RegionRAG: Region-level Retrieval-Augmented Generation for Visual Document Understanding

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu*, Hongtao Xie

University of Science and Technology of China, Hefei, China
{lulupig, arieseirack, liuzhihang, syw95}@mail.ustc.edu.cn, {liucb92, htxie}@ustc.edu.cn

Abstract

Multi-modal Retrieval-Augmented Generation (RAG) has become a critical method for empowering LLMs by leveraging candidate visual documents. However, current methods consider the entire document as the basic retrieval unit, introducing substantial irrelevant visual content in two ways: 1) Relevant documents often contain large regions unrelated to the query, diluting the focus on salient information; 2) Retrieving multiple documents to increase recall further introduces redundant and irrelevant documents. These redundant contexts distract the model's attention and further degrade the performance. To address this challenge, we propose RegionRAG, a novel framework that shifts the retrieval paradigm from the document level to the region level. During training, we design a hybrid supervision strategy from both labeled data and unlabeled data to pinpoint relevant patches. During inference, we propose a dynamic pipeline that intelligently groups salient patches into complete semantic regions. By delegating the task of identifying relevant regions to the retriever, RegionRAG enables the generator to focus solely on concise, query-relevant visual content, improving both efficiency and accuracy. Experiments on six benchmarks demonstrate that RegionRAG achieves state-of-the-art performance. It improves retrieval accuracy by 10.02% in R@1 on average, and boosts question answering accuracy by 3.56% while using only 71.42% visual tokens compared with prior methods.

Code — https://github.com/Aeryn666/RegionRAG
Extended version — https://arxiv.org/pdf/2510.27261

1 Introduction

Retrieval-Augmented Generation (RAG) is a powerful paradigm that equips Large Language Models (LLMs) with external knowledge by retrieving relevant context from a dynamic database (Chen et al. 2022; Yasunaga et al. 2022; Liu et al. 2025b). As RAG achieves significant results on text databases, researchers have shifted their focus more to visual document databases with complex visual layouts that hinder knowledge retrieval and grounding (Yu et al. 2024; Faysse et al. 2024; Dong et al. 2025; Tanaka et al. 2025).

Early visual document RAG approaches relied on brittle pipelines that first extracted text via Optical Character Recognition (OCR) and layout analysis (Zhang et al. 2024). This process is not only complex but also discards crucial visual information (e.g., layout, style) for holistic comprehension. To address these limitations, the development of Vision-Language Models (VLMs) has spurred a new wave of frameworks that operate directly on document images. Representative works such as VisRAG (Yu et al. 2024), ColPali (Faysse et al. 2024), and VDocRAG (Tanaka et al. 2025) leverage VLMs to retrieve documents based on holistic visual and textual features, thereby bypassing the error-prone text extraction stage, preserving richer semantics.

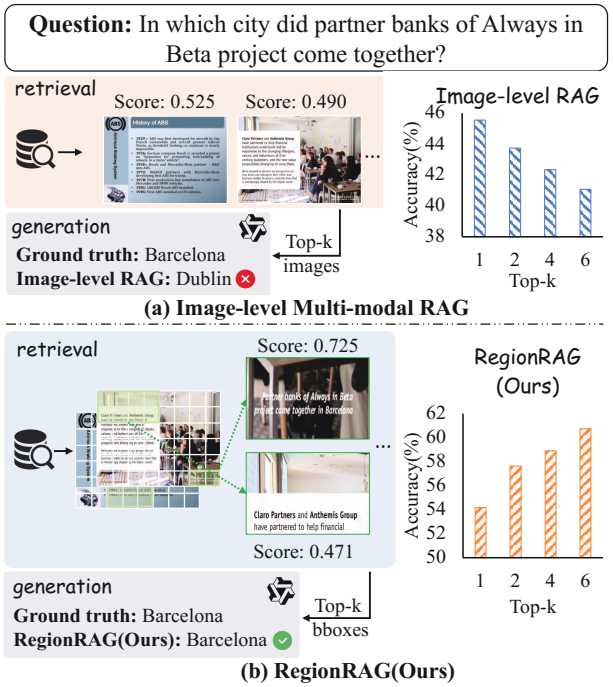

Figure 1: Comparison of (a) Image-level RAG with (b) our proposed RegionRAG. While traditional methods retrieve entire yet coarse-grained document images, our RegionRAG identifies and forwards only the most salient regions to the generator. This focused, region-level approach significantly improves final generation accuracy.

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Despite recent progress, existing VLM-based RAG systems typically operate at the document level, treating the entire image as a retrieval unit. This coarse granularity introduces two major types of noise that degrade generation performance. The first type is from the irrelevant information in a document. Even relevant documents often contain large regions unrelated to the query, diluting the model's focus on truly salient content. The second type is from the multiple documents specially introduced in RAG. A common strategy is to feed the top-k retrieved documents into the generator (Yao et al. 2025) to compensate for retrieval errors. However, this introduces a performance paradox: while intended to increase ground-truth recall, our experiment shown in Figure 1(a) reveals that increasing k actually leads to a decrease in performance. This may be because including redundant and irrelevant documents significantly increases visual tokens and distracts the model's attention, thereby outweighing the limited performance gains provided by the small amount of valid information.

To tackle this problem, an intuition is to shift the retrieval granularity from the entire document to only the most relevant semantic regions. Therefore, we propose a novel framework named RegionRAG that enhances retrieval at both the training and inference stages. During training, we improve the alignment between queries and document regions through a hybrid supervision strategy. Specifically, RegionRAG jointly leverages manually annotated bounding boxes from labeled data and weakly-supervised similarity signals from unlabeled data. A unified loss integrates both supervision types to learn precise query-patch alignment while minimizing annotation cost. During inference, we introduce a dynamic region proposal algorithm based on patch-specific similarity maps. Since relevant information often spans multiple neighboring patches (e.g., parts of a table or paragraph), we apply a neighbor-based grouping strategy that merges patches into coherent regions. This enhances matching precision by isolating concise, context-complete visual segments relevant to the query. As shown in Figure 1(b), our region-level RAG can effectively improve performance as k increases. Extensive experiments demonstrate that RegionRAG achieves a 3.56% average accuracy gain across five document visual question answering (VQA) benchmarks, and cuts 28.58% costs of visual tokens compared to prior methods.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose a region-level multi-modal RAG framework (RegionRAG) for visual documents.
- We introduce a dual-objective training strategy that achieves fine-grained patch alignment using different types of data, and introduce a neighbor-based grouping inference strategy to get visual regions.
- Experiments show our RegionRAG achieves more accurate retrieval and efficient generation compared with previous methods.

2 Related Works

Image-Level Retrieval-Augmented Generation (RAG). RAG enhances Large Language Models (LLMs) by incorporating external knowledge, a paradigm that was initially successful in natural language processing (NLP) tasks (Guu et al. 2020; Borgeaud et al. 2022; Ram et al. 2023). It typically involves retrieving relevant information from an external knowledge base to guide answer generation (Shi et al. 2023; Yu et al. 2023; Liu et al. 2024). Beyond NLP, RAG has been extended to visually rich documents. Early works often relied on brittle OCR pipelines (Yu et al. 2023), which discarded visual cues and introduced fragile pre-processing steps. However, this approach discards visual cues and introduces brittle pre-processing steps. Recently, the rapid development of Vision-Language Models (VLMs) (Liu et al. 2023, 2025a), such as VisRAG (Yu et al. 2024), ColPali (Faysse et al. 2024), and VDocRAG (Tanaka et al. 2025) has enabled the direct retrieval of entire document images. While pioneering, these methods operate at a coarse, document-level granularity. This forces the generator to process entire images, introducing substantial noise from query-irrelevant content, which dilutes contextual relevance, degrades generation quality, and reduces computational efficiency.

General Visual Document Understanding. Another line of research focuses on end-to-end understanding of individual visual documents. Traditional OCR-based pipelines (Zhang et al. 2024) are fragile and discard layout cues crucial for reasoning. With the emergence of multimodal LLMs, models such as DocFormer (Appalaraju et al. 2021), Qwen-VL (Bai et al. 2023), and InternVL (Chen et al. 2024) demonstrate unified reasoning over document images. However, these models are not designed for large-scale retrieval. Retrieval-augmented methods like VisRAG (Yu et al. 2024) and TabPedia (Zhao et al. 2024) leverage visual grounding but still retrieve entire documents, introducing redundant context. In contrast, RegionRAG retrieves fine-grained, query-relevant regions for precise and efficient context construction.

Region-Based VQA and Grounding. Inspired by the need for finer granularity, another category of models explores region-level reasoning. Studies such as RegionGPT (Guo et al. 2024), region-based VQA methods (Wu et al. 2022), VLM-R3 (Jiang et al. 2025), and DeepEyes (Zheng et al. 2025) adopt an end-to-end generate-and-ground paradigm. However, their task also differs crucially from ours: they primarily perform localization or grounding within a single, pre-selected document to answer a query. They are not retrieval systems designed to search a corpus. While one could envision an agentic, multi-step pipeline for corpus-level retrieval using these models, such approaches are prohibitively slow for efficient retrieval. In contrast, RegionRAG is the first to formalize an explicit, decoupled region-retrieval stage for the RAG paradigm, bridging the gap between efficient retrieval and fine-grained understanding.

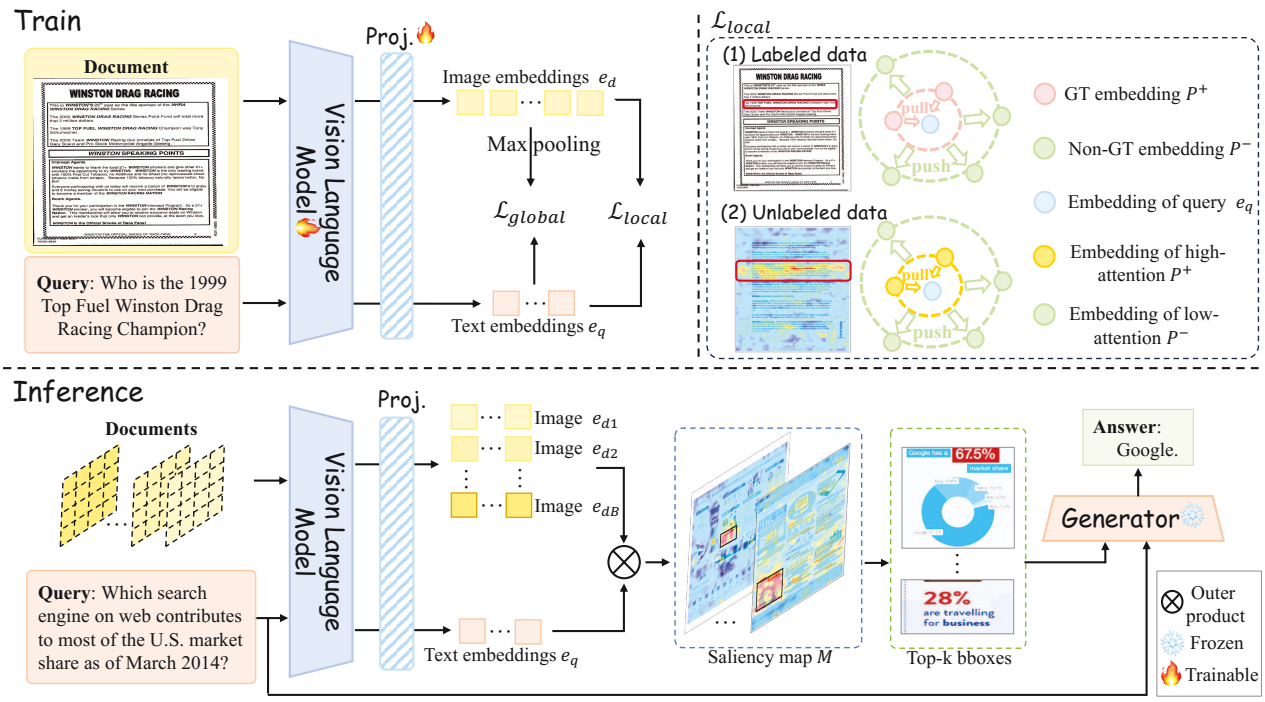

Figure 2: The framework of RegionRAG. During training, the model jointly learns global (document-query) and local (region-query) alignments. At inference time, it first identifies and retrieves the most relevant visual regions and subsequently generates an answer based on this retrieved context.

3 Methods

3.1 Preliminary

Our RegionRAG framework redefines the standard RAG pipeline for visually-rich documents. We begin with a corpus $\mathcal{C}$ containing M documents, $\mathcal{C} = \{D_1, D_2, \ldots, D_M\}$. Each document D_m is segmented into N non-overlapping patches $\{p_1, p_2, \ldots, p_N\}$. These patches serve as our fundamental retrieval units, contrasting with traditional methods that operate on whole documents.

Our pipeline consists of two main components. Firstly, a RegionRetriever $(\mathcal{R})$, is responsible for identifying the most relevant set of patches $P_{\mathcal{R}}$ from a document D_m given a text query q. It uses a shared VLM to embed both the query and the visual patches into a common latent space for fine-grained similarity matching; its function is represented as $P_{\mathcal{R}} = \mathcal{R}(q, D_m)$.

We train $(\mathcal{R})$ with two levels of alignment to develop its region retrieval ability, a global document-query level and a fine-grained region-query level. Secondly, we employ a Generator $(\mathcal{G})$ to synthesize the final textual answer a. As an MLLM, it is conditioned on the original query q and the set of retrieved image patches $P_{\mathcal{R}}$; this process is denoted as $a = \mathcal{G}(q, P_{\mathcal{R}})$.

3.2 Global Document-Query Alignment

While our ultimate goal is to retrieve specific regions, the model must first identify the correct parent document. Our framework first performs Global Document-Query Alignment to obtain a holistic document representation and ensure basic retrieval ability among documents.

We follow (Faysse et al. 2024; Yu et al. 2024) to employ global contrastive learning. To create a holistic document representation from its constituent patches, we aggregate their features into a single global vector. We begin by segmenting a document image D into N patches. A VLM then processes these patches to produce a set of N embeddings, $E_D = \{e_1, e_2, \ldots, e_N\}$, where each $e_k \in \mathbb{R}^d$ and d is the embedding dimension. We apply an element-wise max-pooling operation across the N patch embeddings E_D to obtain the single global representation v_D. During training, we sample a batch of B document-query pairs $\{D_i, q_i\}_{i=1}^{B}$. For each pair, we compute the document embedding v_{D_i} and the query embedding e_{q_i}. The similarity between the i-th query and the j-th document is measured by cosine similarity:

$$s(e_{q_i}, v_{D_j}) = \frac{e_{q_i} \cdot v_{D_j}}{\|e_{q_i}\| \, \|v_{D_j}\|}. \quad (1)$$

The objective function for global contrastive learning is based on InfoNCE loss (He et al. 2020). For the i-th pair in the batch, we treat (e_{q_i}, v_{D_i}) as the positive pair and (e_{q_i}, v_{D_j}) for all $j \neq i$ as the negative pairs. The loss for the entire batch is:

$$\mathcal{L}_{global} = -\frac{1}{b} \sum_{i=1}^{b} \log \left(\frac{\exp(s(e_{q_i}, v_{D_i})/\tau)}{\sum_{j=1}^{b} \exp(s(e_{q_i}, v_{D_j})/\tau)} \right), \quad (2)$$

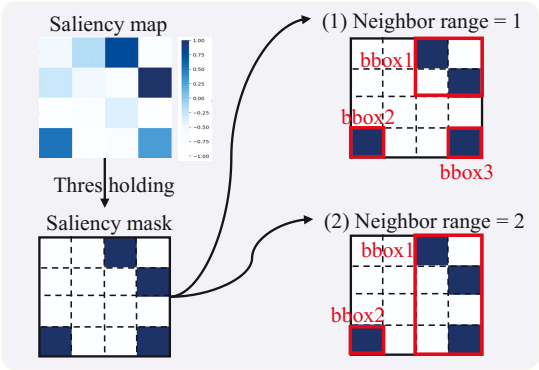

Figure 3: Our neighbor-based grouping strategy. Patches with high saliency are binarized and then grouped into bounding boxes.

where τ is a temperature parameter. This global alignment stage leverages inter-document level supervision, enabling the model to retrieve relevant documents from the corpus.

3.3 Fine-grained Region-Query Alignment

The coarse-grained global alignment only provides inter-document supervision and cannot achieve precise region-level activation within a document, thus failing to support fine-grained region retrieval. Therefore, we introduce Fine-grained Region-Query Alignment to enable intra-document supervision. This is primarily implemented through the local objective function $\mathcal{L}_{local}$, which guides the model to pinpoint the specific regions most relevant to the query. The core of our region-level approach is a hybrid supervision strategy that leverages two distinct signal sources: gold-standard annotations from labeled data (with bounding boxes) and weakly-supervised signals from unlabeled data (without bounding boxes). Both supervision types share a unified loss formulation.

We consider a training batch of B document-query pairs, denoted as $\{D_i, q_i\}_{i=1}^{B}$. For each pair, the VLM provides a query embedding $e_{q_i} \in \mathbb{R}^d$ and a set of N patch embeddings for the document D_i, denoted as $E_{D_i} = \{e_{i,1}, e_{i,2}, ..., e_{i,N}\}$. Each patch embedding $e_{i,k} \in \mathbb{R}^d$ represents a specific document region. The core of our fine-grained stage is to compute the similarity between the query e_{q_i} and each patch embedding $e_{i,k}$ to identify the most relevant regions.

Supervision from Labeled Data. For labeled data, each document-query pair (D_i, q_i) is accompanied by ground-truth bounding boxes that delineate the regions essential for answering the query. We use this precise supervision to construct positive and negative sets of patch embeddings. The positive set, P_i^+, consists of all patch embeddings $\{e_{i,k}\}$ corresponding to image patches centered within any ground-truth box. Conversely, the negative set, P_i^-, consists of all patch embeddings for patches lying entirely outside these boxes. The local loss $\mathcal{L}_{local}^{L}$ for a labeled sample encourages the query embedding e_{q_i} to be closer to the embeddings in P_i^+ than to those in P_i^-.

Supervision from Unlabeled Data. A key advantage of our method is its ability to learn from unlabeled data, which contains only weakly-supervised document-query pairs (D_i, q_i) without region annotations. To extract supervisory signals, we generate pseudo-ground-truth labels by leveraging the model's similarity estimates between the query and document patch embeddings. This is feasible because a well-pretrained VLM, guided by the global loss and supervision from labeled data, naturally exhibits some patch-level attention capability. We compute a saliency map that captures the interaction between the query q_i and the document patch embeddings E_{D_i}. This map is then binarized using a predefined threshold θ. To ensure stable selection of positive samples, we adopt a relatively low value for θ. Patch embeddings with similarity scores above θ are included in the positive set P_i^+, while those below are assigned to the negative set P_i^-.

Given the positive set P_i^+ and negative set P_i^- (derived from either ground-truth or pseudo-labels), the unified contrastive loss $\mathcal{L}_{\text{local}}$ is formulated as:

$$\mathcal{L}_{\text{local}} = -\log\left(\frac{\sum_{p\in P_i^+}\exp(s(e_{q_i},p)/\tau)}{\sum_{p\in\{P_i^+,P_i^-\}}\exp(s(e_{q_i},p)/\tau)}\right), \quad (3)$$

where τ is a temperature parameter.

3.4 Overall objective

The overall RegionRAG learning objective is a weighted sum of the global and fine-grained alignment loss:

$$\mathcal{L}_{\text{RegionRAG}} = \alpha\mathcal{L}_{global} + \beta\mathcal{L}_{local}, \quad (4)$$

where the weights α and β are hyper-parameters.

3.5 Inference

Building upon the model's capacity for local region attention developed during training, our inference process extracts specific areas to perform region-level retrieval and subsequent answer generation. The pipeline unfolds in three steps: first, we obtain a proposal mask from a retrieval saliency map. Second, we identify connected components of salient patches based on a defined neighborhood size. Third, we extract the minimum bounding rectangle for each element. Finally, we feed the retrieved regions into a generator $\mathcal{G}$ to produce the final answer.

Region Proposal via Saliency Mapping. The inference process begins by identifying candidate regions within a document image D that are potentially relevant to the input query q. We first compute a saliency map $S = \{s(e_q, p_k)|k = 1, 2, \ldots, N\}$, which is a 2D grid of cosine similarity corresponding to the layout of the N document patches. To filter out patches with low relevance, we binarize this saliency map into a mask M using a predefined threshold $\eta \in [0, 1]$. Patches with scores above this threshold are considered salient candidates.

$$M_k = \begin{cases} 1, & \text{if } S_k \geq \eta \\ 0, & \text{if } S_k < \eta \end{cases}. \quad (5)$$

This step isolates a sparse set of potentially relevant regions, as shown in Figure 3, preparing them for the subsequent grouping and merging process.

Algorithm 1: Bounding Box Proposal from Saliency Map

Input: Saliency map S, threshold η, neighbor range r, image size $(H_{\text{img}}, W_{\text{img}})$, grid size (H_p, W_p)
Output: A set of bounding boxes $\mathcal{B}$

$M \leftarrow S \geq \eta$ ▷ Binarize saliency map
$\mathcal{R} \leftarrow \text{FindComponentsBFS}(M, r)$ ▷ Find connected components
$\mathcal{B} \leftarrow \emptyset$
for each component R in $\mathcal{R}$ **do**
    $B \leftarrow \text{CalculateMinBbox}(R, (H_{\text{img}}, W_{\text{img}}), (H_p, W_p))$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{B\}$
end for
return $\mathcal{B}$

Region Merging and Bounding Box Generation. To prevent excessive fragmentation of semantically related areas, we group salient patches from the mask M into connected components. Two patches are considered connected if the distance between their spatial coordinates is less than or equal to a neighborhood radius r. We use the Chebyshev distance as our metric:

$$D((x_1, y_1), (x_2, y_2)) = \max(|x_1 - x_2|, |y_1 - y_2|). \quad (6)$$

The radius r controls the merging granularity, allowing for either tight or broad region proposals as illustrated in Figure 3. While exhaustively traversing all patches is inefficient, we employ a Breadth-First Search (BFS) algorithm for efficient traversal. Specifically, we use a queue to manage the search and mark patches as visited to avoid redundant processing. For each connected component found, we compute its minimum bounding rectangle. Therefore, these bounding rectangles can be considered our final retrieved regions, and can be further sent into the generator $\mathcal{G}$ to complete the question answering. For each image, the entire region retrieval process is summarized in the Algorithm 1.

4 Experiments

4.1 Experimental Settings

Datasets. We train our model using a combination of the unlabeled dataset from VisRAG (Yu et al. 2024) in-domain data and the document-focused subset of Visual-CoT (Shao et al. 2024) for labeled, bounding-box level supervision. For evaluation, we test our method on a diverse suite of benchmarks including DocVQA (Tito, Karatzas, and Valveny 2023), PlotQA (Methani et al. 2020), ArxivQA (Li et al. 2024), InfoVQA (Mathew et al. 2022), SlideVQA (Tanaka et al. 2023), and ViDoRe (Faysse et al. 2024).

Evaluation Metrics. We report the retrieval and generation performance on the evaluation sets of the datasets sourced from the VQA datasets. For retrieval, we use Recall@1, Recall@10, and nDCG@5. For generation, we follow VisRAG (Yu et al. 2024) to report the answer accuracy, employing a relaxed extract match metric that allows a 5% error margin for numeric responses.

Implementation Details. We initialized RegionRetriever with Qwen2.5-VL-3B (Bai et al. 2025) for training, while the Generator is an off-the-shelf model without training. The model is trained for five epochs on a mixture of 98k labeled and 122k unlabeled data. The batch size per GPU is set to 64. The temperature parameter τ is set to 0.02 for the global loss and 0.25 for the local loss.

Models	#Para	Arxiv	Doc	Info	Plot	Slide	Avg
BM25	n.a.	54.3	86.8	82.6	76.0	91.6	78.3
SigLIP	883M	45.0	68.0	84.7	58.3	89.0	69.0
BGE(large)	335M	48.7	68.2	88.2	73.1	90.1	73.7
NV-Embed-V2	7.85B	70.1	89.9	95.1	80.9	97.8	86.8
ColPali*	2.92B	79.4	90.2	88.9	31.4	92.8	76.5
VisRAG*	3B	87.6	91.2	97.0	89.3	97.1	92.4
ColQwen2.5*	3B	84.5	97.4	98.6	47.5	97.0	85.0
RegionRAG (ours)	3B	**92.5**	**99.4**	**99.5**	**92.4**	**98.4**	**96.4**

Table 1: Retrieval performance comparison between our RegionRAG and other SOTA methods on multiple types of DocumentVQA benchmarks in Recall@10. $*$ denotes that we reproduce the evaluation using their official checkpoints.

Models	#Data	Arxiv	Doc	Info	Plot	Slide	ViDoRe
		Recall@1					
VisRAG*	362k	69.1	66.2	78.8	49.1	76.4	-
ColQwen2.5*	128k	62.3	80.2	82.3	9.3	76.1	**83.3**
RegionRAG	220k	**78.4**	**86.8**	**88.9**	**55.3**	**80.6**	82.9
		nDCG@5					
VisRAG*	362k	78.8	77.7	88.4	65.5	89.9	-
ColQwen2.5*	128k	71.8	87.8	90.3	19.0	86.7	**89.4**
VDocRAG	541k	-	-	72.9	-	77.3	-
RegionRAG	220k	**84.5**	**93.1**	**94.8**	**70.3**	**90.3**	88.4

Table 2: Retrieval performance comparison in Recall@1 and nDCG@5. $*$ denotes that we reproduce the evaluation based on their official checkpoints.

4.2 Performance Comparison

We compare our RegionRAG with the following state-of-the-art methods, including OCR-based pipelines BM25 (Robertson, Zaragoza et al. 2009), BGE-large-en-v1.5 (Xiao et al. 2024), and NV-Embed-V2 (Lee et al. 2024), and VLM-based methods SigLIP (Zhai et al. 2023), VisRAG (Yu et al. 2024), VDocRAG (Tanaka et al. 2025), ColPali and ColQwen2.5 (Faysse et al. 2024).

Retrieval Performance. As shown in Table 1, our RegionRAG outperforms existing methods on the Recall@10 metric across all benchmarks. Compared to VisRAG, which has a comparable number of model parameters (3B) and is trained on 362k samples, our model achieves superior performance with less data (220k). On average, RegionRAG surpasses VisRAG by 4.0 percentage points in Recall@10. When compared with ColQwen2.5, which also adopts Qwen2.5-VL-3B as its backbone, the advantage of RegionRAG becomes even more pronounced. Its average Recall@10 score is 11.4 percentage points higher (96.4 vs. 85.0), with a huge margin on the Plot benchmark (92.4 vs.

Methods	Region	DocVQA	InfoVQA	Average
ColQwen2.5	Image	0.062	0.109	0.086
	Bbox	0.112	0.138	0.125
RegionRAG (Ours)	Image	0.119	0.226	0.173
	Bbox	**0.271**	**0.346**	**0.309**

Table 3: Comparison of average similarity scores between text queries and different visual scopes (full image vs. ground-truth bounding box) on DocVQA and InfoVQA.

47.5). These results demonstrate that our proposed method of supervising the most query-relevant image regions is even superior on the whole document retrieval than employing image-level supervision. This success can be attributed to our tailored training scheme and loss design. We further compare RegionRAG with recent VLM-based methods on Recall@1 and nDCG@5 metrics in Table 2. Our model attains leading performance on most benchmarks, highlighting its consistent superiority. While its scores on ViDoRe are marginally behind those of the top performer, this evaluation constitutes a partial zero-shot setting for our model due to the limited subset domain coverage in our training data.

Region-Text Alignment Capability. We evaluate the model's ability to localize query-relevant regions by comparing the average retrieval similarity between the full image and the ground-truth bounding box (Bbox) from the Visual-CoT test split. As shown in Table 3, both ColQwen2.5 and our RegionRAG achieve higher similarity on Bbox regions, indicating that relevant information is locally concentrated. Notably, RegionRAG yields a larger similarity gain (0.136 vs. 0.039 for ColQwen2.5), highlighting the effectiveness of our region-level contrastive loss in capturing fine-grained, query-relevant visual cues.

Generation Performance. We evaluate our method under both fixed and dynamic resolution settings. In the fixed setup, we process the retrieved region at either 256^2 or 512^2 resolution. As shown in Table 4, accuracy improves notably at both resolutions, and using retrieved bboxes consistently outperforms full-image inputs under the same resolution. In the dynamic setting, which better reflects real-world inference, the model uses the original resolution with up to 512^2. On average, the bbox input consumes only 71.4% of the image tokens while maintaining or even surpassing full-image performance. For instance, it achieves 16.29% higher accuracy on InfoVQA. Notably, the full-image method is upper-bounded by the "oracle" setting, which inputs the ground-truth image at a fixed resolution. Interestingly, our bounding box approach can even surpass this baseline. For example, on InfoVQA at 512^2 resolution, our top-4 bbox method attains 63.79% accuracy, exceeding the oracle's 51.39%. This is because downscaling affects smaller cropped regions less severely than entire images—preserving finer details that enable the model to better perceive the relevant content. Overall, these results highlight the advantages of our region retrieval in both efficiency and fidelity.

Pixel	Methods	Arxiv	Doc	Info	Plot	Slide
256^2	top1 image	**62.25**	19.97	24.23	16.69	38.49
	top1 bbox	62.13	**31.13**	**39.13**	**17.15**	**39.92**
	top4 image	61.40	18.95	22.42	17.73	38.31
	top4 bbox	**62.62**	**42.3**	**44.15**	**19.00**	**44.42**
	oracle	64.71	19.97	24.37	20.39	39.39
512^2	top1 image	64.46	69.20	50.97	**24.10**	59.17
	top1 bbox	**64.83**	**70.39**	**59.19**	**24.10**	**62.94**
	top4 image	63.73	70.22	47.91	23.52	61.69
	top4 bbox	**65.07**	**74.28**	**63.79**	**24.56**	**66.01**
	oracle	67.77	75.63	51.39	37.78	63.31
dynamic (top4)	image acc	**65.2**	70.22	46.94	**24.10**	62.77
	#tokens	1258	1268	1248	1266	1259
	bbox acc	64.83	**71.24**	**63.23**	23.64	**64.03**
	#tokens	1035	903	661	966	936
	oracle	68.01	75.63	51.39	37.78	63.31

Table 4: Performance comparison under fixed and dynamic resolutions across five VQA benchmarks. "topk image/bbox" denotes processing the retrieved full image or bounding box, "oracle" uses ground-truth images, "dynamic" means using the original resolution with up to 512^2.

Methods	ArxivQA		DocVQA		SlideVQA	
	R@1	N@5	R@1	N@5	R@1	N@5
RegionRAG (Ours)	**78.43**	**84.52**	**86.80**	**93.10**	**80.58**	**90.29**
w/o $\mathcal{L}_{local}$	77.08	83.41	84.94	91.70	78.24	88.90
w/o $\mathcal{L}_{global}$	77.57	81.68	85.45	81.14	78.24	86.85
w/o Labeled data	59.31	84.24	82.40	90.88	77.34	89.19
w/o Unlabeled data	73.90	67.75	71.07	90.83	75.00	88.18

Table 5: Main ablation study (%) of retrieval results for global contrastive learning ($\mathcal{L}_{global}$), regional contrastive learning ($\mathcal{L}_{local}$) with labeled and unlabeled data.

4.3 Ablation

Main Ablation. The core components of our method include the global contrastive loss ($\mathcal{L}_{global}$), the regional contrastive loss ($\mathcal{L}_{local}$), and a semi-supervised strategy that utilizes both labeled and unlabeled data. As shown in the Table 5, removing either $\mathcal{L}_{global}$ or $\mathcal{L}_{local}$ consistently leads to performance degradation across all datasets, indicating that both global document-level and fine-grained region-level contrastive learning are crucial for effective retrieval. Moreover, the results underscore the effectiveness of our semi-supervised data strategy. While discarding labeled data causes a substantial performance collapse, excluding unlabeled data also yields a notable decline (e.g., a 4.53% drop in R@1 on ArxivQA). These findings confirm that our model not only relies on the supervised signal but also effectively leverages large-scale unlabeled data to enhance its generalization and robustness.

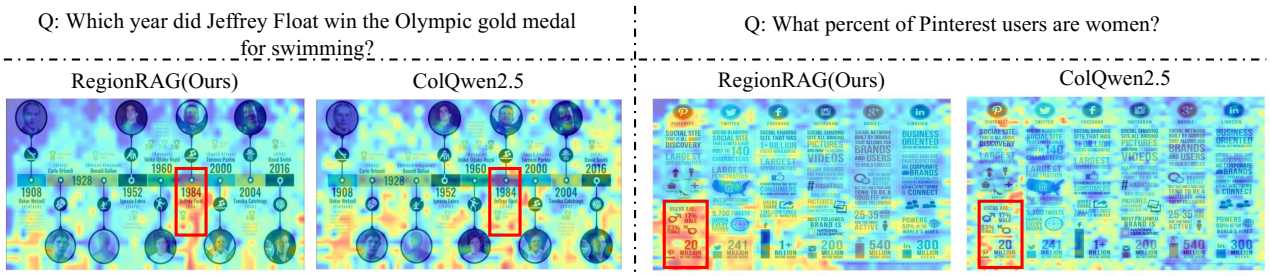

Figure 4: Visualization of similarity maps. The relevant regions are highlighted with red boxes for better reading. Our RegionRAG obtains better localization ability, as it exhibits higher and more concentrated similarity in relevant regions.

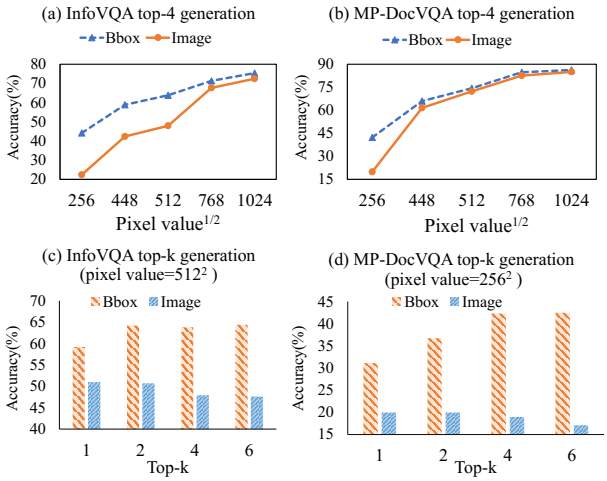

Figure 5: Ablation study on generation performance. We compare using Bbox versus Image as input to the generator. (a) and (b) show accuracy versus input resolution for a fixed $k = 4$. (c) and (d) show accuracy as a function of top-k candidates at a fixed resolution.

Ablation Study on Generation Inputs. We conduct an ablation study to analyze the impact of input granularity on the generator. Specifically, we compare providing cropped bounding box (Bbox) regions against full image pages, focusing on two factors: input resolution (i.e., visual token count) and the number of top-k retrieved candidates. As shown in Figure 5(a) and (b), when inputs are constrained to a fixed resolution, the Bbox method consistently outperforms the Image method. The advantage is most pronounced at lower resolutions like 256^2 and 448^2, where Bbox accuracy exceeds Image by up to 21.73 points. This improvement arises because resizing a small Bbox crop enlarges its informative content compared to a down-sampled full image. Consequently, our approach achieves higher performance with fewer tokens (e.g., Bbox at 256^2 rivals Image at 448^2 on InfoVQA). At very high resolution (1024^2), the gap narrows as Qwen2.5-VL already localizes answers in high-fidelity images. Overall, our method strikes a better balance between performance and efficiency. Figure 5(c) and (d) further analyze the effect of the top-k setting at a fixed resolution. For all k, Bbox accuracy remains consistently higher than Image accuracy. Critically, as k increases, Bbox performance improves while Image performance degrades. This contrast likely stems from attention dilution in the full-image setting, which introduces redundant and query-irrelevant information that distracts the LLM. In contrast, for the Bbox method, a larger k presents an opportunity to provide more regions that are relevant to the query, leading to improved accuracy. These results confirm that our region-based approach effectively leverages multiple inputs, whereas the image-level approach is penalized by them.

4.4 Qualitative Analysis

To qualitatively investigate how RegionRAG achieves its strong performance, we visualize the similarity heatmaps in Figure 4. The figure clearly shows where the model gives the relevant regions to retrieve from a given query, comparing our method with ColQwen2.5, which processes the entire image. The similarity map of RegionRAG is significantly more concentrated on the query-relevant visual parts. In contrast, ColQwen2.5 focuses diffusely across the image, indicating a difficulty in distinguishing important information from background noise. For example, for the query about Jeffrey Float, our model pinpoints the exact '1984' entry on the timeline. Similarly, for the other example, RegionRAG precisely pinpoints the relevant charts and data points. This shows that our region-level retrieval mechanism effectively guides the model to the most salient information, which is key to its superior performance.

5 Conclusion

In this work, we presented RegionRAG, an innovative framework that enhances visual document RAG by retrieving fine-grained semantic regions instead of entire documents. By leveraging a dual-objective training strategy and a neighbor-based region grouping algorithm, RegionRAG effectively filters irrelevant context and provides the generator with precise visual evidence. Our method achieves state-of-the-art performance across six benchmarks, significantly improving both retrieval (+10.02% in R@1) and question answering accuracy (+3.56%) while cutting visual token costs by 28.58%. In the future, we aim to build more efficient and scalable RAG systems, extending to more general scenarios.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (62425114, 62121002, U23B2028, 62272436). We thank the support of the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and USTC super-computing center for providing computational resources for this project.

References

Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 993–1003.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Chen, W.; Hu, H.; Saharia, C.; and Cohen, W. W. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Dong, Y.; Ueda, N.; Boros, K.; Ito, D.; Sera, T.; and Oyamada, M. 2025. SCAN: Semantic Document Layout Analysis for Textual and Visual Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.14381*.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Guo, Q.; De Mello, S.; Yin, H.; Byeon, W.; Cheung, K. C.; Yu, Y.; Luo, P.; and Liu, S. 2024. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13796–13806.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Jiang, C.; Heng, Y.; Ye, W.; Yang, H.; Xu, H.; Yan, M.; Zhang, J.; Huang, F.; and Zhang, S. 2025. VLM-R ^{3}: Region Recognition, Reasoning, and Refinement for Enhanced Multimodal Chain-of-Thought. *arXiv preprint arXiv:2505.16192*.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Li, L.; Wang, Y.; Xu, R.; Wang, P.; Feng, X.; Kong, L.; and Liu, Q. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 3855–3863.

Liu, Z.; Xie, C.-W.; Li, P.; Zhao, L.; Tang, L.; Zheng, Y.; Liu, C.; and Xie, H. 2025a. Hybrid-level instruction injection for video token compression in multi-modal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8568–8578.

Liu, Z.; Xie, C.-W.; Wen, B.; Yu, F.; Li, P.; Zhang, B.; Yang, N.; Gao, Z.; Zheng, Y.; Xie, H.; et al. 2025b. CAPability: A Comprehensive Visual Caption Benchmark for Evaluating Both Correctness and Thoroughness. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; and Jawahar, C. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1697–1706.

Methani, N.; Ganguly, P.; Khapra, M. M.; and Kumar, P. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 1527–1536.

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37: 8612–8642.

Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Tanaka, R.; Iki, T.; Hasegawa, T.; Nishida, K.; Saito, K.; and Suzuki, J. 2025. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24827–24837.

Tanaka, R.; Nishida, K.; Nishida, K.; Hasegawa, T.; Saito, I.; and Saito, K. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13636–13645.

Tito, R.; Karatzas, D.; and Valveny, E. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144: 109834.

Wu, X.; Zheng, D.; Wang, R.; Sun, J.; Hu, M.; Feng, F.; Wang, X.; Jiang, H.; and Yang, F. 2022. A region-based document VQA. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4909–4920.

Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; and Nie, J.-Y. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 641–649.

Yao, J.; Liu, S.; Wang, Y.; Mei, L.; Bi, B.; Ge, Y.; Li, Z.; and Cheng, X. 2025. Who is in the Spotlight: The Hidden Bias Undermining Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2506.11063*.

Yasunaga, M.; Aghajanyan, A.; Shi, W.; James, R.; Leskovec, J.; Liang, P.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Yu, Z.; Xiong, C.; Yu, S.; and Liu, Z. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, G.; Qu, S.; Liu, J.; Zhang, C.; Lin, C.; Yu, C. L.; Pan, D.; Cheng, E.; Liu, J.; Lin, Q.; et al. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv:2405.19327*.

Zhao, W.; Feng, H.; Liu, Q.; Tang, J.; Wu, B.; Liao, L.; Wei, S.; Ye, Y.; Liu, H.; Zhou, W.; et al. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37: 7185–7212.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.