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

Agathoklis Georgiou
Independent Researcher
athrael.soju@gmail.com

Abstract

Late-interaction multimodal retrieval models 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 **59.7% hit rate at IoU@0.5** (84.4% at IoU@0.25, 35.8% at IoU@0.7), with mean IoU of **0.569**. Our approach reduces context tokens by **29%** compared to returning all OCR regions and by **52%** 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.

Late-interaction retrievers 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 late-interaction 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 two-stage architecture that enables efficient candidate retrieval before full-resolution region-level 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 59.7% hit rate at IoU@0.5 with 52% token savings versus full-page retrieval (Section 6). We release Snappy, a complete open-source system implementing this architecture (Section 5).

6. **Model and Training Data Analysis.** We analyze how model architecture and training data affect patch-to-region localization, identifying two distinct regimes: a model-dependent regime where larger models yield substantial gains, and a precision-bound regime where patch quantization limits dominate regardless of model capacity (Section 6.6).

On BBox-DocVQA, 59.7% of retrieved regions achieve IoU ≥ 0.5 with ground-truth evidence bounding boxes, while reducing context tokens by 52% compared to full-page retrieval.

2 Background and Related Work

2.1 Late-Interaction Multimodal 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; region-level retrieval remains an unexplored mystery.

ColPali-Family Models. The ColPali architecture has spawned a family of late-interaction visual retrievers sharing the core patch-embedding approach. ColQwen3-4B (Huang and Tan, 2025) combines the ColPali framework with a Qwen3-based language model, achieving state-of-the-art performance on ViDoRe while maintaining the patch-level embeddings essential to our approach. At the efficiency frontier, ColModernVBERT (Teiletche et al., 2025) is a 250M-parameter variant achieving within 0.6 NDCG@5 of ColPali with $10\times$ fewer parameters. Our approach applies to any model in this family, as all preserve the patch-level similarity structure we exploit for region-level retrieval. We evaluate both ColQwen3-4B and ColModernVBERT to demonstrate this generality and explore the accuracy-efficiency tradeoff.

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. RegionRAG clusters salient patches via

BFS to produce visual region crops, requiring hybrid supervision (bounding box annotations plus pseudo-labels from unlabeled data) and a dual-objective contrastive loss. Our approach differs in three respects: (1) we operate at *inference time* without additional training, (2) we output text with spatial coordinates rather than image crops, enabling direct use by text-only LLMs, and (3) OCR-derived regions align with semantic document structure (paragraphs, tables, captions) rather than visual patch connectivity. As of December 2025, RegionRAG has no public implementation or model weights; we therefore compare against the ColPali-family baselines that both approaches share.

DocVLM. DocVLM (Shpigel Nacson et al., 2024) integrates OCR into vision-language models 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 late-interaction patch embeddings 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 late-interaction 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 N 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 to obtain page-level representations for efficient approximate nearest neighbor search, retrieving top-K candidate pages. This stage uses standard dense retrieval techniques; our contribution focuses on Stage 2.

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:*

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

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 and Signal-to-Noise Improvement

Score-based region ranking improves retrieval precision over random selection whenever patch scores correlate positively with relevance—an assumption validated empirically in Section 6. This precision improvement directly translates to better signal-to-noise ratio in retrieved context: selecting fewer, higher-relevance regions reduces the irrelevant content that downstream LLMs must process.

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×)**	**59.7%**	**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`.

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.

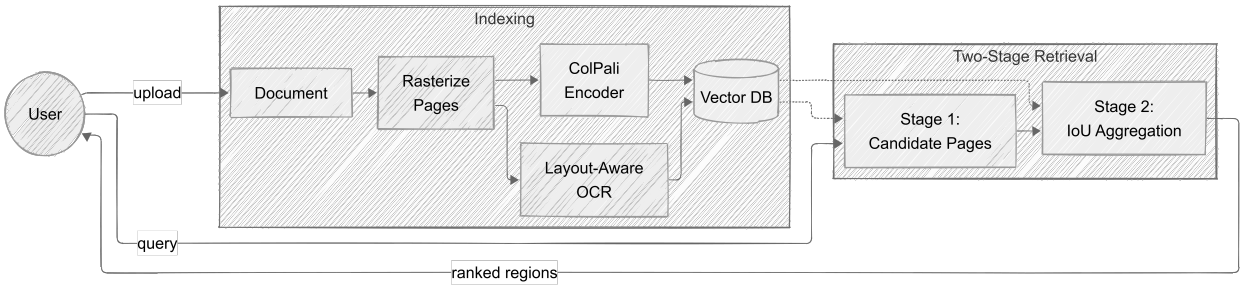

Figure 1: Snappy system architecture. The indexing pipeline processes documents through parallel OCR and embedding branches. The query pipeline retrieves candidates via ANN search, computes patch-level similarities, and filters OCR regions by relevance.

Snappy implements a two-stage retrieval pipeline over PDFs. Each page is rendered as an image, embedded via ColPali to produce patch-level multivectors, and optionally processed by layout-aware OCR to extract text regions with bounding boxes. Both patch embeddings and OCR metadata are stored in a vector database for retrieval.

At query time, the system: (1) encodes the query via ColPali's text encoder, (2) retrieves top-K candidate pages via ANN search on pooled embeddings, (3) computes full patch-level similarity for candidates, (4) propagates scores to OCR regions via IoU-weighted aggregation, and (5) returns ranked regions with text content and bounding boxes.

Unlike standard ColPali deployments that return only page-level scores, Snappy extracts the full $(n \times G^2)$ similarity matrix for downstream region scoring. The vector database stores both page-level pooled embeddings (Stage 1) and full patch multivectors (Stage 2).

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. We evaluate within-page localization (Stage 2); page-level retrieval accuracy (Stage 1) is measured by existing ViDoRe benchmarks and is outside our scope.

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 1,619 samples across eight categories (1,623 total samples minus 4 that consistently failed during OCR processing and are excluded from all reported metrics).

Models. We evaluate three ColPali-family models: ColQwen3-8B (8B parameters) and ColQwen3-4B (4B parameters), representing state-of-the-art accuracy at different scales, and ColModern-VBERT (250M parameters), representing the efficiency frontier.

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

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

$$\text{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{12}$$

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.

ColPali's patch attention localizes relevant regions without additional training. ColModern-VBERT (250M parameters) achieves 45.5% hit rate at IoU@0.5 with the highest token efficiency (38% savings vs OCR, 58% vs full image). Scaling to ColQwen3-4B improves localization substantially (59.7% at IoU@0.5), but further scaling to ColQwen3-8B yields negligible gains (59.8%), demonstrating diminishing returns from parameter scaling within the same architecture.

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

Model	N	Mean IoU	IoU@0.25	IoU@0.5	IoU@0.7
ColModernVBERT (P50)	1619	0.480	72.2%	45.5%	27.9%
ColQwen3-4B (P50)	1619	0.569	84.4%	59.7%	35.8%
ColQwen3-8B (P50)	1619	0.569	83.9%	59.8%	36.1%

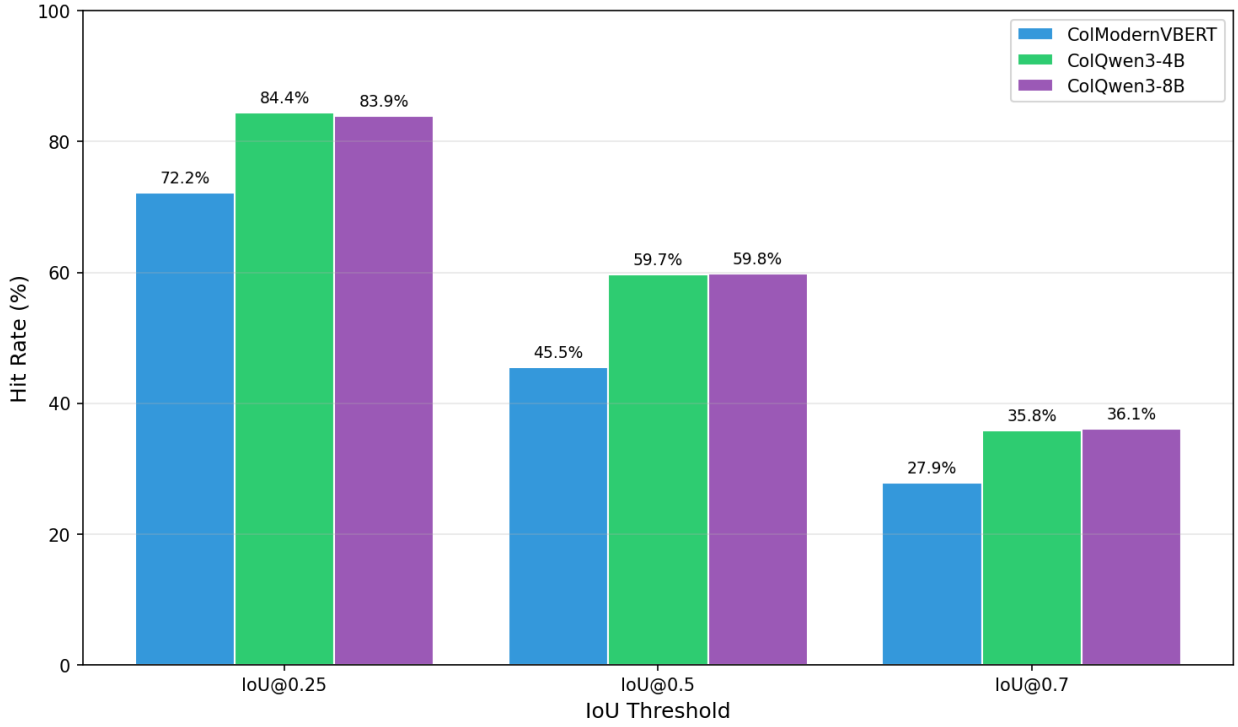

Figure 2: Hit rate comparison across IoU thresholds. ColQwen3-8B and ColQwen3-4B achieve nearly identical performance, while both consistently outperform ColModernVBERT, with the gap widening at stricter thresholds.

6.3 Category Analysis

Tables 4, 5, and 6 present performance variation across document categories for all three models, ordered by model size.

Computer science and electrical engineering documents achieve substantially higher localization accuracy with ColQwen3-4B (Mean IoU ≥ 0.65) than economics (0.513) and mathematics (0.382) documents, with physics (0.579) and quantitative biology (0.616) falling in between. ColModernVBERT shows a similar pattern but with uniformly lower scores; notably, quantitative biology achieves the highest accuracy (0.575) for the smaller model, while computer science drops from 0.697 to 0.527. 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.

Table 4: ColModernVBERT performance by document category. The smallest model (250M parameters) shows consistent performance degradation across all categories compared to ColQwen3 variants, with the largest gaps in computer science and electrical engineering.

Category	N	Mean IoU	IoU@0.25	IoU@0.5	IoU@0.7
q-bio	176	0.575	84.1%	59.1%	39.2%
cs	216	0.527	74.1%	49.5%	38.0%
physics	213	0.521	80.3%	53.5%	28.6%
stat	200	0.510	78.0%	50.0%	30.5%
q-fin	216	0.459	72.7%	45.8%	24.5%
eess	196	0.451	69.9%	42.9%	27.0%
econ	214	0.427	61.2%	36.0%	25.7%
math	188	0.370	58.0%	27.1%	10.1%
Overall	**1619**	**0.480**	**72.2%**	**45.5%**	**27.9%**

Table 5: ColQwen3-4B 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	216	0.697	94.9%	75.5%	56.9%
eess	196	0.656	94.4%	78.1%	46.9%
q-bio	176	0.616	89.8%	66.5%	42.6%
physics	213	0.579	86.9%	63.8%	34.3%
stat	200	0.559	86.5%	57.0%	31.5%
q-fin	216	0.543	86.1%	59.7%	29.2%
econ	214	0.513	75.2%	46.7%	31.8%
math	188	0.382	60.1%	28.7%	11.7%
Overall	**1619**	**0.569**	**84.4%**	**59.7%**	**35.8%**

Table 6: ColQwen3-8B performance by document category. The 8B model achieves nearly identical overall accuracy to the 4B variant, confirming that parameter scaling within the same architecture provides negligible localization gains.

Category	N	Mean IoU	IoU@0.25	IoU@0.5	IoU@0.7
cs	216	0.689	92.6%	74.5%	57.4%
eess	196	0.660	96.4%	79.1%	44.9%
q-bio	176	0.616	90.3%	66.5%	42.6%
physics	213	0.590	88.3%	65.7%	33.8%
stat	200	0.559	86.0%	57.0%	34.0%
q-fin	216	0.545	86.6%	60.2%	29.2%
econ	214	0.509	72.0%	46.3%	33.6%
math	188	0.378	58.5%	27.7%	12.2%
Overall	**1619**	**0.569**	**83.9%**	**59.8%**	**36.1%**

6.4 Token Efficiency

Figure 3 visualizes the token savings achieved by our approach across methods. 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 tokenizer, used as approximation). Token counts are summed across all 1,619 evaluation samples. Table 7 presents token savings under different retrieval strategies.

Table 7: 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)	4,003,039	–	–
All OCR Regions	2,678,723	–	33.1%
ColModernVBERT (P50)	1,661,684	38.0%	58.5%
ColQwen3-4B (P50)	**1,908,329**	**28.8%**	**52.3%**
ColQwen3-8B (P50)	1,974,263	26.3%	50.7%

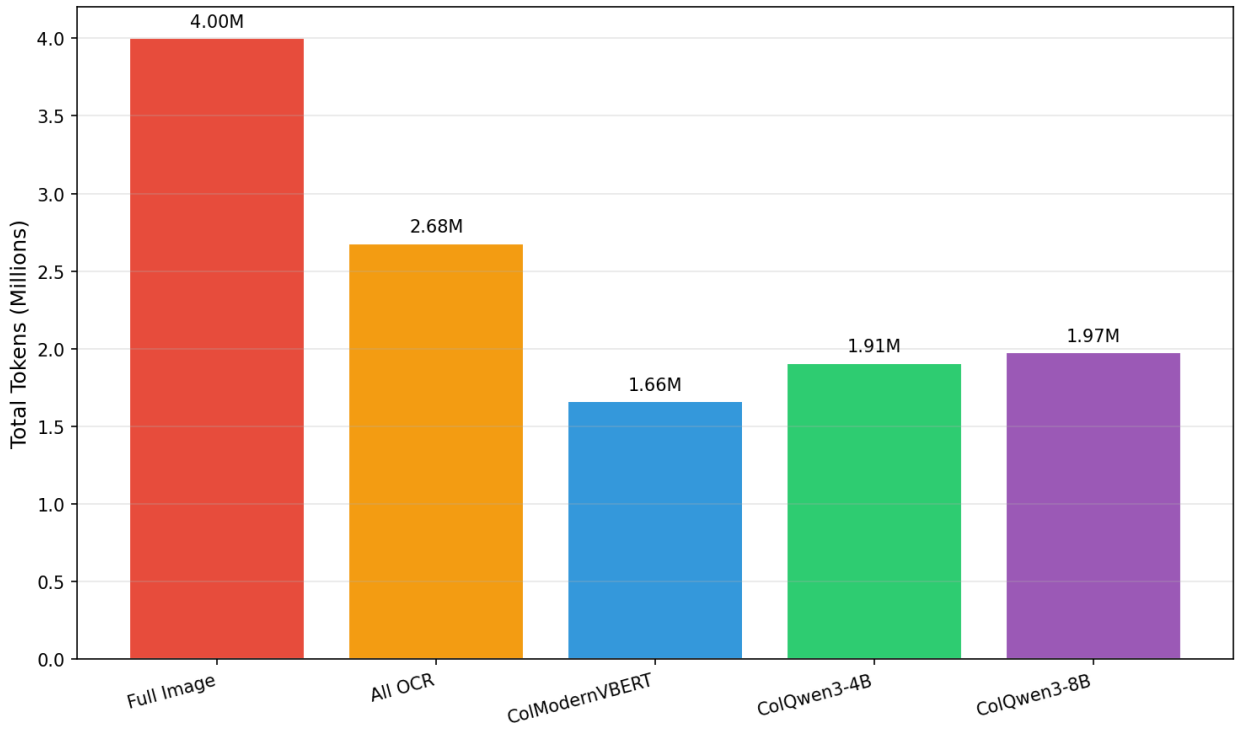

Figure 3: Token usage comparison across methods. All hybrid approaches achieve substantial reductions compared to full image (4.00M) and all OCR (2.68M) baselines, with ColModernVBERT offering the greatest efficiency at 1.66M tokens. ColQwen3-8B (1.97M) and ColQwen3-4B (1.91M) achieve similar token savings.

The hybrid approach with percentile-50 filtering reduces context tokens by 29% compared to

returning all OCR regions and by 52% compared to full-page image tokens. These savings directly reduce LLM inference costs and context window usage.

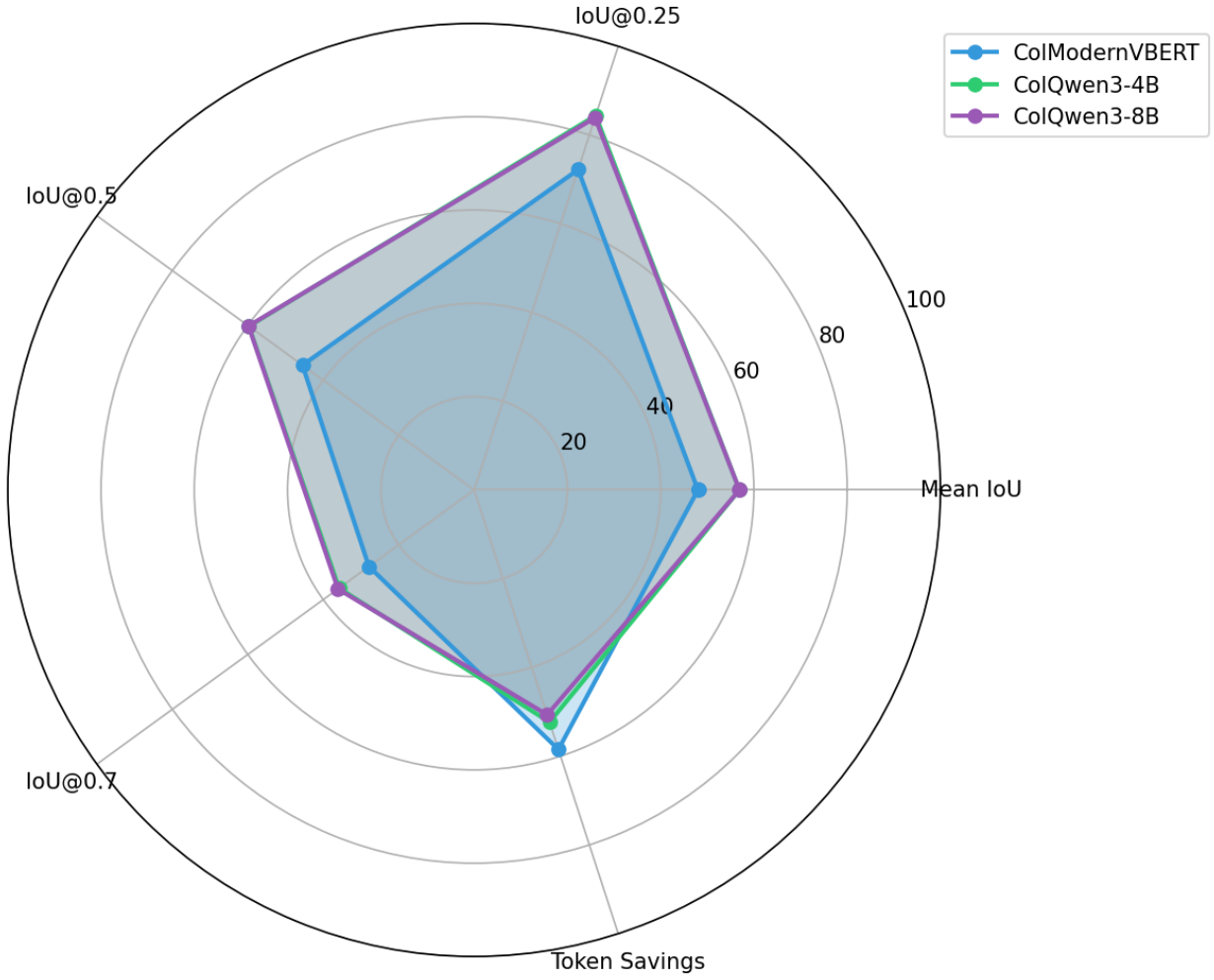

Figure 4: Multi-dimensional comparison of ColQwen3-8B, ColQwen3-4B, and ColModernVBERT. ColQwen3-8B and ColQwen3-4B achieve nearly identical accuracy metrics, while ColModernVBERT offers the best token savings. Mean IoU is scaled to percentage for visual consistency.

6.5 Threshold Selection

The percentile threshold controls the precision-efficiency tradeoff: lower thresholds retain more regions (higher recall), while higher thresholds improve token savings at the cost of potentially filtering relevant content. Table 8 characterizes this tradeoff on a stratified subset (N=405, 25% of dataset).

The tradeoff is monotonic and predictable: [describe observed pattern]. We use P50 throughout this paper as a balanced default, but practitioners should select based on application requirements. High-stakes domains (legal, medical) may prefer P25 to maximize recall; cost-sensitive deployments may prefer P75 for greater token savings.

Table 8: Threshold selection guide for ColQwen3-4B. Applications can tune the threshold based on whether they prioritize localization accuracy or token efficiency.

Threshold	IoU@0.5	Token Savings	Regions/Page
P25	XX.X%	XX.X%	X.X
P50	XX.X%	XX.X%	X.X
P75	XX.X%	XX.X%	X.X

6.6 Model Architecture and Training Data Effects

The substantial performance gap between ColQwen3-4B and ColModernVBERT (14.2 percentage points at IoU@0.5) invites analysis of how model architecture and training data affect patch-to-region relevance propagation. Notably, ColQwen3-8B achieves nearly identical performance to ColQwen3-4B (59.8% vs 59.7% at IoU@0.5), suggesting diminishing returns from parameter scaling within the same architecture family.

Architectural Differences. The models differ fundamentally in attention mechanism and vision encoder design. ColQwen3-4B uses a causal decoder architecture with Qwen3-VL's ∼675M parameter vision encoder (Qwen Team, 2025), which employs *DeepStack* multi-level feature fusion, injecting visual tokens from multiple intermediate ViT layers into corresponding decoder layers to preserve fine-grained details. The encoder uses Interleaved M-RoPE (Multimodal Rotary Position Embedding) that uniformly interleaves temporal, horizontal, and vertical positional components, providing balanced frequency allocation for 2D spatial layouts. ColModernVBERT instead combines a 150M-parameter ModernBERT (Warner et al., 2024) text encoder with SigLIP2-base-16b (Tschannen et al., 2025) (∼100M parameters), using *bidirectional attention* throughout. SigLIP2's vision encoder benefits from captioning-based pretraining, self-distillation, and masked prediction, producing denser feature representations optimized for localization.

Attention Mechanism Impact. The bidirectional versus causal attention distinction substantially affects late interaction quality. Bidirectional attention allows each token embedding to incorporate full context from both directions, producing richer representations for fine-grained MaxSim matching. In controlled experiments, bidirectional encoders demonstrate a +10.6 nDCG@5 advantage over causal decoders at equivalent parameter counts for retrieval tasks. This explains ColModernVBERT's ability to achieve within 0.6 nDCG@5 of models 10× larger on *page-level* retrieval benchmarks.

Patch Attention and Localization. For *region-level* localization, however, attention *sharpness* matters more than global context. Our relevance propagation mechanism (Section 3.3) computes per-patch scores via $\text{score}_{\text{patch}}(j) = \max_i S_{ij}$, then aggregates these to OCR regions. Localization accuracy depends on how tightly the model's attention concentrates on semantically relevant patches versus diffusing across the page. ColQwen3-4B's larger capacity produces sharper attention distributions that better discriminate relevant from irrelevant patches, while DeepStack's multi-scale features preserve both coarse layout structure and fine-grained text details in the patch embeddings. M-RoPE's explicit 2D spatial encoding further helps attention patterns respect spatial boundaries. These factors compound to explain the 14.2 percentage point IoU@0.5 gap: ColQwen3-4B's attention more precisely targets content regions, yielding higher overlap with ground-truth bounding boxes.

Training Data Composition. The models differ substantially in both pre-training scale and fine-tuning data:

- **ColQwen3-4B**: Built on Qwen3-VL pre-trained on ∼1.4 trillion multimodal tokens including

extensive OCR data, document formats, and VQA datasets. Fine-tuned on VDR (Visual Document Retrieval), ViDoRe-ColPali-Training, and VisRAG-Ret-Train (Shi et al., 2024), with 63% academic datasets (DocVQA, InfoVQA, TAT-DQA) and 37% synthetic query-document pairs.

- **ColModernVBERT**: Pre-trained on ∼2 billion text tokens with modality alignment on The Cauldron 2 and Docmatix (21% OCR/document content), then contrastively trained on 118k document-query pairs with 300k text-only pairs using a 2:1 text-to-image ratio.

The 700× pre-training scale difference provides ColQwen3-4B with inherently stronger document parsing and text-in-image understanding before retrieval-specific fine-tuning. However, ColModernVBERT's inclusion of text-only pairs during contrastive training provides cross-modal transfer benefits (+1.7 nDCG@5), demonstrating that the model learns text-image alignment even from pure text supervision.

Scaling Within Architecture Family. The near-identical performance of ColQwen3-8B and ColQwen3-4B (Mean IoU 0.569 for both; 59.8% vs 59.7% at IoU@0.5) reveals that doubling parameters within the same architecture provides negligible localization gains. This suggests that ColQwen3's localization accuracy is bounded by factors other than model capacity—likely patch resolution and attention pattern characteristics rather than representational capacity. The practical implication is clear: ColQwen3-4B offers the same localization quality at lower computational cost.

Category-Specific Performance Gaps. Table 9 reveals that the performance gap between ColQwen3-4B and ColModernVBERT varies systematically across document categories.

Table 9: Performance gap analysis between ColQwen3-4B and ColModernVBERT by category. ColQwen3-8B results (not shown) are within 0.01 Mean IoU of ColQwen3-4B across all categories. The gap narrows substantially for mathematics documents, suggesting both model families approach fundamental precision limits for small-region localization.

Category	ColQwen3-4B	ColModernVBERT	Δ Mean IoU	Δ IoU@0.5
cs	0.697	0.527	−0.170	−26.0 pp
eess	0.656	0.451	−0.205	−35.2 pp
q-bio	0.616	0.575	−0.041	−7.4 pp
physics	0.579	0.521	−0.058	−10.3 pp
stat	0.559	0.510	−0.049	−7.0 pp
q-fin	0.543	0.459	−0.084	−13.9 pp
econ	0.513	0.427	−0.086	−10.7 pp
math	0.382	0.370	−0.012	−1.6 pp
Overall	**0.569**	**0.480**	**−0.089**	**−14.2 pp**

The *smallest* performance gap occurs in mathematics documents ($\Delta = 0.012$ Mean IoU), where both models achieve their lowest absolute performance. This convergence suggests that mathematics documents present challenges that affect both models similarly—whether due to training data composition, difficulty parsing equations and dense notation, or other factors. The precise cause remains an open question for future work.

Conversely, the *largest* gaps appear in computer science ($\Delta = 0.170$) and electrical engineering ($\Delta = 0.205$) documents, where ColQwen3-4B's superior attention discrimination yields substantial improvements. These categories feature larger, well-separated regions (figures, code blocks, prose paragraphs) where model capacity translates directly to localization accuracy.

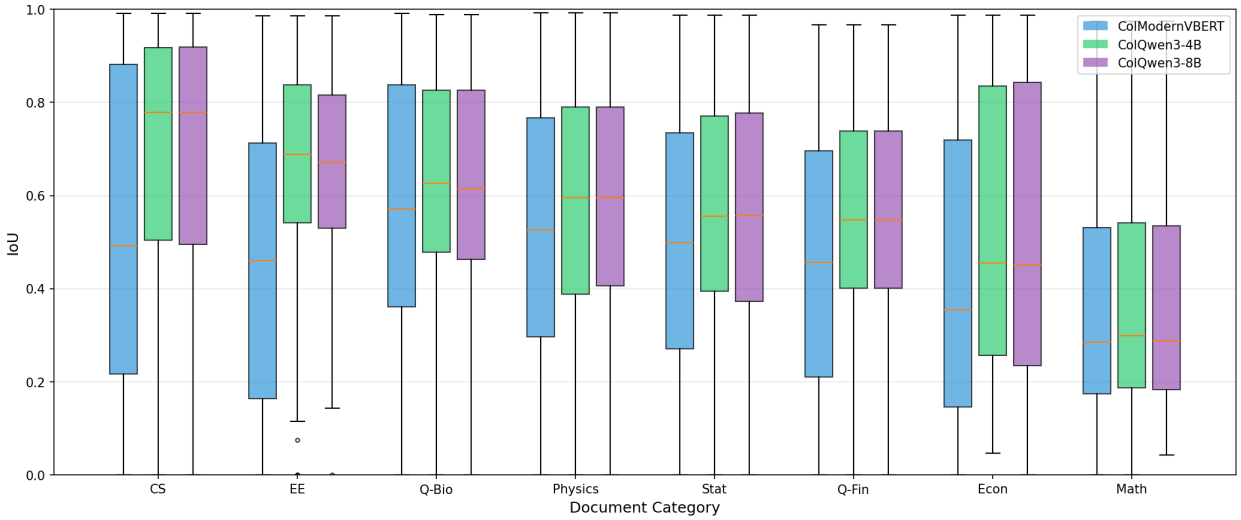

Figure 5: IoU distribution by document category for ColQwen3-8B, ColQwen3-4B, and ColModernVBERT. Box plots show median, quartiles, and outliers. Categories ordered by mean IoU (highest to lowest). Mathematics documents show consistently low performance across all models, while computer science achieves both higher mean and greater spread.

Implications for Model Selection. This analysis suggests three empirically distinct regimes:

1. **Architecture-sensitive documents**: For documents with large, well-separated regions (CS, EESS, Q-Bio), ColQwen3 models substantially outperform ColModernVBERT.
2. **Scale-insensitive within architecture**: ColQwen3-8B and ColQwen3-4B achieve nearly identical performance, indicating that doubling parameters within the same architecture yields negligible localization gains.
3. **Uniformly challenging documents**: For mathematics and dense tabular content, all models converge to similar (lower) performance, suggesting fundamental precision limits.

Applications should consider document characteristics when selecting models: ColModernVBERT's 16× parameter efficiency may be acceptable for mathematics-heavy corpora where all models achieve similar accuracy, while ColQwen3-4B offers the best accuracy-efficiency tradeoff for technical documentation with varied layouts. ColQwen3-8B provides no localization benefit over ColQwen3-4B despite doubled parameters.

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 retrieval model's spatial attention accuracy. Layout-aware OCR systems with visual grounding

mitigate this, but OCR errors propagate to retrieval. Note that OCR segmentation quality affects region boundaries but is orthogonal to our relevance scoring mechanism; our contribution is patch-to-region propagation, not OCR quality.

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.

Category-Specific Performance. Empirical evaluation reveals substantial performance variation across document types. Mathematics documents achieve only 28.7% hit rate at IoU@0.5 compared to 78.1% for electrical engineering documents. This gap likely reflects multiple factors: training data composition favoring certain document types, difficulty parsing equations and dense mathematical notation, and potential patch quantization effects for small regions. Applications targeting specific document types should consider this variation when setting performance expectations.

Within-Document Variance. Variance decomposition reveals that 72% of IoU variance is *within-document* (same document, different questions) versus 28% between documents. This suggests that question-specific semantic factors—rather than inherent document difficulty—dominate localization performance. A document may yield high IoU for one question and low IoU for another, depending on how well the query aligns with the model's learned attention patterns. This finding implies that per-category performance statistics represent averages over highly variable per-question outcomes.

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.

8 Conclusion

We presented a hybrid architecture for spatially-grounded document retrieval that unifies late-interaction retrieval 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 late-interaction retrieval backends.

On BBox-DocVQA, we evaluate three ColPali-family models. ColQwen3-8B and ColQwen3-4B achieve nearly identical localization accuracy (59.8% and 59.7% hit rate at IoU@0.5, respectively), while ColModernVBERT achieves 45.5%. All models provide substantial token savings compared to full-page retrieval (51–58%), consistent with the efficiency gains predicted by Theorem 2.

Category analysis across eight arXiv domains reveals systematic performance variation: computer science and electrical engineering documents achieve higher localization accuracy (Mean IoU ≥ 0.65) than economics (0.51) and mathematics (0.38) documents, with physics (0.58–0.59) and quantitative biology (0.62) falling in between.

Analysis of model architecture and scaling effects reveals three empirically distinct regimes. First, ColQwen3-8B and ColQwen3-4B achieve nearly identical performance, indicating that parameter scaling within the same architecture provides negligible localization gains. Second, for architecture-sensitive document types (computer science, electrical engineering), ColQwen3 models substantially outperform ColModernVBERT (up to 35 percentage points at IoU@0.5). Third, for uniformly challenging document types (mathematics), all models converge to similar performance, suggesting fundamental precision limits. This finding has practical implications: ColQwen3-4B offers the best accuracy-efficiency tradeoff, while ColQwen3-8B provides no localization benefit despite doubled parameters.

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

The author thanks the ColPali team (ILLUIN Technology) for their foundational work on late-interaction document retrieval and for maintaining an open research ecosystem that enabled this work. We also thank TomoroAI for releasing the ColQwen3 embedding models under Apache 2.0 license, and the Qwen team (Alibaba Cloud) for the Qwen3-VL and Qwen3-Embedding foundation models. The Snappy system implementation is available at `https://github.com/athrael-soju/Snappy`.

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.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Jieyu Zhang, Jiawei Gu, Ruhui Xue, Hongyu Lin, Xianpei Han, and Le Sun. VisRAG: Vision-based Retrieval-augmented Generation on Multimodality Documents. arXiv:2410.10594, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786, 2025.

Qwen Team. Qwen3 Technical Report. arXiv:2505.09388, 2025.

Xin Huang and Kye Min Tan. Beyond Text: Unlocking True Multimodal, End-to-end RAG with Tomoro ColQwen3. Tomoro.ai, 2025. https://tomoro.ai/insights/beyond-text-unlocking-true-multimodal-end-to-end-rag-with-tomoro-colqwen3.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv:2412.13663, 2024.