

4.3.3 Ablation Studies (RQ3)

We ablate three core components: (1) **w/o E2E**: uses the Stage-1 retriever to retrieve raw images for a frozen Qwen2.5-VL-3B-Instruct generator;

Method	Vidore				MMLongBench-Doc			
	A@1	A@5	R@1	R@5	A@1	A@5	R@1	R@5
w/o E2E	60.84	60.32	82.60	93.80	26.21	29.16	45.53	74.08
w/o Diff. Top- <i>K</i>	61.40	61.20	83.20	93.80	26.01	29.12	45.58	73.80
w/o GRF	NA	NA	NA	NA	NA	NA	NA	NA
Full model	61.60	62.44	83.60	94.40	26.63	30.79	46.06	74.10

Table 3: Ablation study results. “w/o E2E”: the Stage-1 retriever and Qwen2.5-VL-3B-Instruct generator; “w/o Diff. Top-*K*”: standard hard Top-*K* selection, disabling gradient flow to the retriever; “w/o GRF”: no gated fusion. **Bold**: best; NA: not applicable.

(2) **w/o Diff. Top-*K***: standard hard Top-*K* selection replacing the differentiable selector, blocking gradient flow to the retriever; (3) **w/o GRF**: removing the Gated Residual Fusion module and directly concatenating retrieved features into generation.

End-to-end optimization. Removing joint training degrades all metrics: Vidore Acc@5 drops 2.12 points (62.44%→60.32%) and MMLong Acc@5 falls 1.63 points (30.79%→29.16%), while Recall@1 also decreases from 46.06% to 45.53%, confirming that generation gradients not only improve generation quality but also back-propagate to refine the retrieval representation space beyond decoupled training.

Differentiable Top-*K* selection. Removing the differentiable Top-*K* mechanism hurts both retrieval and generation performance: Vidore Recall@1 drops from 83.60% to 83.20%, MMLong Recall@1 drops from 46.06% to 45.58%, and MMLong Acc@5 falls 1.67 points (30.79%→29.12%). This indicates that without gradient flow to the retriever, evidence selection remains sub-optimal for generation, and the retrieval representation space lacks direct supervision from the generation task, leading to degraded retrieval precision.

Gated Residual Fusion. Directly concatenating retrieved features causes training to diverge (DNC). The gate mechanism is therefore critical: without controlled residual injection, noisy retrieved patches overwhelm the generation pathway, causing model divergence.

4.3.4 Visual Structure Awareness (RQ4)

Figure 4 compares three paradigms on a MMLongBench-Doc example requiring spatial counting of bullet-list items (gold answer: 10).

OCR RAG serializes the image into flat text, losing two entries, and answers 4. **VLM RAG** feeds retrieved images to a frozen generator, partially recognizing the structure and answering 6.

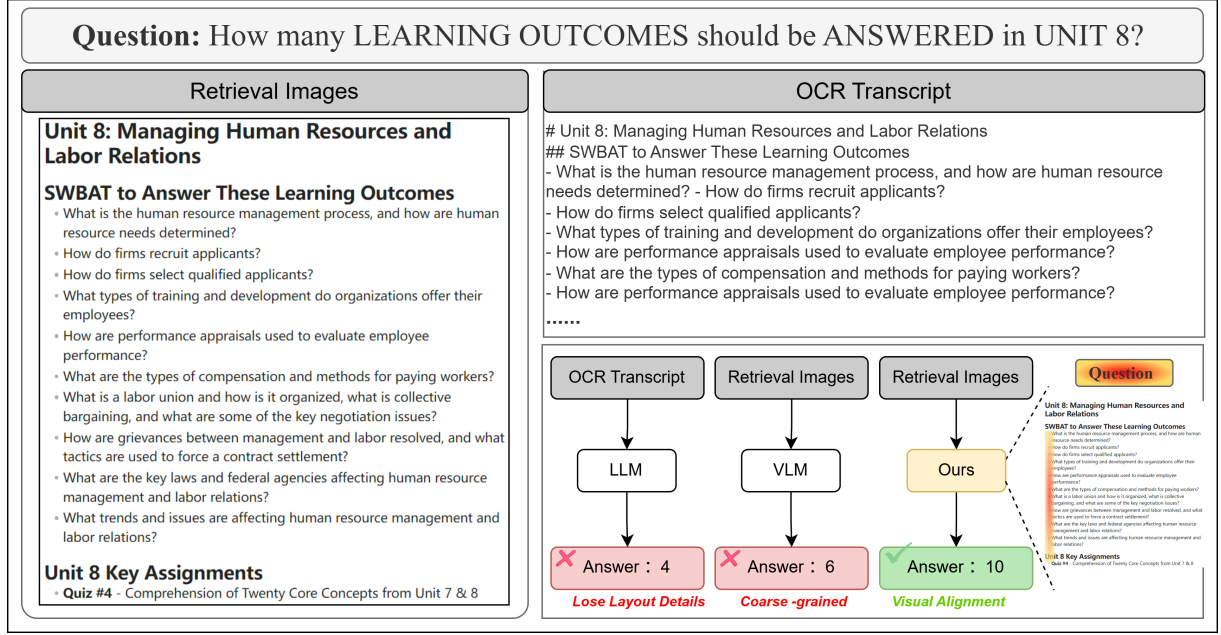


Figure 4: Case study on visual structure awareness (MMLongBench-Doc). **Top-left:** Retrieved image preserving native visual layout. **Top-right:** OCR transcript with destroyed item boundaries and bullet markers. **Bottom:** OCR RAG (Answer: 4); VLM RAG (Answer: 6); E2E-VisRAG (Answer: 10).

Method	Enc. (ms/pg)	Retr. (ms/q)	Gen. (ms/q)	Total (ms/q)	Acc@5 (%)
OCR+BGE _{3B}	2,491	26.2	903	3,420	29.20
OCR+Cap _{3B}	7,475	24.6	946	8,446	45.96
SigLIP _{3B}	39	5.5	1,488	1,533	31.20
ColPali _{3B}	79	160.7	1,461	1,701	<u>58.92</u>
Ours_{3B}	155	96.5	1,310	<u>1,562</u>	62.44

Table 4: Latency–accuracy trade-off. Encoding (per page), retrieval and generation (per query), total latency, and generation Acc@5 on Vidore.

E2E-VisRAG preserves the complete visual layout through native patch processing; the jointly trained generator correctly counts all ten items, validating that end-to-end optimization enables stronger spatial reasoning than OCR serialization or frozen inference.

4.4 Latency Analysis

Table 4 reports the latency–accuracy trade-off across methods. SigLIP achieves the lowest total latency (1,533 ms) yet its Acc@5 is only 31.20%, confirming that global embeddings sacrifice generation quality for speed. In contrast, E2E-VisRAG incurs a comparable latency (1,562 ms) while reaching the highest Acc@5 (62.44%), demonstrating that its performance advantage does not come at the cost of inference speed.

Method	Tokens	Vidore					MMLongBench-Doc				
		A@1	A@5	R@1	R@5		A@1	A@5	R@1	R@5	
ColPali _{3B}	784	58.40	58.92	78.80	90.60	24.82	30.05	45.26	73.72		
ColPali _{3B}	1024	61.20	61.20	78.80	90.60	25.89	31.15	45.26	73.72		
ColPali _{3B}	unlimited	<u>62.40</u>	59.20	78.80	90.60	26.01	<u>31.74</u>	45.26	73.72		
Ours_{3B}	784	61.60	<u>62.44</u>	<u>83.60</u>	<u>94.40</u>	<u>26.63</u>	30.79	<u>46.06</u>	<u>74.10</u>		
Ours_{3B}	1024	64.40	63.60	83.60	94.60	29.83	32.94	47.77	75.40		

Table 5: Effect of visual token budget on retrieval and generation performance. **Bold:** best overall; underline: second best; unlimited: no max visual tokens limitation.

4.5 Visual Token Budget Analysis

To ensure a fair comparison, we unify the number of visual tokens across all methods. Following ColPali (Faysse et al., 2025), we adopt 784 tokens as the default budget, applied consistently to all base-lines and our model. We additionally report results with a 1024-token budget to examine scalability.

Table 5 presents the results. Under the matched 784-token setting, E2E-VisRAG already outperforms the token-limited ColPali on every metric—Vidore Recall@1 improves from 78.80% to 83.60%, and Vidore Acc@5 rises from 58.92% to 62.44%—confirming that the gains stem from end-to-end joint optimization rather than a token-budget advantage.

Increasing our token budget to 1024 yields further improvements across both datasets (Vidore Acc@1: 61.60%→64.40%), indicating that E2E-

VisRAG benefits from richer patch-level representations. The 784-token configuration is retained as the primary setting throughout this paper and is adopted in other experiments.

5 Conclusion

We presented E2E-VisRAG, an end-to-end visual RAG framework that keeps document evidence in native visual token space, performs fine-grained MaxSim late-interaction retrieval, and uses a differentiable Top- K selector to pass generation gradients back into the retriever. A gated residual fusion module further controls how retrieved visual features enter the generation path. Under matched 3B generator capacity, experiments on Vidore and MMLongBench-Doc show that end-to-end joint optimization of retrieval and generation improves both recall and answer accuracy over OCR-based, global-embedding, and patch-level retrieval methods.

Limitations

Training memory and scalability. The differentiable Top- K selector incurs $O(N)$ memory overhead over the in-batch candidate set, and the InfoNCE contrastive loss demands large batch sizes for sufficient negatives, compounding memory pressure. Approximate nearest-neighbor indexing or memory bank mechanisms will be needed for open-domain deployment over large corpora.

Visual token budget and efficiency. Performance improves from 784 to 1024 tokens, yet larger budgets amplify MaxSim computation, selector memory, and generation sequence length. Whether token compression or adaptive-resolution strategies can close this efficiency gap remains unexplored.

Generator scale and module contribution. All experiments use Qwen2.5-VL-3B-Instruct; larger models may shift the relative contribution of Gated Residual Fusion or move the bottleneck from retrieval to generation, an interaction that warrants further study.

Language and domain scope. Current evaluation is limited to English-only benchmarks. Generalization to low-resource languages, non-Latin scripts, handwriting, or low-quality scans remains unverified, and conclusions may not transfer to video-grounded QA or specialized scientific domains.

Evaluation scope and judge reliability. GPT-4o-mini judgments may exhibit systematic bias on layout-sensitive or numerically precise answers. We mitigate this by using deterministic decoding and averaging over five trials, yet ANLS metrics or human annotation should be used to quantify judge error rates. Furthermore, document QA carries the inherent risk of generating unsupported or numerically incorrect answers even when relevant pages are retrieved. Additional safeguards such as answer verification or confidence calibration may be needed for high-stakes applications.

Acknowledgments

This study is partially supported by National Natural Science Foundation of China (62276082), Shenzhen Science and Technology Research and Development Fund (KJZD20240903102802003), Shenzhen Science and Technology Research and Development Fund for Sustainable Development Project (GXWD20231128103819001, 20230706140548006) and Guangdong Provincial Key Laboratory Grant (2023B1212060076).

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.
- Hongyu Cao, Yuxuan Wu, Yucheng Cai, Xianyu Zhao, and Zhijian Ou. 2025. Improving end-to-end training of retrieval-augmented generation models via joint stochastic approximation. *arXiv preprint arXiv:2508.18168*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- 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 *International Conference on Learning Representations*, volume 2025, pages 61424–61449.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7969–7992.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.
- 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: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yugang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. [MMLONGBENCH-DQC: benchmarking long-context document understanding with visualizations](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- OpenAI. 2024. [Gpt-4o system card](#). *CoRR*, abs/2410.21276.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International*

- Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384.
- Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document collection visual question answering. In *International Conference on Document Analysis and Recognition*, pages 778–792. Springer.
- Zehua Wang, Shipeng Li, and Buzhou Tang. 2026a. Cpsrch: A chain progressive search training framework incentivizing the cognitive behaviors for searching in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33755–33763.
- Zehua Wang, Zhaojin Zhang, Boyu Qiu, Xiaolong Weng, Ying Xiong, Buzhou Tang, and Min Zhang. 2026b. Sr-rag: Verifiable multi-hop reasoning via on-the-fly symbolic graph construction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 38588–38606.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *Preprint*, arXiv:2309.07597.
- Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Re-comp: Improving retrieval-augmented llms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pages 43478–43502.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17443–17453.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

A Dataset and Evaluation Details

A.1 Training Data

Our retrieval module is trained on the ColPali training set (Faysse et al., 2025), which comprises 118,695 query–page pairs. The dataset combines openly available academic benchmarks (63%) and synthetic data (37%) generated from web-crawled PDF documents with VLM-generated (Claude-3 Sonnet) pseudo-questions. All training data is English-only by design, enabling zero-shot generalization study to non-English languages. Table 6 provides the composition breakdown.

Source	# Pairs	Language
DocVQA	39,463	English
InfoVQA	10,074	English
TAT-DQA	13,251	English
arXiv-QA	10,000	English
Web-crawled PDFs (synthetic)	45,940	English
Total	118,695	English-only

Table 6: Composition of the ColPali training set.

Data Contamination Prevention. We explicitly verify that no multi-page PDF document appears in both the training set and the Vidore evaluation suite, preventing evaluation contamination.

A.2 Evaluation Data

Vidore. The Vidore benchmark suite (Faysse et al., 2025) comprises five visual document types: Infographic-VQA (Info.), Arxiv-QA (Arxiv.), DocVQA, TAT-DQA (TAT), and PDF-style documents (PDF). The official test set is designed solely for retrieval evaluation and lacks answer annotations; therefore, we use the validation set (500 query–page pairs with ground-truth answers) for both retrieval and generation assessment.

MMLongBench-Doc. MMLongBench-Doc contains 1,091 in-document queries across 135 long documents, averaging 48.3 pages per document. Following the official protocol, we exclude 253 questions deemed unanswerable due to item errors or missing annotations, as the 3B-parameter generator exhibits high uncertainty on such items. The remaining 838 answerable questions are used for evaluation, of which 478 (57.0%) are single-page and 360 (43.0%) are multi-page.

Single-Page vs. Multi-Page Split. All Vidore queries are inherently single-page, with each query

annotated to a single gold page. MMLongBench-Doc queries are categorized by evidence scope:

- **Single-Page:** The answer can be inferred from a single document page (478 questions, 57.0%).
- **Multi-Page:** The answer requires cross-page reasoning or aggregation (360 questions, 43.0%).

A.3 Evaluation Metrics

Retrieval. Recall@ K ($R@K$) measures the proportion of queries where at least one relevant document appears in the Top- K retrieved pages. We report $R@1$ and $R@5$.

Generation. Acc@ K ($A@K$) measures answer accuracy when the Top- K retrieved pages are provided as context. All generation experiments are run with 5 random seeds, and we report the mean.

LLM-as-Judge. Answer correctness is evaluated by GPT-4o-mini serving as an automated judge. The judge is prompted with the following system instruction:

Prompt Template of Answer Verification by LLM

You are a strict but fair evaluation judge for visual document QA.

Your task: decide whether a predicted answer is CORRECT given the question and gold answer.

Rules:

1. Answer ONLY with JSON: {"correct": 1} or {"correct": 0}
2. Mark correct=1 if prediction conveys same key information as gold, even if worded differently or has minor typos.
3. Mark correct=0 if wrong, hallucinated, off-topic, or misses critical info.
4. For multiple choice: correct if the selected option matches gold (letter or text).
5. No explanation, only JSON.

For each query, the model output and ground-truth answer are fed to the judge along with the

original question. The judge returns a binary correctness label, which is aggregated into Acc@1 and Acc@5 scores.

B Baseline Details

B.1 OCR-Based Methods

Document images are parsed by RapidOCR to extract plain text. The extracted text is then indexed for retrieval:

- **OCR+BM25:** Sparse retrieval using BM25 with default parameters ($k_1 = 1.5, b = 0.75$).
- **OCR+BGE:** Dense retrieval using bge-m3 encoded via HuggingFace Transformers. Embeddings are indexed with FAISS for approximate nearest neighbor search.
- **OCR+Caption variants:** Image captions are generated by Qwen2.5-VL-3B-Instruct. Captions are concatenated with OCR text before indexing.

B.2 Global-Embedding Visual Retrieval (SigLIP)

We use the siglip-base-patch16-224 model to encode both query text and document images into a shared embedding space. Global average pooling compresses each image into a single vector. Cosine similarity is computed between query and document embeddings for ranking.

B.3 Patch-Level Visual Retrieval (ColPali)

We use the pre-trained colpali-v1.2 model to compute late-interaction MaxSim scores over visual patch tokens per page. **Crucially**, the ColPali retriever is **not** co-trained with the generator; retrieved images are fed into a frozen generator for answer generation, following the original decoupled paradigm.

B.4 Per-Source Retrieval Results on Vidore

Table 7 reports the complete per-source breakdown of retrieval performance (Recall@1 and Recall@5) across all five document categories in the Vidore benchmark suite, complementing the aggregated results in Table 1 of the main text.

C Algorithm Details

C.1 Differentiable Top- K Selection

Algorithm 1 provides the complete procedure for our sequential soft selection with Straight-Through

Estimator (STE). The key design is that the forward pass performs *hard* selection (one-hot), while the backward pass propagates gradients through *soft* probabilities, enabling end-to-end optimization of the retrieval pathway.

Algorithm 1: Differentiable Top- K Selection with STE

Input: Score vector $\mathbf{s} \in \mathbb{R}^N$, temperature

τ_{sel} , rounds K

Output: Selection weights

$\{\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(K)}\}$

```

1  $\mathbf{z} \leftarrow \mathbf{s} / \tau_{\text{sel}}$  // Temperature scaling
2  $\mathbf{M}^{(0)} \leftarrow \mathbf{0} \in \mathbb{R}^N$  // Initialize mask
3 for  $j \leftarrow 1$  to  $K$  do
4    $\mathbf{p}^{(j)} \leftarrow \text{softmax}(\mathbf{z} + \mathbf{M}^{(j-1)})$ 
   // Masked selection probabilities
5    $i^* \leftarrow \arg \max_i p_i^{(j)}$  // Forward: hard
   selection
6    $\mathbf{K}_{\text{hard}}^{(j)} \leftarrow \text{one\_hot}(i^*) \in \mathbb{R}^N$ 
7    $\mathbf{W}^{(j)} \leftarrow \mathbf{K}_{\text{hard}}^{(j)} + (\mathbf{p}^{(j)} - \text{sg}(\mathbf{p}^{(j)}))$ 
   // STE: hard forward, soft backward
8   Set  $M_{i^*}^{(j)} \leftarrow -\infty$ ;  $M_i^{(j)} \leftarrow M_i^{(j-1)}$ 
   for  $i \neq i^*$ 
9 return  $\{\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(K)}\}$ 

```

Retrieved Representation Aggregation. Given selection weights $\mathbf{W}^{(j)}$, the retrieved document representation at round j is:

$$H_{\text{ret}}^{(j)} = \sum_{i=1}^N W_i^{(j)} \cdot H_{\text{raw},i}, \quad (19)$$

where $H_{\text{raw},i} \in \mathbb{R}^{S \times D}$ is the retrieval-optimized representation of document d_i (see Eq. 2 in the main text). After K rounds, the set $\{H_{\text{ret}}^{(1)}, \dots, H_{\text{ret}}^{(K)}\}$ is passed to the gated generation stage.

C.2 Gated Residual Fusion

The complete computation flow of Gated Residual Fusion follows the implementation in GatedResidualFusion class:

1. Layer normalization of raw and retrieval-optimized features:

$$\tilde{H}_{\text{orig},i} = \text{LN}_{\text{orig}}(H_{\text{orig},i}), \quad (20)$$

$$\tilde{H}_{\text{raw},i} = \text{LN}_{\text{raw}}(H_{\text{raw},i}). \quad (21)$$

Method	Infographic-VQA		arXiv-QA		DocVQA		TAT-DQA		PDF		Overall	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
OCR+BM25	22.22	41.67	15.69	19.61	14.91	24.22	38.10	63.49	48.15	62.96	31.00	44.60
OCR+BGE	83.33	91.67	70.59	78.43	32.30	46.58	39.68	61.90	83.60	89.95	60.20	71.40
OCR+Cap+BM25	55.56	75.00	68.63	80.39	22.36	31.68	39.68	71.43	64.02	78.84	47.40	62.60
OCR+Cap+BGE	88.89	97.22	76.47	90.20	34.78	50.93	42.86	69.84	84.13	91.01	62.60	75.80
SigLIP	63.89	80.56	56.86	80.39	8.70	17.39	3.17	15.87	32.28	46.56	25.80	39.20
ColPali	91.67	100.00	88.24	96.08	63.98	80.12	68.25	93.65	89.95	95.24	78.80	90.60
Stage1	91.67	100.00	96.08	100.00	64.60	84.47	76.19	96.83	94.71	97.88	82.60	93.80
E2E	91.67	100.00	96.08	100.00	67.70	85.71	74.60	96.83	95.24	98.41	83.60	94.40

Table 7: Per-source retrieval performance on the Vidore benchmark. Recall@1 / Recall@5 (%) for each document category.

Parameter	Stage 1	Stage 2
Learning rate	1e−4	2e−5
Batch size	64	16
Epochs	1	1
LoRA rank r / alpha α	32 / 32	32 / 32
LoRA dropout	0.1	0.1
Temperature τ_{cl} / τ_{sel}	0.07	0.07 / 0.1
Loss weight λ_r / λ_g	—	0.1 / 1.0
Top- K retrieved pages		2
Max new tokens / Temperature		256 / 0.7
GPU		1×NVIDIA H100

Table 8: Key hyperparameters for training.

2. Residual increment and projection:

$$\Delta H_i = \tilde{H}_{raw,i} - \tilde{H}_{orig,i}, \quad (22)$$

$$\Delta H'_i = W_p \cdot \Delta H_i, \quad W_p \in \mathbb{R}^{D \times D}. \quad (23)$$

3. Gated fusion (initialized gate $\sigma(\gamma) \approx 0.2$):

$$H_{input,i} = H_{orig,i} + \sigma(\gamma) \cdot \Delta H'_i. \quad (24)$$

In practice, γ is initialized to -1.3863 (before sigmoid, yielding ≈ 0.2) with small random noise ($\mathcal{N}(0, 0.02)$) to break symmetry. The projection matrix W_p is initialized with $\mathcal{N}(0, 10^{-4})$.

D Reproducibility

D.1 Hyperparameters

Table 8 lists the complete hyperparameter configuration for both training stages.

D.2 Prompt Templates

Caption Generation Prompt. The following prompt is used to generate image captions for OCR+Caption baselines:

Caption Generation Prompt

Please describe all the data and information in this chart/table/image in detail, including any numbers, labels, and text visible.

Generation System Prompt. We use dataset-specific system prompts for generation:

RAG Generation System Prompt

Please Answer the question based on the retrieved images.

Rules:

1. Answer based ONLY on the content in the image.
2. For choice questions, output only the correct option text or letter.
3. Do not fabricate information not visible in the retrieved images.
4. The retrieved documents may contain multiple pages; evidence may exist only on a single page or several pages. Please distinguish valid evidence pages and exclude irrelevant noisy content.
5. (MMLongBench-Doc only) If explicit supporting evidence is missing, or the question requires assumption and reasoning beyond the given content, output exactly: Not answerable.