Spatially-Grounded Document Retrieval via Patch-to-Region Relevance Propagation

Agathoklis Georgiou

Independent Researcher

athrael.soju@gmail.com

Abstract

Vision-language models (VLMs) like ColPali achieve state-of-the-art document retrieval by embedding pages as images and computing fine-grained similarity between query tokens and visual patches. However, they return entire pages rather than specific regions, limiting utility for retrieval-augmented generation (RAG) where precise context is paramount. Conversely, OCR-based systems extract structured text with bounding box coordinates but lack semantic grounding for relevance assessment. We propose a hybrid architecture that unifies these paradigms: using ColPali's patch-level similarity scores as spatial relevance filters over OCR-extracted regions. We formalize the coordinate mapping between vision transformer patch grids and OCR bounding boxes, introduce intersection metrics for relevance propagation, and establish theoretical bounds on retrieval precision. We evaluate on BBox-DocVQA with ground-truth bounding boxes. Using ColQwen3-4B with percentile-50 thresholding, our approach achieves **85.8% hit rate at IoU@0.25**, **62.3% at IoU@0.5**, and **41.9% at IoU@0.7**, with mean IoU of **0.597**. Our approach reduces context tokens by **30%** compared to returning all OCR regions and by **54%** compared to full-page image tokens. Our approach operates at inference time without additional training. We release Snappy, an open-source implementation at https://github.com/athrael-soju/Snappy.

1 Introduction

Retrieval-augmented generation (RAG) has emerged as the dominant paradigm for grounding large language models in external knowledge, enabling factual responses without costly retraining. The effectiveness of RAG systems hinges on a fundamental requirement: retrieving *precisely relevant* context while minimizing noise. For text corpora, this challenge is well-studied. Dense retrievers identify semantically similar passages, and chunking strategies control context granularity. However, document collections present a fundamentally harder problem.

Documents are not sequences of tokens but *spatially-organized visual artifacts*. A single page may contain heterogeneous elements, including tables, figures, equations, headers, and footnotes, each carrying distinct semantic content at different spatial locations. When a user queries "What was the Q3 revenue?", the answer likely resides in a specific table cell, not spread across the entire page. Yet current retrieval systems operate at the wrong granularity.

Vision-language models (VLMs) such as ColPali (Faysse et al., 2025) have achieved state-of-the-art performance on document retrieval benchmarks by embedding document pages directly as images. ColPali produces 1,024 patch embeddings (a 32×32 grid) per page, each projected to 128 dimensions. The model computes relevance through late interaction, specifically a MaxSim operation that sums the maximum similarity between each query token and all document patches. This approach elegantly sidesteps OCR errors and preserves layout semantics. However, ColPali

and its variants return *entire pages* as retrieval units. For RAG applications, this is problematic: feeding a full page into a language model's context window introduces irrelevant content, increases latency, inflates costs, and, critically, dilutes the signal that the model must attend to. The retrieval system knows *which page* contains the answer but not *where on the page*.

Conversely, OCR-based pipelines extract text with precise bounding box coordinates, enabling structured representations of document content. Tables become rows and columns; figures receive captions; headers define hierarchy. This structural fidelity is invaluable for downstream processing. Yet OCR systems lack *semantic grounding*. They cannot assess which extracted regions are relevant to a given query. A page with twenty OCR regions offers no ranking mechanism; all regions are treated as equally plausible candidates.

We observe that these paradigms are complementary. ColPali's patch-level similarity scores encode *where* on a page the model attends when processing a query. This information is computed but discarded when returning page-level results. OCR systems know *what* content exists and *where* it is located, but not *why* it matters. By unifying these signals through spatial coordinate mapping, we achieve region-level retrieval: returning only the document regions that are both structurally coherent (via OCR) and semantically relevant (via VLM attention).

Crucially, our approach operates at *inference time* without additional training. Unlike Region-RAG (Li et al., 2025), which uses a hybrid training approach combining bounding box annotations with weakly-supervised signals from unlabeled data, our method leverages ColPali's emergent patch attention as a post-hoc spatial filter. This provides flexibility: the same approach works with any OCR system providing bounding boxes and any ColPali-family model.

1.1 Contributions

This paper presents a hybrid architecture for spatially-grounded document retrieval:

1. **Coordinate Mapping Formalism.** We formalize the mathematical correspondence between vision transformer patch grids and OCR bounding boxes, enabling spatial alignment between heterogeneous representations (Section 3.2).

2. **Relevance Propagation via Interpretability Maps.** We repurpose ColPali's late interaction mechanism to generate per-query-token similarity heatmaps, then propagate these scores to OCR regions through IoU-weighted patch-region intersection (Section 3.3).

3. **Two-Stage Retrieval Architecture.** We introduce a mean-pooling strategy that compresses patch embeddings along spatial axes, enabling efficient candidate retrieval before full-resolution reranking (Section 3.4).

4. **Theoretical Analysis.** We establish bounds on localization precision as a function of patch resolution, derive expected context reduction and precision amplification factors, and analyze computational complexity tradeoffs (Section 4).

5. **Empirical Validation and Open Implementation.** We evaluate on BBox-DocVQA, demonstrating 62.3% hit rate at IoU@0.5 with 54% token savings versus full-page retrieval (Section 6). We release Snappy, a complete open-source system implementing this architecture (Section 5).

Empirical evaluation on BBox-DocVQA validates that region-level retrieval achieves strong spatial grounding, with 62.3% of retrieved regions achieving IoU ≥ 0.5 with ground-truth evidence bounding boxes, while reducing context tokens by over 54% compared to full-page retrieval. These results confirm that retrieval granularity directly impacts both precision and efficiency.

2 Background and Related Work

2.1 Vision-Language Document Retrieval

ColPali and Late Interaction. ColPali (Faysse et al., 2025) represents the state-of-the-art in visual document retrieval. Built on a SigLIP-So400m vision encoder, it produces 1,024 patch embeddings per page (32×32 grid over 448×448 input resolution), each projected to 128 dimensions via a language model projection layer. Unlike single-vector approaches that pool visual features into a single embedding, ColPali preserves patch-level granularity and computes relevance through MaxSim, summing the maximum similarity between each query token and all document patches. This late interaction mechanism, inherited from ColBERT (Khattab and Zaharia, 2020), enables fine-grained matching while remaining computationally tractable for retrieval at scale.

The ViDoRe benchmark (Faysse et al., 2025) evaluates visual document retrieval across diverse domains, measuring NDCG@5 for page-level retrieval. ColPali achieves strong performance, but the benchmark, like the model, operates at page granularity, leaving region-level retrieval unexplored.

ColModernVBERT. Recent work introduces ColModernVBERT (Teiletche et al., 2025), a 250M-parameter late-interaction retriever achieving within 0.6 NDCG@5 of ColPali at over $10\times$ smaller size. The architecture maintains patch-level embeddings, making our approach directly applicable to this more efficient variant.

2.2 Layout-Aware Document Understanding

The LayoutLM family (Xu et al., 2020, 2021; Huang et al., 2022) pioneered joint modeling of text, layout, and visual features for document understanding. LayoutLMv3 introduced Word-Patch Alignment (WPA) pre-training, which predicts whether image patches corresponding to text words are masked. While conceptually related to our patch-OCR alignment, WPA operates at *pre-training time* to improve representations, whereas our approach uses patch similarities at *inference time* for retrieval filtering. Critically, LayoutLM models are designed for document *understanding* tasks (NER, classification) rather than retrieval: they lack late interaction mechanisms and query-conditioned relevance scoring.

OCR-free approaches including Donut (Kim et al., 2022) and Pix2Struct (Lee et al., 2023) perform document understanding directly from pixels. UDoP (Tang et al., 2023) unifies vision, text, and layout modalities. These models excel at understanding but do not address the retrieval problem we target.

2.3 Region-Level Document Retrieval

RegionRAG. The closest existing work is RegionRAG (Li et al., 2025), which shifts retrieval from document-level to semantic region-level granularity. However, RegionRAG uses a hybrid approach combining bounding box annotations with weakly-supervised signals from unlabeled data and a region-level contrastive loss. Our approach is fundamentally different: we achieve region-level retrieval at *inference time* using ColPali's emergent patch attention, requiring no additional training or annotation.

DocVLM. DocVLM (Shpigel Nacson et al., 2024) integrates OCR into VLMs by compressing OCR features into learned queries. This represents the *opposite direction* of our approach. DocVLM adds OCR to enhance VLM understanding, whereas we use VLM patches to filter and score OCR output for retrieval.

Table 1 summarizes the positioning of our approach relative to prior work.

Table 1: Comparison with related approaches. Our approach is unique in achieving region-level retrieval at inference time without additional training, by propagating VLM patch similarities to OCR bounding boxes.

Method	Granularity	OCR Required	Training
ColPali	Page-level	No	Pre-trained
LayoutLM	Understanding	Yes	Pre-trained
RegionRAG	Region-level	Yes	Hybrid supervision
DocVLM	Understanding	Yes	Fine-tuning
Ours	**Region-level**	**Yes**	**Inference-time only**

3 Method

3.1 Problem Formulation

Given a query q and a document corpus $\mathcal{D}$ where each document $d \in \mathcal{D}$ consists of one or more pages, conventional visual document retrieval returns a ranked list of pages. We reformulate the problem as *region-level retrieval*: return a ranked list of (page, region) pairs where each region corresponds to a semantically coherent text block (paragraph, table, figure caption, etc.) extracted via OCR.

Let $\mathcal{P} = \{p_1, \ldots, p_n\}$ denote the set of pages in the corpus, where each page p_i has an associated set of OCR regions $\mathcal{R}(p_i) = \{r_1, \ldots, r_m\}$. Each region r_j is characterized by its bounding box $B(r_j) = (x_1, y_1, x_2, y_2)$ in pixel coordinates and its text content $T(r_j)$. Our goal is to compute a relevance score $\text{rel}(q, r_j)$ for each region that captures both semantic relevance and spatial grounding.

3.2 Coordinate Mapping: Patches to Bounding Boxes

ColPali encodes each page as a grid of $G \times G$ patch embeddings ($G = 32$ for ColPali-v1.3) over an input image of resolution $I \times I$ ($I = 448$). Each patch corresponds to an $s \times s$ pixel region where $s = I/G = 14$ pixels. Patches are indexed in raster scan order (left-to-right, top-to-bottom).

Definition 1 (Patch Coordinate Mapping). *For patch index $k \in \{0, \ldots, G^2 - 1\}$, the corresponding bounding box in pixel coordinates is:*

$$patch\_bbox(k) = (col \cdot s, row \cdot s, (col + 1) \cdot s, (row + 1) \cdot s) \tag{1}$$

where $row = \lfloor k/G \rfloor$ and $col = k \mod G$.

When the original document page has resolution (W, H) different from the model's input resolution $I \times I$, OCR bounding boxes must be scaled to the model's coordinate space:

$$B'(r) = \left(x_1 \cdot \frac{I}{W}, y_1 \cdot \frac{I}{H}, x_2 \cdot \frac{I}{W}, y_2 \cdot \frac{I}{H} \right) \tag{2}$$

3.3 Relevance Propagation via Patch Similarity

Given a query q tokenized into n tokens with embeddings $\{q_1, \ldots, q_n\}$ and a page with patch embeddings $\{d_1, \ldots, d_m\}$ ($m = G^2 = 1{,}024$), we compute the similarity matrix:

$$S \in \mathbb{R}^{n \times m} \quad \text{where} \quad S_{ij} = \text{sim}(q_i, d_j) \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity. Standard ColPali aggregates this into a page score via MaxSim:

$$\text{Score}_{\text{page}}(q, p) = \sum_i \max_j S_{ij} \tag{4}$$

We instead extract the spatial distribution of relevance by computing a per-patch score:

$$\text{score}_{\text{patch}}(j) = \max_i S_{ij} \tag{5}$$

This captures the maximum relevance of patch j to any query token, forming a spatial heatmap over the page.

Definition 2 (Region Relevance Score). *For OCR region r with scaled bounding box $B'(r)$, we propagate patch scores via IoU-weighted aggregation:*

$$rel(q, r) = \sum_j IoU(B'(r), patch\_bbox(j)) \cdot score_{patch}(j) \tag{6}$$

where the sum is over all patches j with non-zero intersection. This weights each patch's contribution by its spatial overlap with the region, ensuring that patches fully contained within the region contribute more than peripheral patches.

3.4 Two-Stage Retrieval Architecture

Computing full patch-level similarity for all pages in a large corpus is prohibitively expensive. We introduce a two-stage architecture that balances efficiency and precision.

Stage 1: Candidate Retrieval. We compress patch embeddings via mean pooling along spatial axes to obtain a single page-level embedding. This enables efficient approximate nearest neighbor search (via Qdrant) to retrieve top-K candidate pages. Note that pooling discards spatial information, which may impact recall for pages with small relevant regions (see Section 7 for discussion of this limitation).

Stage 2: Region Reranking. For each candidate page, we compute full patch-level similarity and propagate scores to OCR regions as described in Section 3.3. Regions are ranked by their relevance scores, and top-k regions are returned.

3.5 Aggregation Strategies

We consider alternative aggregation strategies for propagating patch scores to regions:

Max Aggregation:

$$\text{rel}_{\max}(q, r) = \max_{j \in \text{covered}(r)} \text{score}_{\text{patch}}(j) \tag{7}$$

Mean Aggregation:

$$\text{rel}_{\text{mean}}(q, r) = \frac{1}{|\text{covered}(r)|} \sum_{j \in \text{covered}(r)} \text{score}_{\text{patch}}(j) \tag{8}$$

IoU-Weighted (Default): Uses the region relevance score from Definition 2:

$$\text{rel}_{\text{IoU}}(q, r) = \sum_j \text{IoU}(B'(r), \text{patch\_bbox}(j)) \cdot \text{score}_{\text{patch}}(j) \tag{9}$$

The choice of aggregation strategy affects retrieval quality depending on region size and content density; we use IoU-weighted aggregation as the default based on its principled handling of partial patch overlaps.

4 Theoretical Analysis

4.1 Precision Bounds

The spatial precision of our approach is fundamentally bounded by patch resolution. We formalize this tradeoff.

Theorem 1 (Localization Precision Bound). *For an OCR region with bounding box of width w and height h (in model coordinates), and patch size s, the maximum achievable localization precision is:*

$$precision \leq \frac{w \cdot h}{(w + s) \cdot (h + s)} \tag{10}$$

Proof. Consider a region with bounding box B of dimensions $w \times h$. The set of patches intersecting B depends on B's alignment with the patch grid. In the worst case, B is positioned such that it intersects partial patches on all four edges. Let the region's top-left corner fall at position (x, y) within a patch. The region then spans from patch column $\lfloor x/s \rfloor$ to $\lfloor (x+w)/s \rfloor$ and from patch row $\lfloor y/s \rfloor$ to $\lfloor (y+h)/s \rfloor$. The number of intersecting patches is at most $\lceil w/s + 1 \rceil \cdot \lceil h/s + 1 \rceil$. The total area covered by these patches is at most $(w + s) \cdot (h + s)$, achieved when the region is maximally misaligned with patch boundaries. Since the region's area is $w \cdot h$, the precision (ratio of relevant area to retrieved area) is bounded by $(w \cdot h)/((w + s) \cdot (h + s))$. □

Corollary 1. *For ColPali with $s = 14$ pixels at 448×448 resolution:*

- *A typical paragraph region (200×50 pixels): precision $\leq 73\%$*

- *A table cell (100×30 pixels): precision $\leq 60\%$*

- *A small label (50×20 pixels): precision $\leq 46\%$*

This analysis reveals that smaller regions suffer disproportionately from patch quantization. Applications requiring fine-grained localization (e.g., form field extraction) may benefit from higher-resolution patch grids or multi-scale approaches.

4.2 Computational Complexity

Let N = number of pages, M = average OCR regions per page, G^2 = patches per page, n = query tokens, d = embedding dimension.
   Page-level retrieval (baseline): $O(N \cdot n \cdot G^2 \cdot d)$ for full MaxSim over all pages.
   Our two-stage approach:

- Stage 1: $O(N \cdot d)$ for ANN search with pooled embeddings

- Stage 2: $O(K \cdot n \cdot G^2 \cdot d + K \cdot M \cdot G^2)$ for full similarity on K candidates plus region scoring

For $K \ll N$, the two-stage approach provides substantial speedup. With typical values ($N = $ 100,000 pages, $K = 100$, $G = 32$, $n = 20$, $d = 128$, $M = 15$), Stage 1 reduces the search space by $1000\times$ before the more expensive region-level computation.

4.3 Expected Performance Bounds

4.3.1 Context Reduction

Let A_p denote the total page area and A_r the area of the relevant region containing the answer. For a page with M OCR regions of average area $\bar{A}$, we have $A_p \approx M \cdot \bar{A}$.

Theorem 2 (Context Reduction Bound). *Let k be the number of top-scoring regions returned by our hybrid approach. The expected context reduction factor relative to page-level retrieval is:*

$$CRF = \frac{A_p}{\sum_{i=1}^{k} A_{r_i}} \geq \frac{M}{k} \tag{11}$$

with equality when all regions have equal area.

Proof. Page-level retrieval returns context proportional to A_p. Our approach returns context proportional to the total area of the top-k regions, $\sum_{i=1}^{k} A_{r_i}$. Since each region has area at most A_p and there are M regions total, selecting k regions yields area at most $k \cdot \bar{A} = k \cdot A_p/M$. The ratio $A_p/(k \cdot A_p/M) = M/k$ provides the lower bound. □

Corollary 2 (Token Savings). *For typical document parameters ($M = 15$ regions per page, $k = 3$ returned regions), the expected context reduction factor is at least $5\times$. This translates directly to proportional reductions in:*

- *LLM inference cost (tokens processed)*

- *Response latency (context length)*

- *Attention dilution (irrelevant content in context window)*

4.3.2 Precision Amplification

We model retrieval precision as the probability that a returned region contains relevant content.

Definition 3 (Region Relevance). *Let $R_j \in \{0, 1\}$ indicate whether region j contains content relevant to query q. For a typical factoid query, we assume $\sum_j R_j \in \{1, 2\}$ (one or two relevant regions per page).*

Theorem 3 (Precision Comparison). *Let $\rho = \sum_j R_j/M$ denote the fraction of relevant regions on a page. Then:*
(i) OCR-only (random selection of k regions):

$$\mathbb{E}[Precision_{OCR}] = \rho \tag{12}$$

(ii) ColPali page-level:

$$Precision_{ColPali} = \rho \quad \text{(all regions returned implicitly)} \tag{13}$$

(iii) Hybrid with score-based ranking: If patch scores s_j satisfy $\mathbb{E}[s_j|R_j = 1] > \mathbb{E}[s_j|R_j = 0]$ (relevant regions score higher on average), then selecting top-k by score yields:

$$\mathbb{E}[Precision_{Hybrid}] > \rho \tag{14}$$

with the improvement depending on the score separation between relevant and irrelevant regions.

Proof. Parts (i) and (ii) follow directly from the definition of ρ. For (iii), score-based ranking is equivalent to a classifier with ROC curve above the diagonal when relevant regions have higher expected scores. By the Neyman-Pearson lemma, thresholding on any statistic positively correlated with relevance improves precision over random selection. □

4.3.3 Signal-to-Noise Ratio

We formalize the intuition that region-level retrieval improves the "signal" available to downstream LLMs.

Definition 4 (Retrieval SNR). *For retrieved context C, define:*

$$SNR(C) = \frac{|C \cap Relevant|}{|C \cap Irrelevant|} \tag{15}$$

where $|\cdot|$ denotes token count.

Theorem 4 (SNR Improvement). *Under the assumption that relevant and irrelevant regions have similar average token counts:*
  (i) Page-level retrieval: $SNR_{page} = \rho/(1-\rho)$
  (ii) Hybrid top-k retrieval with precision P_k: $SNR_{hybrid} = P_k/(1-P_k)$
  The SNR improvement factor is:

$$\frac{SNR_{hybrid}}{SNR_{page}} = \frac{P_k(1-\rho)}{\rho(1-P_k)} \tag{16}$$

Proof. For page-level retrieval, the fraction of relevant tokens equals ρ by definition. For hybrid retrieval with precision P_k, the expected fraction of relevant tokens among returned regions is P_k. The SNR ratio follows from algebraic manipulation. $\square$

4.3.4 Summary of Expected Gains

Table 2 provides a unified view of the cost-quality tradeoff, showing that our hybrid approach achieves the context efficiency of OCR-based selection while providing the relevance ranking that OCR alone cannot offer.

Table 2: Combined efficiency-quality comparison. The hybrid approach uniquely achieves both low context cost and high precision. Precision values for our approach are empirically measured (IoU@0.5 on BBox-DocVQA).

Method	Context Cost	Precision (IoU@0.5)	Best Use Case
ColPali (page-level)	High (1.0×)	N/A (no grounding)	Page identification
OCR + Random	Low (0.2×)	~6.7% (1/15 regions)	Baseline
Hybrid (ours)	**Low (0.2×)**	**62.3%**	**Precise RAG**

5 Implementation: Snappy System

We implement our approach in Snappy, an open-source document retrieval system available at `https://github.com/athrael-soju/Snappy`. The system demonstrates the practical applicability of spatially-grounded hybrid retrieval.

5.1 Architecture Overview

Figure 1 illustrates the Snappy system architecture, showing the parallel document indexing pipeline and query processing pipeline that converge at region-level filtering.
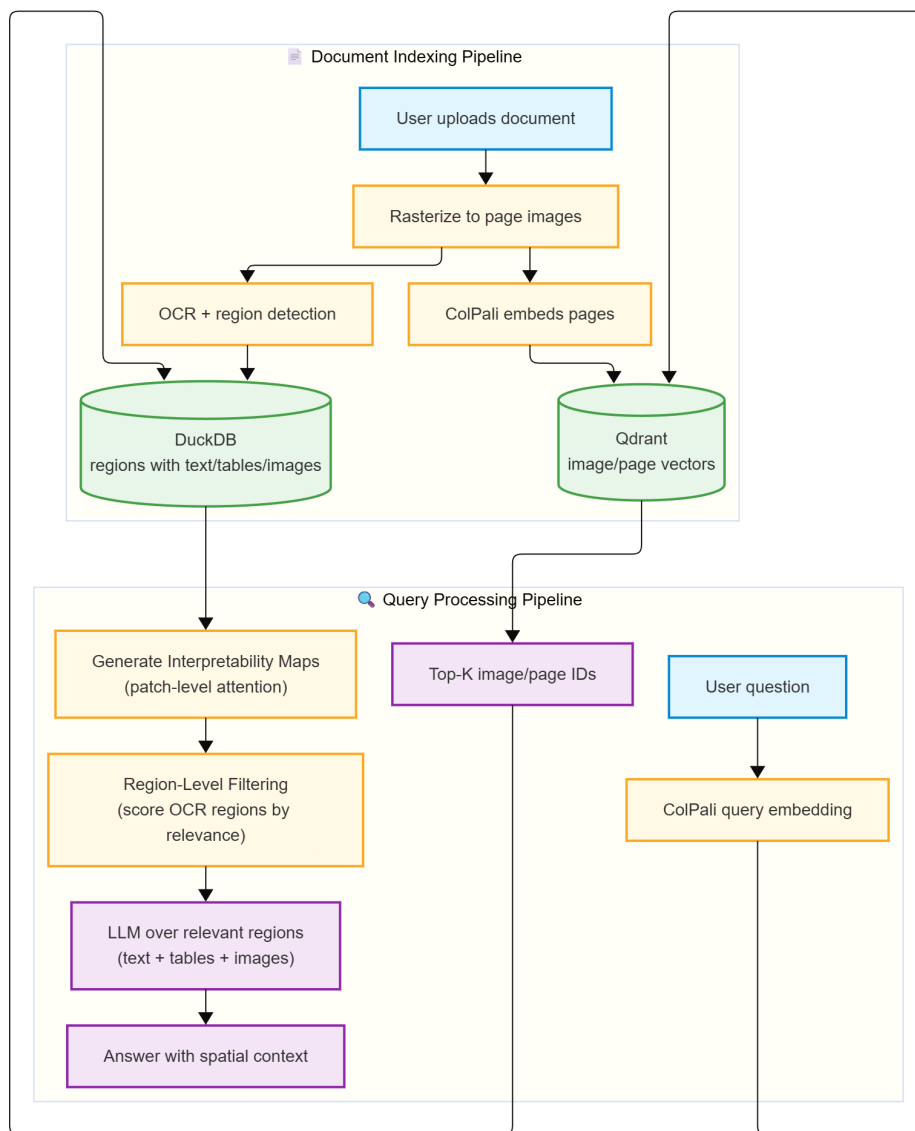

Figure 1: Snappy system architecture. The document indexing pipeline processes uploads through parallel OCR and ColPali embedding branches, storing regions in DuckDB and patch vectors in Qdrant. The query processing pipeline retrieves top-K candidates, generates interpretability maps from patch-level attention, and filters OCR regions by relevance to produce answers with spatial context.

Snappy pairs a FastAPI backend with a ColPali embedding service, DeepSeek OCR for visual grounding, DuckDB analytics for metadata storage, and a Next.js frontend. The system delivers hybrid vision+text retrieval over PDFs, where each page is rasterized, embedded as multivectors, and stored alongside images and OCR-extracted text in Qdrant.

The indexing pipeline processes documents as follows: (1) render each page as an image using

Poppler, (2) extract patch embeddings via the ColPali service, (3) optionally extract text regions with bounding boxes via DeepSeek OCR with visual grounding enabled, (4) store patch embeddings as multivectors and OCR metadata in Qdrant with page-level pooled embeddings for efficient retrieval, and (5) persist images to MinIO object storage for visualization.

At query time: (1) encode the query via ColPali's text encoder, (2) retrieve top-K candidate pages via ANN search on pooled embeddings in Qdrant, (3) compute full patch-level similarity for candidates using the stored multivectors, (4) propagate scores to OCR regions via IoU-weighted aggregation, and (5) return ranked regions with text content, bounding boxes, and optional visual highlighting.

5.2 Component Integration

The ColPali embedding service runs as a dedicated container exposing patch-level embeddings over HTTP, supporting both GPU and CPU deployment. Unlike standard ColPali deployments that return only page-level scores, Snappy's integration extracts the full ($n \times G^2$) similarity matrix stored as multivectors in Qdrant for downstream region scoring.

DeepSeek OCR provides text extraction with visual grounding capabilities. When enabled, the OCR service extracts markdown-formatted text alongside bounding box coordinates for each text region. OCR results are stored in DuckDB for SQL-based querying.

Qdrant stores both page-level pooled embeddings (for Stage 1 candidate retrieval) and full patch embeddings as multivectors (for Stage 2 region scoring).

6 Empirical Evaluation

We evaluate our approach on BBox-DocVQA, measuring spatial grounding accuracy and token efficiency. Our experiments address three questions: (1) How accurately does patch-to-region relevance propagation localize relevant content? (2) How does performance vary across document categories? (3) What token savings does the approach achieve?

Evaluation Scope. Our evaluation targets the core contribution: *region-level localization within a page*. Given a page containing the answer, can our approach identify the specific region? This complements page-level retrieval benchmarks like ViDoRe (Faysse et al., 2025), which measure whether the correct page is retrieved. We use IoU-based spatial grounding metrics to directly measure localization accuracy against ground-truth evidence bounding boxes.

6.1 Experimental Setup

Dataset. BBox-DocVQA (Yu et al., 2025) provides question-answer pairs across documents from arXiv categories including computer science (cs), economics (econ), electrical engineering (eess), mathematics (math), physics, quantitative biology (q-bio), quantitative finance (q-fin), and statistics (stat). Each QA pair includes ground-truth bounding boxes marking the evidence region containing the answer. We evaluate on 1,240 samples across eight categories.

Model. We evaluate ColQwen3-4B, a state-of-the-art 4B parameter model from the ColPali family.

Configuration. We use DeepSeek OCR with visual grounding in markdown mode. For region scoring, we apply percentile-based thresholding at percentile 50 with max token aggregation and max region scoring.

Metrics. For predicted region B_p and ground-truth bounding box B_g, we compute:

$$\text{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{17}$$

We report:

- **Mean IoU**: Average IoU between selected regions and ground-truth bounding boxes

- **Hit Rate@τ**: Fraction of samples where IoU $\geq \tau$, for $\tau \in \{0.25, 0.5, 0.7\}$

- **Token Savings**: Percentage reduction in context tokens, computed as $(T_{\text{baseline}} - T_{\text{method}})/T_{\text{baseline}}$, compared to (a) all OCR regions and (b) full image tokens

6.2 Main Results

Table 3 presents localization accuracy.

Table 3: Spatial grounding accuracy on BBox-DocVQA (n=1240). Hit rates indicate the fraction of samples achieving IoU at or above the specified threshold.

Model	Mean IoU	IoU@0.25	IoU@0.5	IoU@0.7
ColQwen3-4B (P50)	0.597	85.8%	62.3%	41.9%

The results demonstrate that ColPali's patch attention effectively localizes relevant regions without additional training. Over 85% of retrieved regions have meaningful overlap (IoU ≥ 0.25) with ground-truth evidence boxes, and 62.3% achieve substantial overlap (IoU ≥ 0.5).
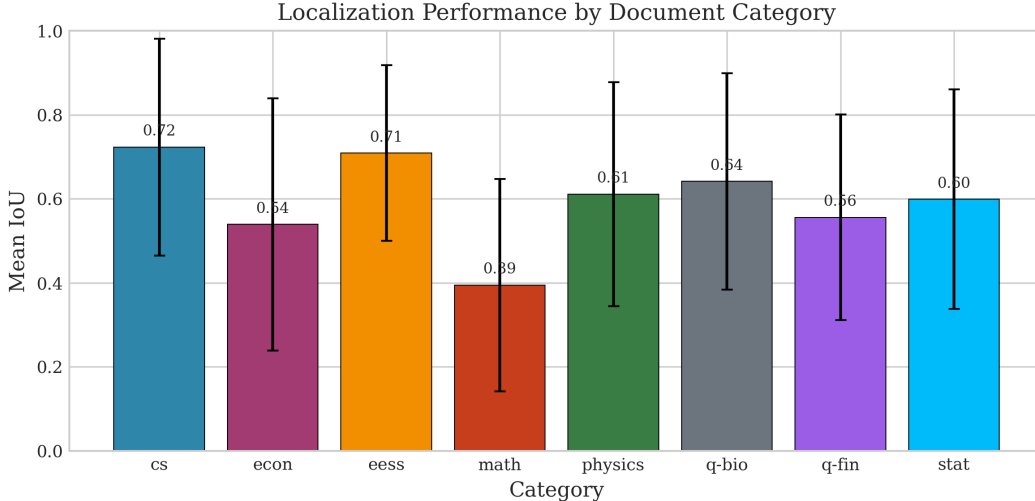

Figure 2: Localization accuracy by document category. Mean IoU varies substantially, with computer science and electrical engineering achieving highest accuracy while mathematics shows lowest performance due to dense equations and small regions.
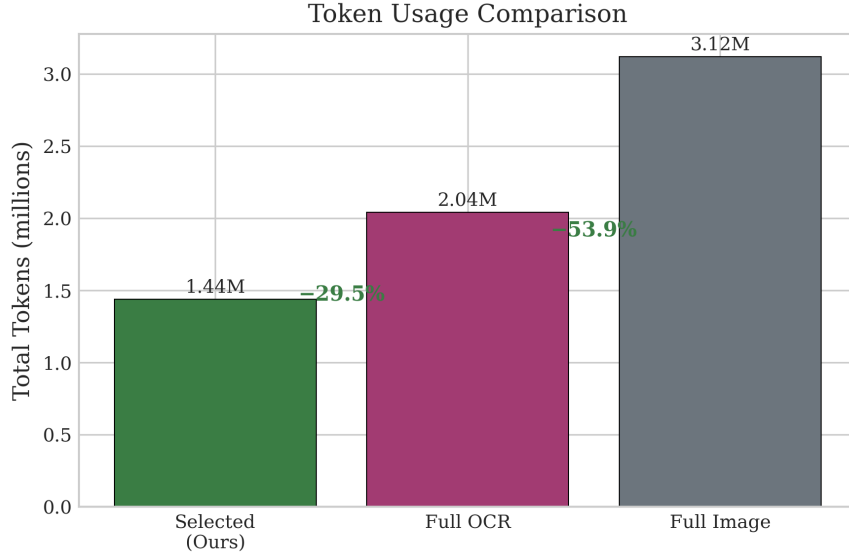
Figure 3: Token usage comparison. Our hybrid approach reduces context tokens by 30% vs. full OCR and 54% vs. full image tokens.

6.3 Category Analysis

Figure 2 and Table 4 present performance variation across document categories.

Table 4: Performance by document category. Mathematics and economics documents show notably lower accuracy, likely due to denser tabular content and smaller region sizes where patch quantization effects are more pronounced.

Category	N	Mean IoU	IoU@0.25	IoU@0.5	IoU@0.7
cs	176	0.723	94.9%	75.6%	65.3%
eess	156	0.709	96.8%	86.5%	58.3%
q-bio	137	0.641	91.2%	67.9%	46.0%
physics	173	0.611	88.4%	66.5%	39.9%
stat	101	0.588	89.1%	60.4%	39.6%
q-fin	176	0.556	86.4%	61.4%	32.4%
econ	174	0.539	77.6%	48.3%	36.8%
math	148	0.395	62.2%	29.7%	13.5%
Overall	**1240**	**0.597**	**85.8%**	**62.3%**	**41.9%**

Computer science and electrical engineering documents achieve substantially higher localization accuracy (Mean IoU > 0.7) than economics (0.539) and mathematics (0.395) documents, with physics (0.611) and quantitative biology (0.641) falling in between. This performance gap likely reflects differences in layout complexity: economics papers frequently contain dense tables with many small cells, while mathematics papers feature equations and formulas that may span multiple regions. These smaller regions suffer disproportionately from patch quantization effects, consistent with the theoretical precision bounds in Section 4.

6.4 Token Efficiency

Figure 3 visualizes the token savings achieved by our approach. We measure context tokens that would be passed to a downstream LLM. For full image retrieval, we estimate tokens using Claude's approximation: images are resized to fit within 1568×1568 pixels while maintaining aspect ratio, then token count is computed as $\lfloor (w \times h)/750 \rfloor$. For OCR-based methods, we tokenize extracted text using tiktoken's `cl100k_base` encoding (GPT-4/Claude compatible). Token counts are summed across all 1,240 evaluation samples. Table 5 presents token savings under different retrieval strategies.

Table 5: Token efficiency comparison. Our hybrid approach with percentile-50 filtering achieves substantial savings over both returning all OCR regions and using full image tokens.

Method	Total Tokens	vs All OCR	vs Full Image
Full Image (baseline)	3,068,485	–	–
All OCR Regions	2,006,575	–	34.6%
Hybrid (P50, filtered)	**1,411,708**	**29.6%**	**54.0%**

The hybrid approach with percentile-50 filtering reduces context tokens by 30% compared to returning all OCR regions and by 54% compared to full-page image tokens. These savings directly translate to reduced LLM inference costs, lower latency, and decreased attention dilution in downstream RAG applications.

7 Discussion

7.1 Limitations

Patch Resolution Bound. As analyzed in Section 4, small regions suffer from limited localization precision due to fixed patch granularity. For ColPali's 14-pixel patches, regions smaller than approximately 35×35 pixels achieve less than 50% precision. Future work could explore multi-scale patch embeddings or dynamic resolution based on document density.

OCR Dependency. Region quality depends on OCR accuracy and segmentation. Poorly segmented regions (merged paragraphs, split tables) degrade retrieval quality regardless of the VLM's spatial attention accuracy. Layout-aware OCR systems like DeepSeek OCR with visual grounding mitigate this, but OCR errors propagate to retrieval.

Emergent Attention Limitations. ColPali's patch attention is optimized for page-level retrieval through training, not explicitly for region-level localization. Attention may diffuse across semantically related but spatially distant regions (e.g., a header and its corresponding paragraph). We rely on OCR region boundaries to constrain this diffusion.

Two-Stage Information Loss. The mean-pooling in Stage 1 discards spatial information, potentially missing pages where relevant content is concentrated in small regions. Relevant pages could be filtered out before Stage 2 region scoring. Alternative pooling strategies (max pooling, attention-weighted pooling) may mitigate this.

Category-Specific Performance. Empirical evaluation reveals substantial performance variation across document types. Mathematics documents achieve only 29.7% hit rate at IoU@0.5 compared to 86.5% for electrical engineering documents. This gap reflects the patch quantization effects predicted by our theoretical analysis: mathematics papers contain equations and dense notation where relevant content spans small regions that are poorly aligned with the fixed patch

grid. Applications targeting specific document types should consider this variation when setting performance expectations.

7.2 Future Directions

Cross-Page Region Linking. Extending region-level retrieval to link semantically related regions across pages could enable quantitative search capabilities. For example, tracking a financial metric across quarterly reports by linking table cells containing that metric across documents.

Elimination of Image Storage. By storing only patch embeddings and OCR-extracted text with bounding boxes, our approach potentially eliminates the need for raw image storage after indexing. This could significantly reduce infrastructure costs for large document collections while maintaining retrieval capability.

Region-Aware Fine-Tuning. While our approach operates without additional training, fine-tuning ColPali with region-level supervision could improve spatial attention alignment. This would sacrifice our training-free advantage but potentially yield higher localization accuracy.

Legal Document Retrieval and Citation Grounding. In legal contexts, retrieval precision directly impacts downstream reliability. Recent empirical work demonstrates that even retrieval-augmented legal AI tools hallucinate 17–33% of responses, with errors compounded by coarse retrieval granularity (Magesh et al., 2025). Region-level retrieval with bounding box coordinates provides citation-bounded context: each retrieved region carries verifiable provenance, constraining generation to specific, locatable sources.

Model-Agnostic Scaling. Because our approach operates at inference time by composing outputs from independent late-interaction and OCR components, it automatically benefits from advances in either domain. More capable vision-language models with finer patch resolution or improved attention alignment will yield better spatial relevance signals, while more accurate OCR systems will produce higher-quality region boundaries, both translating directly to improved retrieval precision and token efficiency without architectural changes.

8 Conclusion

We presented a hybrid architecture for spatially-grounded document retrieval that unifies vision-language late interaction with structured OCR extraction. By formalizing the coordinate mapping between ColPali's patch grid and OCR bounding boxes, we enable region-level retrieval without additional training. Our approach operates entirely at inference time, providing flexibility across OCR systems and VLM backends.

Empirical evaluation on BBox-DocVQA validates our approach. Using ColQwen3-4B with percentile-50 thresholding, we achieve 62.3% hit rate at IoU@0.5, demonstrating that ColPali's patch attention effectively localizes relevant regions without additional training. Token savings of 54% compared to full-page retrieval confirm the practical efficiency gains predicted by our theoretical analysis (Theorem 2).

Category analysis across eight arXiv domains reveals systematic performance variation: computer science and electrical engineering documents achieve higher localization accuracy (Mean IoU > 0.7) than economics (0.54) and mathematics (0.40) documents, with physics and quantitative biology falling in between. This gap aligns with our theoretical precision bounds: smaller regions in dense tables and equations suffer disproportionately from patch quantization effects.

The two-stage architecture balances computational efficiency with region-level granularity, making the approach practical for large-scale document collections. We release Snappy as an open-source

implementation at `https://github.com/athrael-soju/Snappy`, demonstrating practical applicability for retrieval-augmented generation with reduced context windows and improved precision.

Acknowledgments

References

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.01449.

Omar Khattab and Matei Zaharia. 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, 2020.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2579–2591, 2021.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free Document Understanding Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–517, 2022.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18893–18912, 2023.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying Vision, Text, and Layout for Universal Document Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. RegionRAG: Region-level Retrieval-Augmented Generation for Visually-Rich Documents. arXiv:2510.27261, 2025.

Wenhan Yu, Wang Chen, Guanqiang Qi, Weikang Li, Yang Li, Lei Sha, Deguo Xia, and Jizhou Huang. BBox-DocVQA: A Large Scale Bounding Box Grounded Dataset for Enhancing Reasoning in Document Visual Question Answering. arXiv:2511.15090, 2025.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. DocVLM: Make Your VLM an Efficient Reader. arXiv:2412.08746, 2024.

Paul Teiletche, Quentin Macé, Max Conti, Antonio Loison, Gautier Viaud, Pierre Colombo, and Manuel Faysse. ModernVBERT: Towards Smaller Visual Document Retrievers. arXiv:2510.01149, 2025.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, 22:216, 2025.