EVIDENCE-AWARE MAXSIM RERANKING FOR VISUALIZED DOCUMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Late-interaction retrieval models such as ColPali and ColQwen2 have recently shown strong performance for Visualized Document Retrieval (VDR) by matching query tokens against patch embeddings of document pages via MaxSim. However, we observe that MaxSim-based scoring often exhibits systematic ranking failures, where pages with peaked similarity or insufficient supporting evidence are over-ranked, leading to unreliable retrieval results. In this work, we propose **EA-MaxSim**, an evidence-aware reranking framework that augments MaxSim with two complementary signals: *query-level evidence coverage*, which measures how well a page supports the full query intent, and *page-level evidence richness*, which characterizes the structural sufficiency of supporting evidence within a page by leveraging spatial layout cues. EA-MaxSim operates as a lightweight reranking module, requiring no changes to model training, document encoding, or indexing. Extensive experiments on ViDoRe v2 and ViDoSeek demonstrate that EA-MaxSim consistently improves ranking quality across different backbones and datasets, particularly at top ranks. Further analyses show that evidence coverage provides a robust calibration signal, while evidence richness offers complementary improvements. Code is available at `https://anonymous.4open.science/r/EA-MaxSim-654D/`.

1 INTRODUCTION

Large language models (LLMs) increasingly rely on external retrieval to ground their generations and mitigate hallucinations (Lewis et al., 2020; Niu et al., 2024; Edge et al., 2024; Guo et al., 2024). While text-based retrieval has been extensively studied (Karpukhin et al., 2020; Reimers & Gurevych, 2019), many real-world documents—such as scientific papers, financial reports, and presentation slides—encode critical information in their visual presentation, including tables, figures, charts, and layout structure. This has motivated growing interest in *Visualized Document Retrieval* (VDR), which directly operates on document images to preserve visual semantics that are often lost during text extraction (Faysse et al., 2025; Yu et al., 2024).

Recent VDR systems built on large vision–language models have adopted late-interaction architectures to enable fine-grained visual–semantic matching. Methods such as ColPali and ColQwen2 (Faysse et al., 2025) encode each document page into hundreds of patch-level embeddings and compute relevance via *maximum similarity* (MaxSim) aggregation over query tokens. Despite their strong performance, we observe that vanilla MaxSim scoring exhibits systematic biases that limit retrieval reliability.

Evidence Coverage Bias. Under MaxSim, strong matches for a few query tokens can compensate for weak support for other query aspects. As a result, pages that only partially address a query may be ranked above those providing more comprehensive coverage. For example, a page matching only "neural networks" from the query "neural networks for time series forecasting" may outrank a page covering all query components with moderate similarity.

Evidence Richness Bias. Different query tokens can attain their maximum similarity within the same small visual region, yet MaxSim does not distinguish such concentrated matches from spatially distributed support. Consequently, a title-only slide or section header may obtain a high aggregate score while providing limited substantive information.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prior work on improving MaxSim-based retrieval has largely focused on efficiency, aiming to reduce memory or computation by pruning uninformative patches (Ma et al., 2025) or merging similar ones (Clavié et al., 2024). While effective for improving scalability, these approaches do not fundamentally alter how relevance is computed. Recent analyses (Ma et al., 2025) further indicate that patch-level informativeness is highly query-dependent, limiting the effectiveness of static pruning strategies. As a result, existing methods largely preserve the original MaxSim aggregation mechanism and do not explicitly model evidence sufficiency during scoring.

In this work, we propose **EA-MaxSim**, an *evidence-aware MaxSim reranking* framework that explicitly models evidence sufficiency while preserving the simplicity of late interaction. EA-MaxSim augments MaxSim with two complementary signals: **(1) Evidence Coverage**, which measures how comprehensively a page supports the full query intent by analyzing query token matches; and **(2) Evidence Richness**, which characterizes the structural sufficiency of supporting content within a page by leveraging spatial layout cues to account for redundancy. Our approach operates purely at the scoring level and requires no changes to model training, document encoding, or indexing infrastructure.

Our contributions are threefold: **(1)** We identify and formalize fundamental evidence sufficiency biases in MaxSim-based VDR systems. **(2)** We propose EA-MaxSim, a lightweight scoring framework that corrects these biases through evidence coverage and richness modeling. **(3)** We demonstrate improvements over strong VDR baselines across multiple benchmarks, compare against alternative aggregation and a dedicated VLM reranker, and characterize the dependence on calibration parameters.

2 Related Work

2.1 Visualized Document Retrieval

VDR has become essential for applications requiring understanding of document layout, structure, and visual elements (Mathew et al., 2021; Masry et al., 2022). Early approaches rely on OCR to extract text before retrieval, but often lose critical visual information such as tables, charts, and spatial layouts (Tanaka et al., 2023). Recent advances leverage vision-language models to directly process document images. Ma et al. (2024) propose DSE, which encodes entire pages into single vectors for efficient retrieval but lacks fine-grained matching. In contrast, ColPali (Faysse et al., 2025), inspired by ColBERT (Khattab & Zaharia, 2020), uses multi-vector representations with patch-level embeddings and MaxSim scoring, achieving state-of-the-art performance on benchmarks such as ViDoRe and ViDoSeek (Faysse et al., 2025; Wang et al., 2025).

Recent work extends VDR to more complex settings, including multi-page understanding, multi-document retrieval, and multimodal RAG (Chen et al., 2025; Cho et al., 2024; Yu et al., 2024). Among these systems, those based on late interaction retain MaxSim without explicitly calibrating evidence sufficiency.

2.2 Scoring Mechanisms in Multi-Vector Retrieval

The MaxSim aggregation strategy is effective for text retrieval in late-interaction models such as ColBERT (Khattab & Zaharia, 2020), where token embeddings correspond to distinct semantic units and maximum similarity captures the strongest relevance signal. However, extending MaxSim to visual documents introduces challenges due to *spatial redundancy* (e.g., repeated visual patterns across patches) and *compositional semantics*, where relevance depends on multiple complementary visual cues rather than a single strong match.

Despite these differences, most prior work retains the original MaxSim formulation. ColBERTv2 (Santhanam et al., 2022) improves efficiency via quantization without modifying scoring, while recent VDR systems (Faysse et al., 2025; Chen et al., 2025) directly apply MaxSim for patch-level matching, ignoring spatial redundancy and query-level evidence sufficiency. As a result, these approaches remain prone to ranking errors such as over-emphasizing partial matches or visually salient but low-content regions.

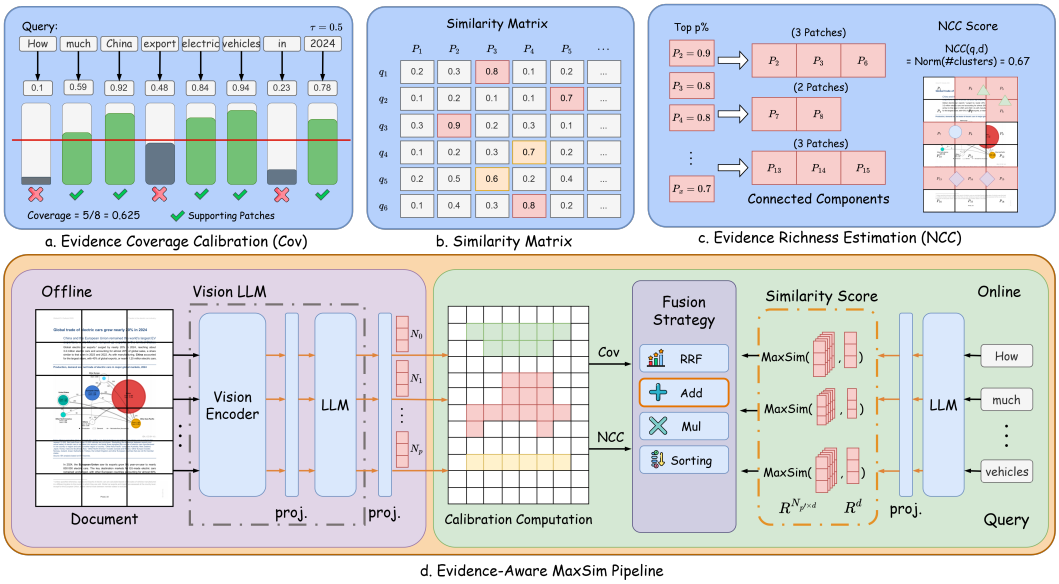


Figure 1: Overview of EA-MaxSim: (a) coverage, (b) similarities, (c) richness, and (d) scoring. Additive fusion is the default.

The closest work is Ma et al. (2025), which studies patch-level embedding reduction for VDR. Their analysis demonstrates that patch-level informativeness is highly query-dependent, explaining the limited effectiveness of static pruning strategies, and shows that token merging is more suitable for VDR. However, their focus is on document-side compression, without revisiting the MaxSim scoring function or modeling query-level coverage and redundancy at inference time.

Beyond patch reduction, other efficiency-focused methods explore quantization (Santhanam et al., 2022), adaptive token reduction (Chen et al., 2024; Xing et al., 2024; Yang et al., 2024), and token merging (Shang et al., 2024; Li et al., 2024; Clavié et al., 2024), while system-level works (Bach, 2025; Vespa Team, 2024; Scheerer et al., 2025) design specialized retrieval engines with two-stage pipelines for scalable multi-vector search: coarse retrieval using lightweight approximate computations, followed by exact multi-vector similarity computation and ranking (exact re-ranking) over a limited candidate set. While effective for improving scalability, these methods primarily target representation efficiency and leave relevance aggregation unchanged.

To our knowledge, this is the first systematic study of evidence-sufficiency-related ranking failures in MaxSim-based visualized document retrieval, coupled with scoring-level calibration through query coverage and spatial evidence richness.

3 Method

We present **EA-MaxSim**, an evidence-aware reranking framework that calibrates MaxSim with query-level evidence coverage (§3.2) and page-level evidence richness (§3.3). Figure 1 provides an overview. We first formalize the retrieval setting and scoring limitations, then define the two signals and their integration (§3.4).

3.1 Preliminaries and Problem Formulation

Problem Setup. Given a query q and a corpus of document pages $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, the goal of visualized document retrieval is to rank pages by their relevance to q. Following recent VDR systems (Faysse et al., 2025), we employ a vision–language model (e.g., ColQwen2) to encode the query into token-level embeddings $\mathbf{Q} = \{q_1, q_2, \ldots, q_{|Q|}\}$ and each document page into patch-level embeddings $\mathbf{P}_d = \{p_1^d, p_2^d, \ldots, p_{|P_d|}^d\}$, where $|Q|$ and $|P_d|$ denote the numbers of query tokens and patches of page d, respectively.

Vanilla MaxSim Scoring. The standard MaxSim scoring function computes relevance as

$$\text{MaxSim}(q, d) = \sum_{i=1}^{|Q|} \max_{j=1}^{|P_d|} \text{sim}(q_i, p_j^d), \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. For each query token q_i, MaxSim selects the most similar document patch and aggregates these maxima across all tokens. While effective in text-based late-interaction retrieval, this formulation exhibits systematic limitations when applied to visual documents.

Evidence Coverage Bias. Under MaxSim, strong matches for a few query tokens can compensate for weak support for other query aspects. As a result, pages that partially match a query may be ranked above those that address the full query intent.

Evidence Richness Bias. MaxSim sums one maximum per query token, rather than summing all matching patches. However, maxima for different query tokens can concentrate in adjacent patches of the same visual element (e.g., a title or figure). The resulting score does not explicitly distinguish concentrated matches from spatially distributed supporting content.

Together, these biases motivate the need for scoring mechanisms that explicitly account for *evidence sufficiency*, beyond raw similarity aggregation. Concrete examples of both failure modes are provided in Appendix C.

3.2 Evidence Coverage Calibration

To address the evidence coverage bias, we introduce a threshold-based coverage metric that explicitly measures how comprehensively a document addresses the complete query intent.

Coverage Metric. We define the coverage score as:

$$\text{Cov}(q, d) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbb{1}\left[\max_{j=1}^{|P_d|} \text{sim}(q_i, p_j^d) > \tau\right] \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function. This metric computes the proportion of query tokens that receive sufficient support (similarity exceeding τ) in the document, where τ is a similarity threshold. By counting supported tokens rather than their similarity magnitudes, the coverage term prevents exceptionally strong matches to a few tokens from compensating for missing support elsewhere in the query.

Query tokens, however, differ in how directly they express retrieval intent. Appendix B.8 therefore examines an optional query-aware token selection variant that focuses the coverage and richness signals on the tokens of LLM-extracted query keywords while leaving the original MaxSim score unchanged. Our main configuration retains all query tokens.

3.3 Evidence Richness Estimation

To address the evidence richness bias of MaxSim, we introduce a patch-centric mechanism to estimate the *effective evidence richness* of a document page. Our goal is to quantify the *spatially distinct supporting regions* a page provides for a given query, while discounting spatial redundancy caused by visually salient or repeated elements.

A straightforward estimate of richness is the number of patches whose query similarity exceeds a threshold. Yet neighboring patches may cover the same title, table header, or figure region, so counting them separately can overestimate the amount of distinct supporting content. We therefore estimate richness from spatially connected support regions rather than individual patches. Selecting this support set is a separate design choice: because absolute similarity magnitudes vary across pages, we use a fixed top percentage of patches ranked within each page instead of an absolute threshold.

Patch-Centric Support Selection. Given the query–page similarity matrix $\mathbf{S} \in \mathbb{R}^{|Q| \times |P_d|}$ with entries $s_{i,j} = \text{sim}(q_i, p_j^d)$, we assign each patch a salience score by taking its strongest alignment

with the query:

$$m_j(d) = \max_{i \in [1,|Q|]} s_{i,j}. \tag{3}$$

This score reflects how strongly a patch matches *any* aspect of the query. We then select the top-$p\%$ patches ranked by $m_j(d)$ as supporting patches:

$$\mathcal{S}(d) = \left\{ p_j^d : j \in \mathrm{TopIdx}_{p\%}\left(\{m_k(d)\}_{k=1}^{|P_d|} \right) \right\}, \tag{4}$$

where $\mathrm{TopIdx}_{p\%}$ returns the indices of the top $p\%$ scores; we fix $p = 10$ in our experiments. This selection identifies relatively salient patches within each page; their spatial grouping, defined next, determines the richness signal.

Spatial Redundancy Removal via Clustering. We group the selected supporting patches by spatial proximity to avoid counting adjacent patches of the same region independently. We construct an undirected graph over $\mathcal{S}(d)$, where each patch is a node and edges connect patches that are adjacent in the document layout. Specifically, let $\mathrm{pos}(p_j) = (x_j, y_j)$ denote the 2D grid coordinates of patch p_j; we define edges as

$$(p_j, p_k) \in E \quad \text{if} \quad \|\mathrm{pos}(p_j) - \mathrm{pos}(p_k)\|_\infty \le \epsilon, \tag{5}$$

where ϵ controls neighborhood connectivity (we use $\epsilon = 1$, corresponding to 8-connected adjacency). Connected components of this graph identify spatially contiguous support regions, which serve as a structural proxy rather than an exact segmentation of semantic evidence units. We discard components consisting of a single patch as well as components corresponding to background-only regions, as they typically do not convey meaningful information (Appendix A.1).

Normalized Evidence Richness. Let $c(q, d) = \#\text{clusters}(\mathcal{S}(d))$ denote the number of retained connected components. Since evidence richness is used as a *relative calibration signal* for reranking rather than an absolute relevance score, we normalize cluster counts across candidate pages:

$$\mathrm{NCC}(q, d) = \frac{c(q, d) - c_{\min}(q)}{c_{\max}(q) - c_{\min}(q)}, \tag{6}$$

where $c_{\min}(q)$ and $c_{\max}(q)$ are the minimum and maximum counts over the candidate set $\mathcal{C}(q)$, i.e., the top-K pages ranked by MaxSim ($K = 500$ in our experiments; Appendix A.2), with $c_{\max}(q) > c_{\min}(q)$. Pages with more retained support regions receive higher NCC scores. This layout-based proxy complements MaxSim similarity and query coverage without treating each selected patch as an independent evidence unit.

3.4 Integration into Scoring

EA-MaxSim integrates evidence coverage and evidence richness as auxiliary signals to calibrate the vanilla MaxSim score.

Additive Calibration. Among additive, multiplicative, reciprocal-rank, and constrained tie-breaking fusion (Appendix B.7), additive fusion performs most consistently across the evaluated backbones and cutoffs. We therefore use:

$$\mathrm{Score}(q, d) = \widehat{\mathrm{MaxSim}}(q, d) + \alpha \cdot \mathrm{Cov}(q, d) + \beta \cdot \mathrm{NCC}(q, d), \tag{7}$$

where $\widehat{\mathrm{MaxSim}}$ is the MaxSim score normalized by the number of query tokens, and α and β weight coverage and richness, respectively.

In this formulation, MaxSim serves as the primary relevance signal, while coverage and richness act as calibration terms that adjust score distributions. This design favors pages with more comprehensive support over partial matches driven by a few highly similar patches or spatially redundant evidence.

The two auxiliary signals summarize different aspects of the same query–page similarity matrix. Coverage asks whether each query token finds a sufficiently strong match anywhere on the page, whereas richness asks how the query-aligned patches are spatially organized. A page may contain several support regions that all address only one part of a query; conversely, a compact region may match several query tokens. Neither signal therefore determines the other. Their joint use lets the calibration distinguish support across query aspects from the spatial organization and redundancy of that support.

4 EXPERIMENTS

Table 1: **Performance comparison on the ViDoRe v2 benchmark.** We evaluate all methods using nDCG@K, Recall@K, and MRR@K with $K \in \{1, 3, 5\}$ across four datasets. Results are reported in percentage (%). For each backbone and each setting, the best performance is highlighted in **bold**.

Method	Biomedical Lectures			Economics Reports			ESG Reports (Human)			ESG Reports			Mean		
	nDCG	Recall	MRR	nDCG	Recall	MRR	nDCG	Recall	MRR	nDCG	Recall	MRR	nDCG	Recall	MRR
K=1															
ColPali	51.9	31.3	51.9	53.9	6.2	53.9	51.9	34.7	51.9	54.8	28.5	54.8	53.1	25.2	53.1
+ EA-MaxSim	**54.2**	**32.9**	**54.2**	**63.8**	**10.8**	**63.8**	**59.6**	**39.5**	**59.6**	**58.8**	**31.5**	**58.8**	**59.1**	**28.6**	**59.1**
ColQwen2	53.6	31.9	53.6	54.7	7.1	54.7	57.7	41.3	57.7	50.0	23.3	50.0	54.0	25.9	54.0
+ EA-MaxSim	**55.8**	**34.0**	**55.8**	**62.1**	**11.2**	**62.1**	**61.5**	**43.1**	**61.5**	**55.7**	**27.3**	**55.7**	**58.8**	**28.9**	**58.8**
K=3															
ColPali	53.8	**50.7**	61.9	51.3	18.4	65.7	60.0	61.9	65.1	53.7	47.5	65.1	54.7	44.6	64.4
+ EA-MaxSim	**54.5**	50.5	**63.2**	**56.2**	**19.1**	**71.1**	**62.9**	**62.7**	**68.3**	**55.4**	**47.7**	**67.2**	**57.2**	**45.0**	**67.5**
ColQwen2	54.2	50.6	63.0	54.2	20.0	66.3	59.1	56.2	65.7	51.0	45.6	59.7	54.6	43.1	63.7
+ EA-MaxSim	**55.4**	**51.4**	**64.7**	**55.8**	**20.1**	**69.3**	**62.0**	**59.7**	**68.9**	**53.1**	**46.6**	**63.5**	**56.6**	**44.4**	**66.6**
K=5															
ColPali	56.0	59.1	63.8	50.1	26.7	67.2	60.8	68.2	65.9	55.9	56.3	65.7	55.7	52.6	65.7
+ EA-MaxSim	**57.3**	**60.6**	**65.6**	**54.6**	**27.5**	**72.8**	**63.7**	**68.9**	**68.8**	**57.8**	**57.3**	**68.1**	**58.4**	**53.5**	**68.8**
ColQwen2	56.0	58.6	64.4	52.5	**29.0**	68.4	61.4	66.2	67.1	54.0	56.7	61.6	56.0	52.6	65.4
+ EA-MaxSim	**57.7**	**60.1**	**66.1**	**54.4**	28.2	**71.9**	**64.0**	**69.1**	**70.5**	**56.4**	**58.2**	**65.4**	**58.1**	**53.9**	**68.5**

Table 2: **Effect of EA-MaxSim on ViDoSeek.** We compare MaxSim and EA-MaxSim under two backbones. All metrics are reported in percentage (%). N@K, R@K, and M@K denote nDCG@K, Recall@K, and MRR@K, respectively.

Method	K	ColQwen2			ColPali		
		N@K	R@K	M@K	N@K	R@K	M@K
MaxSim	1	74.3	74.3	74.3	67.3	67.3	67.3
	3	83.8	**90.3**	81.5	78.0	85.4	75.4
	5	85.5	**94.5**	82.5	80.0	90.3	76.5
EA-MaxSim	1	**75.2**	**75.2**	**75.2**	**69.1**	**69.1**	**69.1**
	3	**84.1**	90.1	**82.0**	**78.6**	**85.6**	**76.3**
	5	**85.9**	94.4	**83.0**	**80.8**	**90.8**	**77.4**

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate our method on two recent benchmarks for visualized document retrieval, both of which require retrieval over a collection of documents rather than within a single document.

ViDoRe v2 is a benchmark constructed and released by the authors of ColPali for evaluating visualized document retrieval over PDF pages. It consists of four subsets covering diverse document categories, including structured reports and visually complex layouts. Each query is issued against a document collection and may be associated with *multiple* relevant ground-truth pages. We follow the official evaluation protocol and report results separately for each subset, as well as their macro-average.

ViDoSeek is a recently proposed benchmark designed for large-scale retrieval over a corpus of visual documents. In this setting, each query is evaluated against the entire document collection rather than a specific document. Compared to ViDoRe v2, queries in ViDoSeek are typically longer and more compositional, and each query is associated with a *single* ground-truth relevant page. This setting emphasizes the ability to identify pages that provide sufficient supporting evidence under more complex query semantics.

Baselines. We consider two strong patch-level late-interaction baselines: **ColPali v1.3**, a visualized document retrieval model based on patch-level MaxSim matching, and **ColQwen2 v1.0**, which adopts

a similar late-interaction mechanism with a different backbone. To assess the effectiveness of our reranking strategy, we apply **EA-MaxSim** on top of each baseline without modifying the underlying encoders or embeddings. All methods are evaluated under identical retrieval settings.

Evaluation Metrics. Following prior work, we report Recall@K, nDCG@K, and MRR@K with $K \in \{1, 3, 5\}$. For ViDoRe v2, metrics are computed for each subset individually, and we report their unweighted mean so that all four subsets contribute equally. For ViDoSeek, metrics are computed over the full corpus with a single relevant page per query.

Implementation Details. EA-MaxSim is a training-free reranking module applied to the top-500 MaxSim candidates of each query, which yields the same results as reranking all candidates. The support selection ratio is fixed at 10% across datasets. Coverage and richness parameters are selected on a held-out validation split (a random 20% of queries) for each dataset and then fixed for all queries, with no per-query or per-document tuning. Main results do not use query-aware token selection. Appendix A details the candidate pool and complexity; Appendix B.2 reports parameter and reranking-depth sensitivity, and Appendix B.5 reports efficiency.

4.2 Main Results

Overall Performance. Tables 1 and 2 summarize the main results on ViDoRe v2 and ViDoSeek, respectively. Across both benchmarks and backbones, **EA-MaxSim consistently improves ranking quality over the original MaxSim scoring**. On ViDoRe v2, macro-averaged nDCG@1 improves from 54.0 to 58.8 (+4.8 percentage points) for ColQwen2 and from 53.1 to 59.1 (+6.0 points) for ColPali, with gains also observed at larger cutoffs. These results indicate that the proposed evidence-aware reranking generalizes well across different patch-level late-interaction backbones. On ViDoRe v2, the nDCG and MRR improvements at $K \in \{1, 3, 5\}$ are significant for both backbones under paired randomization tests over queries ($p < 0.01$; Appendix B.6).

Dataset-Specific Analysis on ViDoRe v2. The magnitude of improvements varies across subsets, reflecting different prevalence of MaxSim failure modes. *Economics Reports* exhibit the largest gains (e.g., nDCG@1: 54.7 $\rightarrow$ 62.1, +7.4 points for ColQwen2), where queries often involve multiple semantic constraints and vanilla MaxSim tends to over-score pages matching only partial keywords. *Biomedical Lectures* show moderate but consistent improvements, benefiting from evidence sufficiency calibration under sparse-content pages with visually prominent titles. For both *ESG Reports (Human)* and *ESG Reports*, EA-MaxSim yields consistent improvements across cutoffs, suggesting that modeling evidence sufficiency improves top-ranked quality without relying on dataset-specific heuristics.

Results on ViDoSeek. ViDoSeek presents a more challenging setting with longer, more compositional queries and a single ground-truth relevant page per query. As shown in Table 2, EA-MaxSim achieves consistent improvements in nDCG@K and MRR@K for both ColQwen2 and ColPali across all evaluated cutoffs. For ColQwen2, EA-MaxSim improves nDCG@1 from 74.3 to 75.2 and MRR@5 from 82.5 to 83.0. For ColPali, the improvement is more pronounced, with nDCG@1 increasing from 67.3 to 69.1 and MRR@5 from 76.5 to 77.4. Recall@K remains largely comparable, with minor trade-offs at larger K, indicating that EA-MaxSim primarily corrects top-ranked false positives caused by peaked or insufficient evidence rather than expanding the recall pool. Moreover, the MaxSim baseline already reaches a Recall@5 of 94.5 (ColQwen2) and 90.3 (ColPali) on ViDoSeek, leaving little headroom at larger cutoffs.

4.3 Ablation Study

We conduct an ablation study on ViDoRe v2 to analyze the individual and combined effects of the two components in EA-MaxSim: *evidence coverage* and *evidence richness*. For clarity, all results are reported using the macro-average over the four ViDoRe v2 subsets.

Figure 2 shows that incorporating either component alone already improves over the original MaxSim scoring. Using only evidence richness consistently improves nDCG@K across both ColQwen2 and ColPali backbones, indicating that penalizing overly sparse or localized evidence helps mitigate peakiness caused by visually salient but information-poor regions. Similarly, modeling evidence coverage alone yields noticeable gains, suggesting that enforcing coverage over multiple query aspects improves overall ranking robustness.

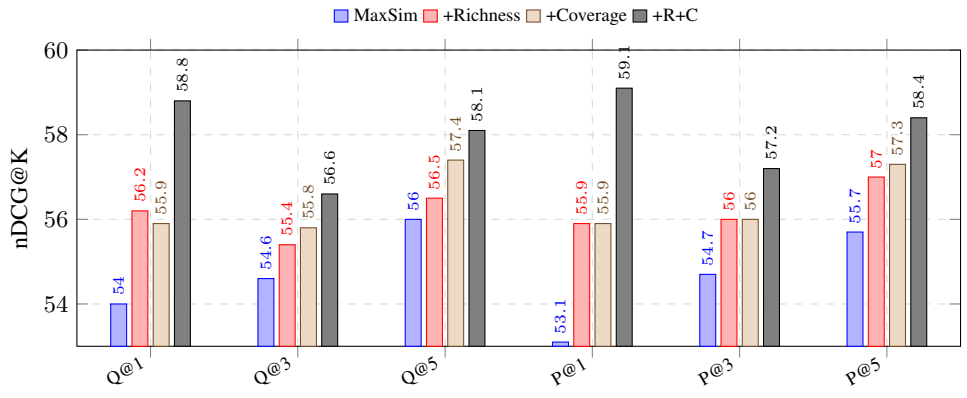


Figure 2: **Ablation on ViDoRe v2 (nDCG@K).** Q and P denote ColQwen2 and ColPali backbones. Coverage and richness provide consistent improvements, with their combination yielding the best performance across all K.

Importantly, combining evidence coverage and evidence richness consistently achieves the best performance across all reported metrics and cutoffs. Compared to using either component in isolation, EA-MaxSim yields additional improvements for both backbones, demonstrating that the two signals capture complementary aspects of evidence sufficiency. At $K = 1$, the combined score exceeds the stronger single component by 2.6 points for ColQwen2 (58.8 versus 56.2) and 3.2 points for ColPali (59.1 versus 55.9), quantifying their complementary contributions.

Alternative richness definitions and their different support-selection rules are examined in Appendix B.1. With one shared configuration across the four ViDoRe subsets for each backbone, nDCG@1 improves from 54.0 to 57.3 for ColQwen2 and from 53.1 to 56.8 for ColPali. Thus, gains do not require a separate configuration for every subset, although per-subset calibration provides a further margin (Appendix B.2).

4.4 Comparison with Alternative Aggregation Operators

Table 3: Alternative aggregation operators on ViDoRe v2. Entries are nDCG@1 / @3 / @5 (%). Softmax pooling and top-k averaging are shown at their best tested settings ($T = 0.02$ and $k = 2$ for both backbones); complete sweeps are provided in the appendix.

Method	ColQwen2	ColPali
MaxSim	54.0 / 54.6 / 56.0	53.1 / 54.7 / 55.7
Softmax pooling, $T = 0.02$	55.6 / 54.8 / 56.2	54.8 / 55.6 / 56.1
Top-k average, $k = 2$	53.9 / 54.3 / 55.6	52.9 / 53.8 / 55.2
Hierarchical merging, $f = 4$	49.0 / 51.3 / 52.3	47.0 / 49.0 / 51.7
EA-MaxSim	58.8 / 56.6 / 58.1	59.1 / 57.2 / 58.4

To test whether the gains arise from evidence-aware calibration rather than generic changes to aggregation, we compare EA-MaxSim with softmax-weighted patch pooling, top-k average pooling, and Light-ColPali hierarchical token merging (Ma et al., 2025) with compression factor four, using the same MaxSim candidate set and no query-token pruning. Pooling replaces the per-query-token maximum; hierarchical merging compresses document representations rather than estimating query-dependent spatial evidence. Softmax pooling takes a weighted average $\sum_j \mathrm{softmax}(s_j/T)\, s_j$ over the patch similarities s_j of each query token, whereas top-k pooling averages only its k highest patch similarities. Both operators recover MaxSim in the limit ($T \to 0$ or $k = 1$), so we sweep them from near-MaxSim settings ($T = 0.005$, $k = 2$) to strong smoothing ($T = 0.5$, $k = 10$). Table 3 reports the best-performing setting of each, alongside hierarchical merging at compression factor four; complete sweeps are in Appendix B.3. Mild softmax smoothing ($T = 0.02$) improves nDCG@1 over MaxSim by 1.6 points for ColQwen2 and 1.7 points for ColPali, with smaller changes at @3 and @5, whereas stronger smoothing, top-k averaging, and hierarchical merging all fall below MaxSim. EA-MaxSim exceeds the best pooling setting by 3.2 (ColQwen2) and 4.3 (ColPali) points

in nDCG@1 and by 1.6–2.3 points at @3 and @5. Thus, adjusting patch aggregation alone yields at most a small gain, whereas supplementing MaxSim with evidence signals yields a substantially larger one.

4.5 Comparison with a Dedicated VLM Reranker

We compare EA-MaxSim with MonoQwen2-VL-v0.1 (Chaffin & Lac, 2024), a dedicated document reranker, using ColQwen2 v1.0 on ViDoRe v2; all methods rerank the same MaxSim candidate list. For each query–page pair, the reranker performs query-conditioned VLM inference at retrieval time and scores relevance using normalized $P(\mathrm{True})$. Unlike EA-MaxSim, which calibrates existing query–patch similarities without another encoder pass, it jointly processes the query and candidate image with an additional model. Table 4 compares retrieval quality and added online latency (measurement details in Appendix B.4). EA-MaxSim reaches nDCG@1 of 58.8, versus 57.5–58.1 for the VLM at the displayed depths. Under the measured implementation and hardware settings, this query-conditioned inference incurs substantial online latency: reranking 5–20 candidates takes 1.74–9.82 s/query on an H100 with batch size eight, compared with 260 ms/query for EA-MaxSim on a single CPU thread. Reranking 20 candidates with MonoQwen2-VL achieves higher nDCG@5 (61.9 versus 58.1), but at a measured added latency of 9.82 s/query. The two approaches are also complementary: applying MonoQwen2-VL to the top candidates of EA-MaxSim achieves better results at lower cost. With only five candidates (2.00 s in total), the cascade reaches the highest nDCG@1 overall (59.6), and with ten candidates (4.98 s) it attains 58.2/61.2 in nDCG@1/5, surpassing MonoQwen2-VL applied directly to 100 candidates (57.6/61.1, 52.69 s; Table 9) at about one tenth of the latency.

Table 4: Dedicated VLM comparison on ViDoRe v2 with ColQwen2. Latencies are online increments beyond the shared MaxSim retrieval stage. Cascade (EA → VLM) latencies include both the EA and VLM stages.

Method	nDCG@1	nDCG@5	Stage hardware	Added latency /query
MaxSim	54.0	56.0	–	–
EA-MaxSim	58.8	58.1	CPU, one thread	260 ms
MonoQwen2-VL, top-5	58.1	57.7	H100-80GB	1.74 s
MonoQwen2-VL, top-10	57.5	60.2	H100-80GB	4.72 s
MonoQwen2-VL, top-20	57.8	61.9	H100-80GB	9.82 s
EA → MonoQwen2-VL, top-5	59.6	59.0	H100-80GB	2.00 s
EA → MonoQwen2-VL, top-10	58.2	61.2	H100-80GB	4.98 s

5 Conclusion

In this work, we revisit MaxSim-based late-interaction retrieval for visualized document retrieval and identify systematic ranking failures caused by peaked similarity and insufficient evidence. While MaxSim enables fine-grained matching, it tends to overemphasize partial matches and spatial redundancy, leading to unreliable ranking when queries require comprehensive coverage.

To address this, we propose **EA-MaxSim**, an evidence-aware reranking framework that augments MaxSim with two complementary signals: query-level evidence coverage and page-level evidence richness. These signals provide a simple yet effective calibration that encourages both comprehensive query support and non-redundant evidence.

EA-MaxSim requires no changes to encoder training, document encoding, or indexing. Experiments on ViDoRe v2 and ViDoSeek show improvements across backbones, especially at top ranks. Comparisons with pooling and token merging support the value of evidence-aware calibration, while shared-configuration experiments and the dedicated-reranker comparison delineate its parameter sensitivity and quality–computation trade-offs.

These findings motivate further exploration of evidence-aware scoring for late-interaction retrieval and its integration into downstream multimodal RAG systems.

Ethics Statement

This work uses only publicly available document retrieval benchmarks (ViDoRe v2 and ViDoSeek) and publicly released models, and it involves no human subjects or newly collected personal data. EA-MaxSim reranks retrieved pages using signals derived from existing model representations and does not generate content. We do not foresee risks beyond those of the underlying retrieval models, whose biases it may inherit. All authors have read and adhere to the ICLR Code of Ethics.

Reproducibility Statement

EA-MaxSim is training-free and builds on publicly released models and benchmarks. Section 4.1 specifies the backbones (ColPali v1.3[1] and ColQwen2 v1.0[2]), benchmarks, evaluation metrics, and parameter-selection protocol. Section 3 and Appendix A describe each component of the scoring function, including support selection, spatial clustering, background filtering, the candidate pool, and normalization, together with its complexity. Appendix B reports the shared configuration, the sensitivity to the weights, coverage threshold, and reranking depth, and the measurement protocols for the timing and VLM comparisons. Our code is available at the anonymous repository linked in the abstract.

AI Use Statement

Generative AI tools were used to assist with language editing to improve the readability and clarity of the manuscript, and minor LaTeX formatting and typesetting. The authors reviewed and verified all AI-assisted text and modifications to ensure that they accurately reflect the intended technical content. Generative AI tools were not used to develop the research methodology, translation, design experiments, interpret results, formulate scientific claims, or draw conclusions. We have reviewed all AI-assisted work through manual inspection by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Duong Bach. Hierarchical patch compression for colpali: Efficient multi-vector document retrieval with dynamic pruning and quantization. *arXiv preprint arXiv:2506.21601*, 2025.

Antoine Chaffin and Aurélien Lac. MonoQwen: Visual document reranking, 2024. URL `https://huggingface.co/lightonai/MonoQwen2-VL-v0.1`.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A Rossi, Changyou Chen, and Tong Sun. LoRA-contextualizing adaptation of large multimodal models for multi-page document understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 134–150, 2024.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3DocRAG: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Benjamin Clavié, Antoine Chaffin, and Griffin Adams. Reducing the footprint of multi-vector retrieval with minimal performance impact via token pooling. *arXiv preprint arXiv:2409.14683*, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

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*, 2025.

[1] `https://huggingface.co/vidore/colpali-v1.3-merged`

[2] `https://huggingface.co/vidore/colqwen2-v1.0-merged`

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

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*, pp. 39–48, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient visual projector for multimodal LLM. *arXiv preprint arXiv:2407.02392*, 2024.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505, 2024.

Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, and Aixin Sun. Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19568–19580, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. DocVQA: A dataset for VQA on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022.

Jan Luca Scheerer, Matei Zaharia, Christopher Potts, Gustavo Alonso, and Omar Khattab. Warp: An efficient engine for multi-vector retrieval. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*, pp. 2504–2512, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A dataset for document visual question answering on multiple images. In *AAAI Conference on Artificial Intelligence*, 2023.

Vespa Team. Vespa: The open big data serving engine. `https://vespa.ai`, 2024.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. PyramidDrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

A Implementation Details and Complexity

A.1 Background Filtering

For background filtering, we compute the mean and standard deviation of pixel intensities within each 28×28-pixel patch. A patch is treated as blank if its mean is at least 252 and its standard deviation is at most 1.0. A connected component is removed as background-only when all its patches satisfy this criterion.

A.2 Candidate Pool

For each query, EA-MaxSim reranks the candidate pool $\mathcal{C}(q)$ formed by the top-K pages under MaxSim, with $K = 500$ in all experiments. Coverage and richness are computed for every page in this pool, and cluster counts are min–max normalized over the same pool to obtain NCC. Appendix B.2.4 shows that $K = 500$ yields the same results as reranking all candidates.

A.3 Time Complexity Analysis

We analyze the time complexity of **EA-MaxSim**, which augments the standard MaxSim scoring with two auxiliary signals: *query-level evidence coverage* and *evidence richness*.

Notation. Let N denote the number of pages in the collection, over which MaxSim is computed, and let $K \leq N$ be the size of the reranking pool, over which coverage and richness are computed. Let Q be the number of query tokens (or query patches). Let D be the number of document patches per page after padding to a fixed length. Let d be the embedding dimension. For a page i, let $S_i \leq D$ denote the number of *supporting patches* selected by a similarity threshold or a top-$p\%$ rule.

Baseline MaxSim. For each candidate page, MaxSim computes a query–page similarity matrix via dot products, followed by a max-reduction over document patches for each query token. The dominant cost comes from similarity computation:

$$T_{\text{MaxSim}} = \mathcal{O}(N \cdot Q \cdot D \cdot d)\,, \tag{8}$$

while the max-reduction adds an additional $\mathcal{O}(N \cdot Q \cdot D)$, which is lower-order when d is non-trivial.

Query Coverage. Query coverage is computed by thresholding the per-token MaxSim scores and averaging (See Equation 2). Since EA-MaxSim reuses the per-token maxima already produced by MaxSim, the additional cost per page is $\mathcal{O}(Q)$, yielding

$$T_{\text{coverage}} = \mathcal{O}(K \cdot Q), \tag{9}$$

which is negligible compared to the baseline MaxSim cost.

Evidence Richness. To estimate information richness, EA-MaxSim analyzes the spatial distribution of supporting patches. Specifically, a connectivity graph is constructed among the S_i supporting patches based on their 2D layout proximity, and the number of connected components is computed. Using a naive pairwise distance computation, the cost is

$$T_{\text{richness}} = \sum_{i=1}^{K} \mathcal{O}\big(S_i^2\big)\,, \tag{10}$$

with a worst-case upper bound of $\mathcal{O}(K \cdot D^2)$. In our setting, the top-10% rule gives $S_i \approx 0.1D$, which keeps this term small in practice.

Overall Complexity. Combining all components, the total per-query complexity of EA-MaxSim is

$$T_{\text{EA-MaxSim}} = \mathcal{O}(N \cdot Q \cdot D \cdot d) + \mathcal{O}(K \cdot Q) + \sum_{i=1}^{K} \mathcal{O}\big(S_i^2\big)\,. \tag{11}$$

In practice, the N-dependence of MaxSim is largely amortized by batched matrix computation, where similarities between a query and multiple pages are evaluated jointly on GPU. The evidence-aware terms depend only on the pool size K, not on the collection size N, and the min–max statistics of the cluster counts are obtained over the pool in an additional linear pass. Appendix B.5 reports end-to-end retrieval timings.

B Additional Results

These additional experiments use ColQwen2 v1.0 and ColPali v1.3 on ViDoRe v2 and, where stated, ViDoSeek. ViDoRe values are macro-averaged over four subsets. Metrics are percentages. Offline parameter sweeps reuse recorded similarities and auxiliary statistics without re-encoding. Unless otherwise noted, EA-MaxSim entries use the main-result configurations and thus match Tables 1 and 2.

B.1 Richness Design and Background Filtering

Tables 5 and 6 compare alternative richness constructions on ColQwen2 and ColPali, respectively. The patch-count variant counts patches above a fixed absolute similarity threshold, since a top-percentage rule would yield identical counts for pages with equal-sized patch grids. The component-count variants instead group the top-10% salient patches into spatially connected regions. The no-filter variant retains spatial clustering without background removal (Appendix A.1).

For a fair comparison, each variant is tuned with its own richness signal under two protocols: *shared* rows use one configuration (α, β, τ) across the four subsets, and *per-subset* rows use subset-specific configurations as in the main results. For the full model, these rows correspond to Table 7 and Table 1, respectively.

The full design achieves the highest nDCG@1 under both protocols and the best nDCG@1/3/5 under the per-subset protocol on both backbones. Replacing connected-component counting with patch counting, or removing background filtering, lowers nDCG@1 by 0.2–0.9 points, confirming that both spatial grouping and background filtering contribute to the richness signal.

Table 5: Richness variants: ColQwen2, ViDoRe v2. Shared rows list the shared (α, β, τ); per-subset rows use subset-specific configurations.

Richness definition	protocol	@1	@3	@5
MaxSim baseline	–	54.0	54.6	56.0
No clustering (patch count)	shared (0.5, 0.1, 0.4)	56.8	55.8	56.7
No clustering (patch count)	per-subset	57.9	55.3	56.9
No bg filter (cluster count)	shared (0.5, 0.01, 0.4)	56.4	55.8	56.7
No bg filter (cluster count)	per-subset	58.2	55.4	57.0
Full (cluster + bg filter)	shared (0.6, 0.1, 0.4)	57.3	55.7	56.6
Full (cluster + bg filter)	per-subset	58.8	56.6	58.1

B.2 Hyperparameter Sensitivity and Robustness

B.2.1 Shared Configuration

This section expands on the dataset-level parameter selection described in Section 4.1. The four ViDoRe v2 subsets differ substantially in domain (biomedicine, economics, and corporate sustainability) and in document type (lecture slides versus long-form reports), so their best-performing settings naturally differ. We therefore also evaluate a much stricter setting: one shared configuration per backbone across all four subsets. Table 7 shows that even this single configuration consistently

Table 6: Richness variants: ColPali, ViDoRe v2. Shared rows list the shared (α, β, τ); per-subset rows use subset-specific configurations.

Richness definition	protocol	@1	@3	@5
MaxSim baseline	–	53.1	54.7	55.7
No clustering (patch count)	shared (0.6, 0.1, 0.4)	56.6	55.0	56.1
No clustering (patch count)	per-subset	58.2	56.2	57.2
No bg filter (cluster count)	shared (1.0, 0.01, 0.4)	56.3	54.5	55.6
No bg filter (cluster count)	per-subset	58.2	56.1	57.3
Full (cluster + bg filter)	shared (0.6, 0.1, 0.5)	56.8	55.9	57.3
Full (cluster + bg filter)	per-subset	59.1	57.2	58.4

improves over MaxSim at all three cutoffs, raising nDCG@1 by 3.3 points for ColQwen2 and 3.7 points for ColPali, while per-subset calibration provides a further margin.

Table 7: Shared versus per-subset configurations on ViDoRe v2. The per-subset row corresponds to the main results (Table 1).

Setting	ColQwen2 @1/@3/@5	ColPali @1/@3/@5
MaxSim	54.0 / 54.6 / 56.0	53.1 / 54.7 / 55.7
EA-MaxSim, one shared config	57.3 / 55.7 / 56.6	56.8 / 55.9 / 57.3
EA-MaxSim, per-subset configs	58.8 / 56.6 / 58.1	59.1 / 57.2 / 58.4

B.2.2 Coverage and Richness Weights

Figure 3 summarizes the sensitivity of EA-MaxSim to the weights of *evidence coverage* and *evidence richness* across datasets and backbones. All analyses are conducted on ViDoRe v2 using nDCG@5, which provides a balanced view between top-ranked precision and ranking stability.

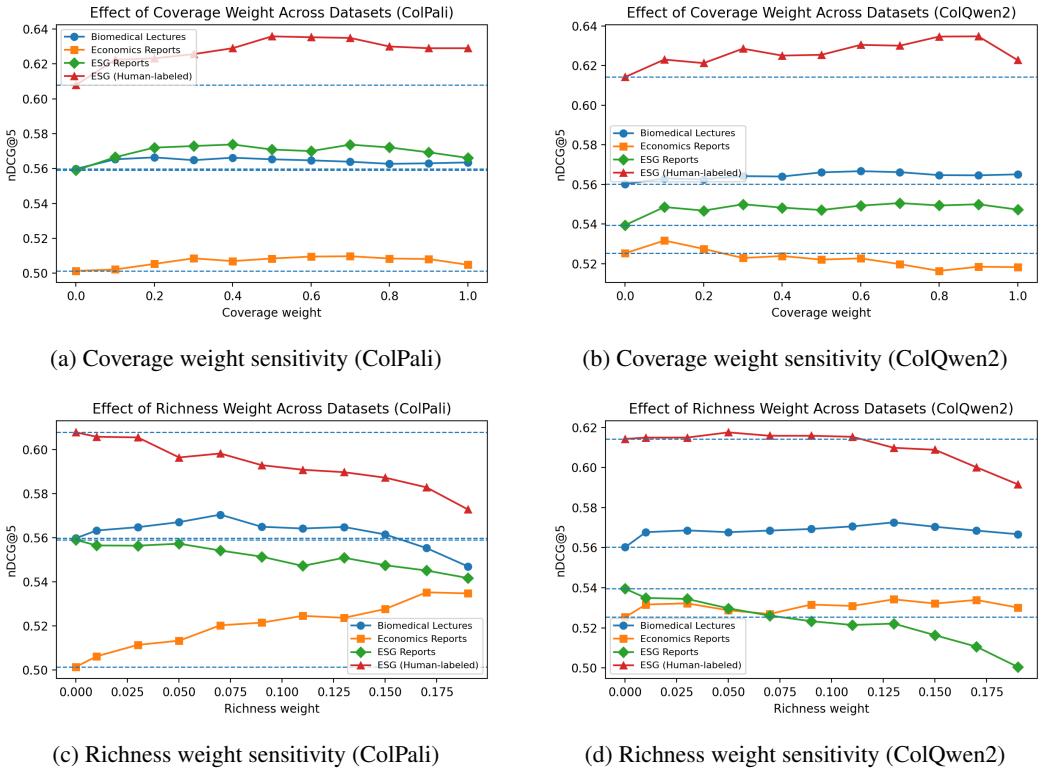
Effect of Coverage Weight. Figures 3a and 3b show the effect of the coverage weight α under a fixed coverage threshold (τ=0.5) for ColPali and ColQwen2, respectively. Across both backbones and all ViDoRe v2 subsets, coverage exhibits a high degree of robustness: performance remains consistently above the MaxSim baseline over a wide range of weights. In particular, moderate weights ($\alpha \in [0.2, 0.8]$) yield stable improvements without noticeable degradation, indicating that coverage acts as a reliable calibration signal rather than a dominant ranking factor.

The effect is especially pronounced on ESG Reports (Human), where increasing coverage weight consistently improves nDCG@5. This suggests that enforcing query-level evidence sufficiency effectively suppresses peaked similarity arising from sparse but visually salient regions. A single shared configuration across all subsets likewise improves over MaxSim (Appendix B.2.1).

Effect of Richness Weight. In contrast, evidence richness demonstrates a markedly different behavior. As shown in Figures 3c and 3d, the optimal richness weight varies substantially across datasets and backbones. While moderate richness weights improve performance on certain subsets (e.g., Biomedical Lectures and Economics Reports), overly large weights lead to rapid degradation—particularly on ESG-style datasets.

This dataset-dependent behavior reflects the nature of richness as a page-level structural signal: different document genres exhibit distinct density patterns, so richness benefits from dataset-aware weighting. Accordingly, we treat richness as an auxiliary calibration signal that complements coverage.

Summary. Taken together, the hyperparameter analysis reveals a clear distinction between the two components of EA-MaxSim. Coverage provides a stable, backbone-agnostic signal that generalizes well across datasets, whereas richness offers complementary benefits that are best realized with dataset-aware calibration. This complementary behavior motivates our design choice to combine coverage and richness while assigning them asymmetric roles in the final scoring function.



(a) Coverage weight sensitivity (ColPali)

(b) Coverage weight sensitivity (ColQwen2)

(c) Richness weight sensitivity (ColPali)

(d) Richness weight sensitivity (ColQwen2)

Figure 3: **Hyperparameter sensitivity analysis on ViDoRe v2.** We report nDCG@5 under different coverage and richness weights for ColPali and ColQwen2. Dashed horizontal lines indicate MaxSim baselines for each dataset.

B.2.3 Coverage Thresholds and Alternative Forms

All other components and weights are fixed at the main-result values; only the coverage form and its threshold vary. The absolute threshold is shared across the four ViDoRe subsets. Percentile coverage sets a query-specific threshold from the distribution of token-wise maximum similarities over candidate pages. Soft coverage replaces the indicator by $\sigma((s_i - \tau)/T)$, with $T = 0.1$.

Figures 4 and 5 show all tested settings for both backbones and all three nDCG cutoffs. The hard and soft variants use absolute thresholds, whereas the percentile variant is parameterized by a percentile. On ViDoRe v2, all three formulations match or exceed MaxSim at every cutoff for $\tau \in [0.2, 0.5]$ and for all tested percentiles; soft coverage remains above MaxSim over the entire tested range. On ViDoSeek, where the overall gains are smaller, hard coverage performs best at $\tau = 0.4$ for ColQwen2 and $\tau = 0.5$ for ColPali, while soft and percentile coverage vary more smoothly with their parameters.

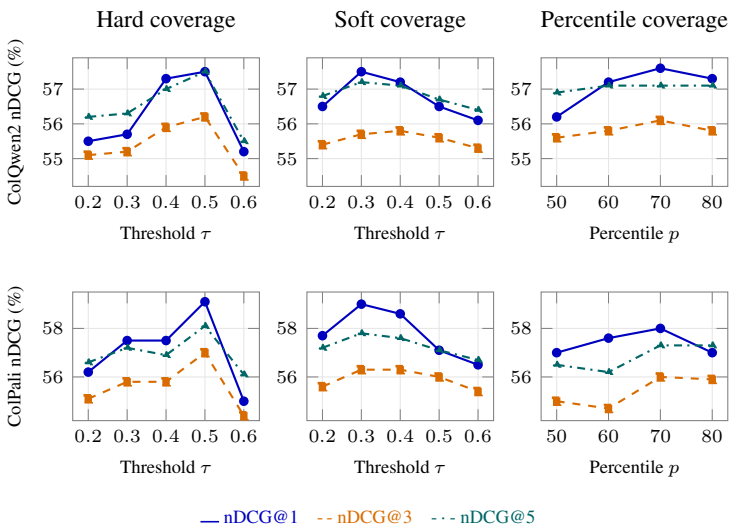


Figure 4: Coverage threshold and formulation sweeps on ViDoRe v2. Rows show the two backbones; columns show hard, soft, and percentile coverage. The vertical scale is shared within each row.

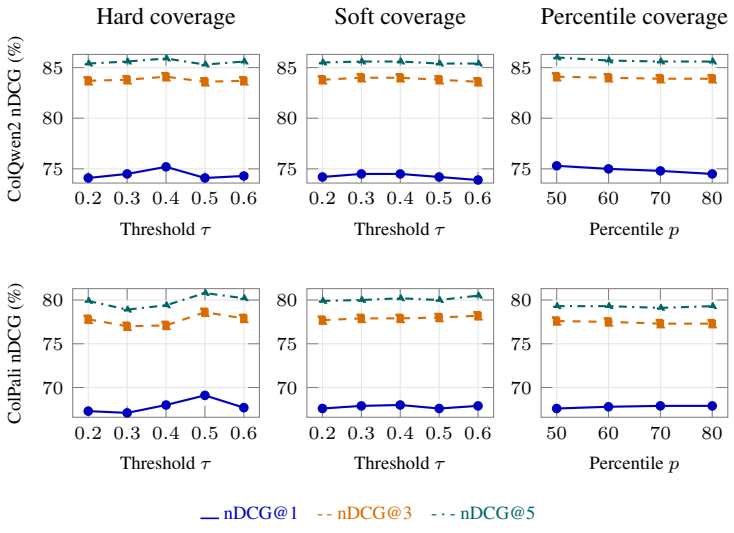


Figure 5: Coverage threshold and formulation sweeps on ViDoSeek. Rows show the two backbones; columns show hard, soft, and percentile coverage. The vertical scale is shared within each row.

B.2.4 Reranking Depth

We vary the reranking depth K, i.e., the size of the MaxSim candidate pool over which coverage and richness are computed and normalized (Appendix A.2). Figure 6 reports nDCG@1, @3, and @5 for $K \in \{20, 50, 100, 200, 300, 400, 500\}$, with MaxSim as a reference. EA-MaxSim improves over MaxSim at every depth and every cutoff, and a small pool already recovers most of the gain: at $K = 100$, nDCG@1 reaches 58.3 for ColQwen2 and 57.6 for ColPali, compared with 54.0 and 53.1 for MaxSim. Performance generally increases with K and saturates at $K = 500$, which yields exactly the same results as reranking all candidates. We therefore use $K = 500$ in all experiments.

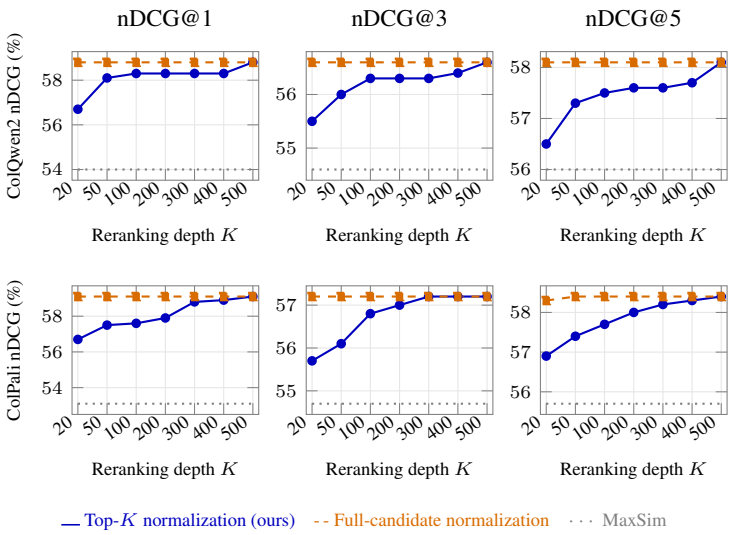


Figure 6: Reranking-depth sweep on ViDoRe v2, separated by backbone and nDCG cutoff. Top-K normalization computes the normalization statistics within the reranked pool; full-candidate normalization computes them over all candidates. The two coincide at $K = 500$. The horizontal axis lists the tested depths (not to scale).

B.3 Pooling and Token Merging

Softmax pooling uses the weighted patch average $\sum_j \mathrm{softmax}(s_j/T)\, s_j$ for each query token; top-k pooling averages the highest k similarities. Page embeddings are L2-normalized, so s_j are cosine similarities. Both operators reduce to MaxSim as $T \to 0$ or $k = 1$. Hierarchical merging uses compression factor four. All comparisons use the same MaxSim candidate set with query-token pruning disabled and auxiliary weights set to zero. Table 8 gives the complete temperature and pooling-depth sweeps for both backbones.

On both backbones, softmax pooling peaks at $T = 0.02$, where it exceeds MaxSim in nDCG@1 (+1.6 for ColQwen2, +1.7 for ColPali) and changes nDCG@3/@5 by at most 0.9 points. As T decreases further, performance approaches MaxSim; as T increases to 0.1 and 0.5, the operator approaches a near-uniform average over the page's patches and performance drops sharply. Top-k averaging stays below MaxSim for all tested $k \geq 2$, and hierarchical merging is below MaxSim at every cutoff. EA-MaxSim remains above the best setting of every alternative at all three cutoffs.

Table 8: Pooling baselines on ViDoRe v2 (nDCG, %).

	ColQwen2			ColPali		
Method	@1	@3	@5	@1	@3	@5
MaxSim	54.0	54.6	56.0	53.1	54.7	55.7
Softmax pooling, T=0.005	54.9	54.6	56.1	53.5	54.4	55.7
Softmax pooling, T=0.01	55.4	54.7	56.2	54.2	54.7	55.9
Softmax pooling, T=0.02	55.6	54.8	56.2	54.8	55.6	56.1
Softmax pooling, T=0.05	53.3	54.5	55.5	54.0	53.4	55.0
Softmax pooling, T=0.1	47.1	47.6	48.4	42.2	46.2	47.7
Softmax pooling, T=0.5	22.9	25.1	26.8	20.7	23.8	25.8
Top-k average, k=2	53.9	54.3	55.6	52.9	53.8	55.2
Top-k average, k=3	52.5	53.4	54.3	51.1	53.4	54.7
Top-k average, k=5	50.4	52.0	52.6	49.8	52.2	53.3
Top-k average, k=10	47.1	48.2	48.7	47.1	50.1	51.6
Light-ColPali hierarchical clustering, f=4	49.0	51.3	52.3	47.0	49.0	51.7
EA-MaxSim	58.8	56.6	58.1	59.1	57.2	58.4

B.4 Comparison with VLM Rerankers

All VLM experiments use ColQwen2 v1.0 on ViDoRe v2, with nDCG macro-averaged over the four subsets, and all methods rerank the same MaxSim candidate list. We consider two VLM rerankers: Qwen2-VL-2B-Instruct, a generic VLM without reranking-specific fine-tuning, and MonoQwen2-VL-v0.1, a dedicated reranker LoRA-adapted using ColPali training data and hard negatives. Both score query–image pairs by normalized $P(\mathrm{True})$, obtained by applying a two-way softmax to the output logits for `True` and `False` and taking the probability assigned to `True`. Cascades first apply EA-MaxSim and then rerank its top m candidates with MonoQwen2-VL.

VLM timing uses one H100-80GB, bf16, SDPA, batch size eight, and a maximum image size of $512 \cdot 28 \cdot 28$ pixels, with no cache. The measurements include JPEG decoding, preprocessing, and model forward computation, averaged over 20 queries balanced across subsets. Reported peak GPU memory is approximately 61.5 GB. EA's cold, single-thread CPU timing measures coverage and richness, including connected-component clustering, over the top-500 candidate pool: 264/235/266/281 ms per query across the four subsets, summarized as approximately 260 ms/query. These are additional online costs beyond the shared MaxSim retrieval stage.

Table 9 reports retrieval quality and added online latency for the generic VLM, the dedicated reranker (depths 5 to 100), and the cascades. Cascade latencies include both the EA and VLM stages. Three observations emerge. First, reranking alone does not guarantee improvement: the generic Qwen2-VL-2B underperforms MaxSim at every tested depth and degrades further as the depth increases. Second, the dedicated MonoQwen2-VL reaches nDCG@1 of 57.5–58.1, below the 58.8 of EA-MaxSim, and achieves higher nDCG@5 from depth ten onward, at the cost of an additional GPU and seconds of latency per query. Third, the two approaches are complementary: the cascade improves over direct MonoQwen2-VL reranking at the same depth for depths five and ten, attains the highest nDCG@1 overall (59.6) with only five VLM calls, and with ten candidates (58.2/61.2) outperforms MonoQwen2-VL on 100 candidates (57.6/61.1) in both nDCG@1 and nDCG@5 at about one tenth of the latency.

Table 9: Full VLM quality and timing record. Cascade latencies include both the EA and VLM stages.

Reranker	nDCG@1	nDCG@5	extra hardware	added latency/query
MaxSim	54.0	56.0	none	0
EA-MaxSim, top-500	58.8	58.1	none (CPU)	260 ms
Generic Qwen2-VL-2B, top-5	53.4	55.6	1×H100-80GB	1.74 s
Generic Qwen2-VL-2B, top-10	49.0	54.4	1×H100-80GB	4.72 s
Generic Qwen2-VL-2B, top-20	47.2	51.9	1×H100-80GB	9.82 s
Generic Qwen2-VL-2B, top-50	41.9	46.5	1×H100-80GB	26.38 s
MonoQwen2-VL, top-5	58.1	57.7	1×H100-80GB	1.74 s
MonoQwen2-VL, top-10	57.5	60.2	1×H100-80GB	4.72 s
MonoQwen2-VL, top-20	57.8	61.9	1×H100-80GB	9.82 s
MonoQwen2-VL, top-50	58.0	61.4	1×H100-80GB	26.38 s
MonoQwen2-VL, top-100	57.6	61.1	1×H100-80GB	52.69 s
EA-MaxSim → MonoQwen2-VL top-5	59.6	59.0	1×H100-80GB	2.00 s
EA-MaxSim → MonoQwen2-VL top-10	58.2	61.2	1×H100-80GB	4.98 s
EA-MaxSim → MonoQwen2-VL top-20	57.7	61.4	1×H100-80GB	10.08 s

B.5 Efficiency Analysis

Table 10: End-to-end online retrieval time on ViDoRe v2 (seconds for all queries of each subset). Page embeddings are pre-computed offline, so the GPU is used only for query encoding (a single GPU with 8-way data parallelism); the remaining computation runs on CPU. MaxSim is the original MaxSim inference of colpali-engine; EA-MaxSim includes MaxSim scoring over the full collection and evidence-aware reranking of the top-500 candidates.

Dataset	**#Queries**	**#Pages**	**MaxSim (s)**	**EA-MaxSim (s)**
Biomedical Lectures	640	1016	11.0	24.2
Economics Reports	232	452	1.7	7.8
ESG Reports	228	1538	5.7	8.8
ESG Reports (Human)	52	1538	1.4	2.2

Table 10 reports the end-to-end online retrieval time of MaxSim and EA-MaxSim on ViDoRe v2. All page-level patch embeddings are pre-computed and stored offline, so the reported time covers only online scoring and reranking.

EA-MaxSim keeps the end-to-end cost within seconds per subset. Even on the largest query set, Biomedical Lectures, answering all 640 queries takes 24.2 s (about 38 ms per query), compared with 11.0 s for MaxSim. Because coverage and richness are computed only for the top-500 candidates, the reranking cost is largely independent of the collection size. For example, Economics Reports and ESG Reports have similar numbers of queries (232 vs. 228) but collections that differ by more than a factor of three (452 vs. 1,538 pages), yet their EA-MaxSim times are similar (7.8 vs. 8.8 s).

Overall, EA-MaxSim improves retrieval accuracy at a small online cost, making it a practical drop-in reranking module for late-interaction retrieval systems.

B.6 Statistical Significance

Since our reranking procedure is deterministic given fixed cached embeddings, repeated runs do not provide meaningful variance estimates. We therefore assess statistical significance with paired randomization tests over queries, applied to the per-query differences between EA-MaxSim and the corresponding MaxSim baseline. On ViDoRe v2 (Table 11), the nDCG and MRR gains are significant at all cutoffs for both backbones ($p < 0.01$), and four of the six Recall gains are also significant. On ViDoSeek (Table 12, with query-aware token selection), ColPali improves significantly on all listed metrics ($p < 0.001$), and ColQwen2 improves significantly at rank one ($p < 0.05$).

Table 11: Statistical significance on ViDoRe v2 overall results. Each cell reports MaxSim → EA-MaxSim, with the absolute improvement in parentheses.

Backbone	Metric	@1	@3	@5
ColQwen2	nDCG	$54.01 \rightarrow 58.77^{\ddagger}$ (+4.77)	$54.62 \rightarrow 56.57^{\ddagger}$ (+1.96)	$55.98 \rightarrow 58.09^{\ddagger}$ (+2.11)
	Recall	$25.89 \rightarrow 28.89^{\ddagger}$ (+3.01)	$43.09 \rightarrow 44.44^{\dagger}$ (+1.34)	$52.63 \rightarrow 53.91$ (+1.28)
	MRR	$54.01 \rightarrow 58.77^{\ddagger}$ (+4.77)	$63.68 \rightarrow 66.60^{\ddagger}$ (+2.92)	$65.37 \rightarrow 68.46^{\ddagger}$ (+3.10)
ColPali	nDCG	$53.13 \rightarrow 59.10^{\ddagger}$ (+5.97)	$54.68 \rightarrow 57.22^{\ddagger}$ (+2.54)	$55.70 \rightarrow 58.35^{\ddagger}$ (+2.65)
	Recall	$25.18 \rightarrow 28.64^{\ddagger}$ (+3.46)	$44.63 \rightarrow 45.00$ (+0.37)	$52.57 \rightarrow 53.54^{\dagger}$ (+0.97)
	MRR	$53.13 \rightarrow 59.10^{\ddagger}$ (+5.97)	$64.44 \rightarrow 67.45^{\ddagger}$ (+3.01)	$65.66 \rightarrow 68.80^{\ddagger}$ (+3.14)

† and ‡ indicate statistically significant improvements over the corresponding MaxSim baseline under a paired randomization test over queries with $p < 0.05$ and $p < 0.01$, respectively.

Table 12: Statistical significance on ViDoSeek with query-aware token selection. The p column reports exact p-values; † and ‡ denote $p < 0.05$ and $p < 0.01$, as in Table 11.

Backbone	Metric	MaxSim	EA-MaxSim	p
ColQwen2	nDCG@1	74.34	75.83	0.034 †
ColQwen2	nDCG@5	85.51	86.00	0.161
ColQwen2	Recall@1	74.34	75.83	0.034 †
ColQwen2	Recall@5	94.48	94.22	0.712
ColPali	nDCG@1	67.34	69.88	<0.001 ‡
ColPali	nDCG@5	79.97	82.05	<0.001 ‡
ColPali	Recall@1	67.34	69.88	<0.001 ‡
ColPali	Recall@5	90.28	92.21	<0.001 ‡

B.7 Fusion Strategy Analysis

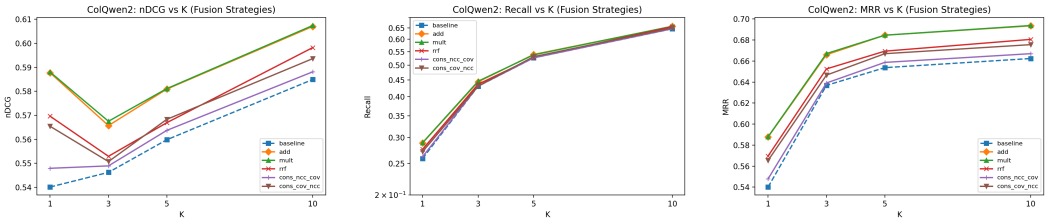
We analyze the impact of different fusion strategies for incorporating evidence coverage and evidence richness into MaxSim scoring. Specifically, we compare additive fusion (*add*), multiplicative fusion (*mult*), reciprocal rank fusion (*rrf*), and two constrained tie-breaking variants (*cons_ncc_cov* and *cons_cov_ncc*). All results are reported on ViDoRe v2 using macro-averaged metrics, with trends visualized in Figure 7.

Additive and multiplicative fusion. Both additive and multiplicative fusion consistently outperform rank-based strategies across metrics and backbones. Among them, additive fusion yields the most stable and robust improvements, achieving the best or near-best performance at almost all cutoffs for both ColQwen2 and ColPali. This behavior suggests that evidence coverage and richness serve as complementary signals that are best incorporated as calibrated adjustments to the original MaxSim score, rather than as dominant multiplicative factors. In contrast, multiplicative fusion exhibits slightly higher variance across K, indicating increased sensitivity to score magnitude and scaling.

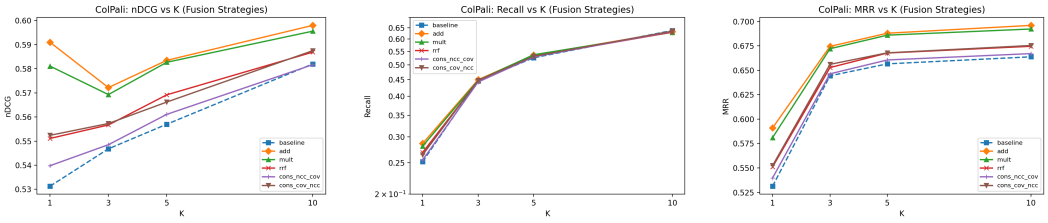
Rank-based fusion. Reciprocal rank fusion performs consistently worse than score-based fusion methods. This can be attributed to the loss of fine-grained score information: by operating solely on rank positions, RRF ignores the magnitude differences between candidates and cannot adequately reflect subtle evidence sufficiency signals. As a result, RRF is less effective at correcting peaked similarity cases where MaxSim assigns disproportionately high scores to evidence-poor pages.

Constrained tie-breaking strategies. The constrained ranking variants, which resolve near-ties in MaxSim scores by prioritizing coverage or richness, achieve modest improvements over the MaxSim baseline but remain inferior to score-based fusion. These methods only influence ordering within narrow score bands and thus have limited capacity to correct systematic ranking errors. Moreover, their asymmetric treatment of coverage and richness leads to less consistent behavior across datasets and cutoffs.

Summary. Overall, these results indicate that explicitly adjusting MaxSim scores using additive fusion provides the most effective and stable mechanism for modeling evidence sufficiency. Consequently, we adopt additive fusion as the default instantiation of EA-MaxSim in all main experiments.



(a) ColQwen2: nDCG@K (b) ColQwen2: Recall@K (c) ColQwen2: MRR@K



(d) ColPali: nDCG@K (e) ColPali: Recall@K (f) ColPali: MRR@K

Figure 7: **Comparison of fusion strategies on ViDoRe v2 (macro-average).** We plot nDCG@K, Recall@K, and MRR@K as a function of $K \in \{1, 3, 5, 10\}$ for different fusion strategies under ColQwen2 (top) and ColPali (bottom) backbones.

B.8 Effect of Query-aware Token Selection

Evidence coverage and evidence richness are both computed with respect to query tokens. However, not all query tokens contribute equally to retrieval intent. In real-world queries, especially long or compositional ones, function words and generic modifiers may dilute the signal from semantically salient terms.

To address this, we optionally introduce a query-aware token selection step that identifies a subset of *important* query tokens and restricts evidence computation to this subset. Conceptually, this step highlights key query concepts (illustrated as emphasized words in Figure 1), while treating other tokens as auxiliary context.

Formally, given the original query token set $\mathbf{Q}$, we extract a subset $\mathbf{Q}^\star \subseteq \mathbf{Q}$, and compute both evidence coverage and evidence richness using only tokens in $\mathbf{Q}^\star$. In our implementation, $\mathbf{Q}^\star$ consists of the query tokens that belong to keywords extracted from the query by an LLM. The underlying MaxSim score remains unchanged and is always computed over the full query.

This design allows EA-MaxSim to focus its evidence sufficiency signals on the most informative parts of the query, without altering the base similarity computation. Importantly, query-aware token selection is not required for EA-MaxSim to function and is treated as an optional enhancement.

We study the effect of computing evidence coverage and richness over different subsets of query tokens. Specifically, we compare using all query tokens versus using a selected subset of query-aware tokens when estimating coverage and richness, while keeping the original MaxSim scores unchanged.

Results on ViDoRe v2. As shown in Table 13, using all query tokens improves the reported ViDoRe v2 nDCG values on both backbones, while Recall and MRR show some ties or gains from token selection. The performance gap is most pronounced at ColQwen2, with up to 2.6 nDCG@1 degradation when restricting to selected tokens. This suggests that for ViDoRe v2, where queries are typically short and focused, most query tokens already carry meaningful retrieval signals. Removing tokens in this setting may inadvertently discard complementary evidence cues, leading to reduced coverage estimates and weaker ranking calibration.

Results on ViDoSeek. In contrast, query-aware token selection improves the reported ViDoSeek nDCG and MRR values on both backbones, although ColQwen2 Recall@5 decreases slightly and Recall@10 is unchanged. For ColQwen2, nDCG@1 improves from 75.2 to 75.8, with similar gains observed at larger cutoffs. A similar trend holds for ColPali. These gains indicate that

Table 13: **Effect of query-aware token selection.** We compare computing evidence coverage and richness over all query tokens (All) versus selected query tokens (Sel). All metrics are reported in percentage (%). Each cell is shown as All/Sel.

Dataset	Backbone	Metric	K=1	K=3	K=5	K=10
ViDoRe v2	ColQwen2	nDCG	**58.8**/56.2	**56.6**/56.3	**58.1**/57.8	**60.7**/60.3
		Recall	**28.9**/28.4	44.4/44.4	**53.9**/53.8	65.7/65.7
		MRR	**58.8**/56.2	**66.6**/65.4	**68.5**/67.2	**69.4**/68.1
ViDoRe v2	ColPali	nDCG	**59.1**/58.8	**57.2**/56.8	**58.4**/57.9	**59.8**/59.7
		Recall	**28.6**/28.3	**45.0**/44.7	**53.5**/53.3	63.3/63.3
		MRR	**59.1**/58.8	67.5/**68.5**	68.8/**69.5**	69.6/**70.1**
ViDoSeek	ColQwen2	nDCG	75.2/**75.8**	84.1/**84.5**	85.9/**86.0**	86.9/**87.1**
		Recall	75.2/**75.8**	90.1/**90.6**	**94.4**/94.2	97.6/97.6
		MRR	75.2/**75.8**	82.0/**82.4**	83.0/**83.2**	83.4/**83.7**
ViDoSeek	ColPali	nDCG	69.1/**69.9**	78.6/**79.8**	80.8/**82.1**	82.5/**83.6**
		Recall	69.1/**69.9**	85.6/**86.9**	90.8/**92.2**	96.2/**96.9**
		MRR	69.1/**69.9**	76.3/**77.4**	77.4/**78.6**	78.2/**79.3**

ViDoSeek queries, which are generally longer and more complex, contain a substantial fraction of low-informative or auxiliary tokens. Restricting coverage and richness computation to a selected subset allows the model to better focus on salient query intents, resulting in more accurate evidence sufficiency estimation.

Summary. Taken together, these results show that the benefit of query-aware token selection depends on query characteristics. For concise queries, computing coverage and richness over all tokens provides more stable and complete evidence estimation. For long and compositional queries, selective token usage can mitigate noise and improve ranking quality. Accordingly, we treat query-aware token selection as an optional instantiation of EA-MaxSim rather than a mandatory component, and report results under both settings for completeness.

C Case Study

Figure 8 presents two representative examples illustrating how EA-MaxSim corrects systematic ranking errors produced by vanilla MaxSim by explicitly modeling evidence coverage and evidence richness.

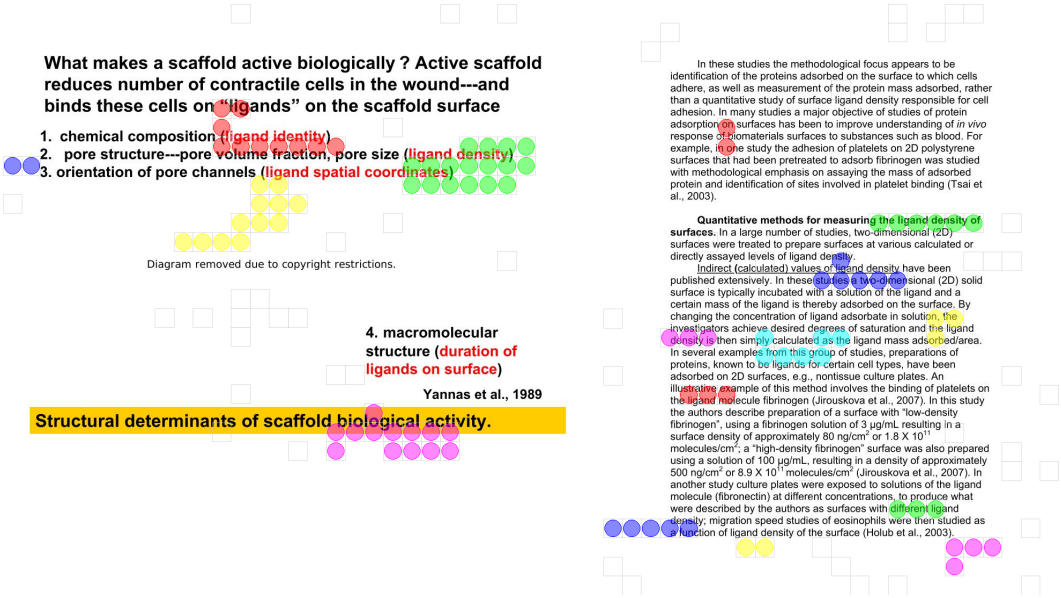
Coverage-driven correction (Figure 8a). The query "MYT group turnover" requires pages that jointly address *MYT group* and *turnover*. The left page, selected as top-1 by vanilla MaxSim, achieves a slightly higher raw MaxSim score due to strong visual matches to prominent brand elements (e.g., large logos and headers). However, it fails to provide evidence for the key concept "turnover," resulting in incomplete query coverage. In contrast, the page selected by EA-MaxSim explicitly covers all query terms, yielding a higher coverage score despite a marginally lower MaxSim score. By calibrating MaxSim with query coverage, EA-MaxSim correctly prioritizes the semantically complete page.

Richness-driven correction (Figure 8b). The second example demonstrates the effect of evidence richness modeling. Both pages contain patches that match query keywords related to ligand density, leading to comparable MaxSim scores. However, the left page primarily matches isolated keywords without providing concrete tabular or comparative evidence needed to answer the query. As shown by the clustered support patches, many matches on the left correspond to the same visual regions, yielding a small number of effective evidence units. In contrast, the page selected by EA-MaxSim contains multiple spatially distinct regions with quantitative ligand density information, resulting in a larger number of evidence clusters. By down-weighting spatially redundant matches and rewarding independent evidence units, EA-MaxSim correctly ranks the information-rich page higher.

Overall, these examples highlight that raw similarity aggregation alone is insufficient for visualized document retrieval. Explicitly modeling evidence sufficiency—through both query coverage and



(a) Query: MYT group turnover. Left: `maxsim` = 12.562, `coverage` = 0.889. Right: `maxsim` = 12.500, `coverage` = 1.0.



(b) Query: In what ways do the ligand density values for different biomaterials compare in the tables? Left: `maxsim` = 19.252, `cluster_count` = 5, `norm_cluster_count` = 0.238. Right: `maxsim` = 19.000, `cluster_count` = 11, `norm_cluster_count` = 0.571.

Figure 8: Case studies illustrating the limitations of vanilla MaxSim and the effectiveness of EA-MaxSim. Red heatmaps indicate patch-level similarity to the query. Small squares denote supporting patches whose similarity falls within the top 10%, while colored circles represent patches that actively contribute to evidence richness after filtering isolated and blank noise. Patches that are spatially adjacent and share the same color belong to the same spatial cluster (evidence unit). Although vanilla MaxSim assigns higher raw similarity scores to the incorrect pages in both cases, EA-MaxSim correctly ranks the relevant pages by calibrating MaxSim with query coverage (a) and effective evidence richness (b).

effective information richness—enables EA-MaxSim to distinguish superficial matches from pages that truly satisfy complex information needs.