REGIONMAX-WA: SUPERVISING MAXSIM WINNERS FOR VISUAL DOCUMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual document retrieval preserves the layout and visual context needed to search complex pages, but its usual supervision identifies only *which page* is relevant, not *which evidence* should support the match. This gap matters for late-interaction retrievers, whose page score sums a separate patch decision for each query token. We introduce **RegionMax-WA**, which supervises evidence membership in the same full-page token-wise competition used at retrieval time. OCR text and boxes propose query-dependent candidate evidence; Winner Allocation encourages each content token's MaxSim winner to fall within it; and the auxiliary loss supplements standard page-ranking training. Across four ViDoRe V2 collections, RegionMax-WA improves macro nDCG@5 (0–100 scale) by 2.21 and 4.21 absolute points over matched page-only PaliGemma-3B and Qwen2.5-VL-3B baselines. It also exceeds a full-complement aggregate-loss control by 0.49 and 0.95 points. On 250 independently annotated query–page pairs, MaxSim winners show greater spatial agreement with human evidence boxes. OCR is used only during training; inference retains the same index format and MaxSim scorer.

1 INTRODUCTION

Scientific, financial, and administrative collections contain pages whose meaning depends on the arrangement of text, tables, and figures. Retrieving the right page is often the first step toward answering a question or tracing a claim back to its source. In report analysis, scientific review, and evidence-based question answering, the retrieved page is a gateway to a particular statement, cell, or figure, rather than the final answer itself. Image-based document retrievers preserve this visual context and can index pages once for many future queries (Faysse et al., 2025; Macé et al., 2025; Dong et al., 2025). Their success makes a more precise question important: *what evidence within a retrieved page actually drives the match?*

The usual training signal identifies a relevant page, not the evidence that makes it relevant. This gap is particularly consequential in late-interaction retrieval, where different query terms independently select their strongest matching page patches. Dense reports often repeat entities, labels, and numerical values, giving a query several plausible but unequal local matches. A page may be ranked correctly even when some terms attach to a repeated number, a neighboring table cell, or unrelated text. Thus, good page ranking leaves token-wise winner locations unconstrained (Figure 1a). Restricting a training comparison to selected regions still leaves the other page patches available to win at inference (Figure 1b).

Prior work has improved page representations and introduced finer-grained supervision through teacher attention, OCR-aware pretraining, and region-grounded descriptions (Faysse et al., 2025; Cui et al., 2026; Masry et al., 2025; Yang et al., 2026). These approaches show the value of local information, but their targets differ from the individual maxima that form the deployed MaxSim score. An aggregate local score, for example, may improve while a poorly matched token is compensated by another. We therefore ask whether supervising evidence membership *before* token aggregation, against the full-page complement, improves retrieval (Figure 1c).

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.

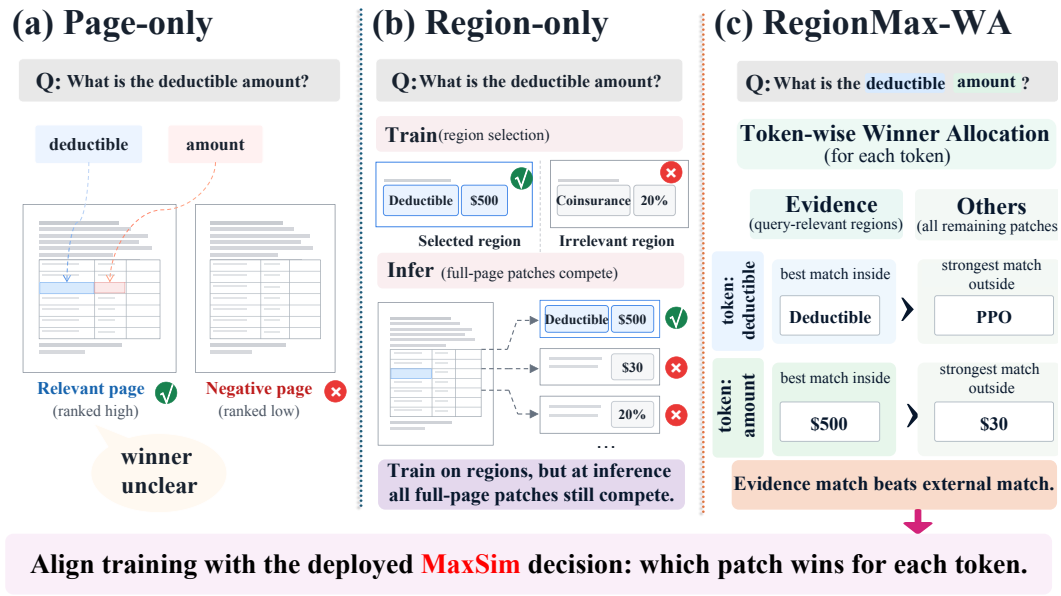

Figure 1: **From page ranking to token-wise winner supervision.** (a) A page-level label can rank the relevant page without specifying which patch wins for each query token. (b) A region-restricted training comparison can favor the selected region while unselected full-page patches still compete at inference. (c) RegionMax-WA compares each content token's best match in proposed evidence with its strongest match among all remaining patches.

We address this question with **RegionMax-WA**. OCR provides a text–location interface: transcriptions identify query-relevant candidate regions, and boxes map them to visual patches. For each content token, an auxiliary loss compares the best match inside the proposed evidence with the best match in the full-page complement. The standard page-ranking loss remains necessary because local evidence membership alone cannot rank relevant pages above irrelevant ones. Training thus links page relevance with token-wise winner location without adding an inference-time module.

Matched experiments on four ViDoRe V2 collections and two 3B backbones show gains over page-only training of 2.21 and 4.21 macro nDCG@5 points. A full-complement aggregate-loss control is lower by 0.49 and 0.95 points, while corrupting the selected evidence also reduces performance. On independently annotated pages, the model's winning matches align more often with human evidence boxes.

Our contributions are:

1. We formulate Winner Allocation as a token-wise evidence-versus-complement objective on the retriever's full-page patch grid, and show that its loss upper-bounds the fraction of MaxSim winners outside the proposed evidence.
2. We combine query-conditioned OCR evidence with this objective and the standard page-ranking loss, obtaining consistent retrieval gains across four collections and two backbones. Aggregate-loss and evidence-corruption controls probe the contribution of the local target.
3. We evaluate the location of deployed token winners against a separate 250-pair human-box set, complementing page-level retrieval metrics with a direct spatial diagnostic.

2 Related Work

Visual document retrieval and late interaction. Document retrieval must preserve signals that disappear when a page is reduced to plain text, including layout, typography, tables, plots, and spatial relations. Image-based retrievers therefore encode rendered pages directly and have become a

strong basis for document search (Faysse et al., 2025; Macé et al., 2025; Dong et al., 2025; Wasserman et al., 2025); description-based systems such as SERVAL offer a complementary semantic route (Nguyen et al., 2025). ColPali adapts multi-vector late interaction (Khattab & Zaharia, 2020; Santhanam et al., 2022) so that each query token can select its strongest visual patch while pages remain independently indexable (Faysse et al., 2025). This flexibility is well suited to dense documents, but it also separates the supervision signal from the scoring decision: training labels identify a relevant page, whereas the score is assembled from local token–patch maxima. Analyses of visual late interaction show that strong page retrieval can coexist with unstable or unintuitive patch matches (Qiao et al., 2025). Page representation and local decision alignment are therefore complementary problems.

Fine-grained supervision for visual retrieval. Recent work uses several routes to local supervision. AGREE aligns a query-averaged patch-similarity map with teacher attention, including full-map and salient-versus-complement variants (Cui et al., 2026). ColMate combines OCR-based spatial pretraining with a late-interaction objective that uses TopKSim during training and MaxSim at inference (Masry et al., 2025). ReAlign uses region-grounded descriptions to align query-induced and description-induced ranking distributions over candidate documents (Yang et al., 2026). RegionRAG contrasts relevant and other regions to support region-level retrieval and generation (Li et al., 2025). All make local information useful, but they supervise different quantities. Our target is the membership of *each* content token's hard full-page MaxSim winner in query-dependent candidate evidence; Appendix Table 7 compares the targets compactly.

OCR and structured evidence. OCR-aware document models combine recognized text, two-dimensional position, and visual appearance to organize page content (Xu et al., 2020; 2021; Huang et al., 2022; Appalaraju et al., 2021); OCR-free models instead learn to read or parse pixels directly (Kim et al., 2022; Lee et al., 2023). RegionMax-WA uses OCR in a narrower, training-only role. The transcription supports efficient query–evidence matching, and its box anchors that cue to the visual page. This pairing provides candidate supervision without replacing the image representation or requiring region labels in the training set. The local target is supplied by OCR, but the optimized decision is made by the original visual retriever.

3 METHOD

We study page-level visual document retrieval: given a text query q and a collection of page images, the model ranks pages by relevance. Training provides query–positive-page pairs, but no ground-truth evidence locations. A late-interaction retriever encodes the query as token vectors $\mathbf{q}_t$ ($t = 1, \ldots, N_q$) and each page d as visual patch vectors $\mathbf{d}_j$ ($j = 1, \ldots, N_d$). Its page score is

$$S_{\text{page}}(q, d) = \sum_{t=1}^{N_q} \max_{1 \le j \le N_d} \mathbf{q}_t^\top \mathbf{d}_j. \tag{1}$$

For token t, let $j_t^*(q, d) = \arg\max_j \mathbf{q}_t^\top \mathbf{d}_j$ denote its winning patch. MaxSim chooses this winner independently for each token, then sums the winning scores to rank pages. The same page score can therefore arise from different collections of local matches: a correct term may support the answer-bearing area while another term wins on a distractor elsewhere. Page-level training constrains the final ranking but leaves the locations of these winners unspecified. RegionMax-WA uses weak evidence to supervise those locations without changing Eq. 1 at inference.

Given a query and a relevant page, we first use OCR text to select candidate evidence regions and map their boxes to the retriever's patch vectors (Figure 2, top). The same full-page encoder then supplies the visual vectors used by MaxSim (bottom). For each content token, we compare its strongest match in the selected set with its strongest match among all other page patches. An auxiliary loss encourages the evidence match to win, while the page-ranking loss preserves the retrieval task. The evidence selector and local loss are used only during training.

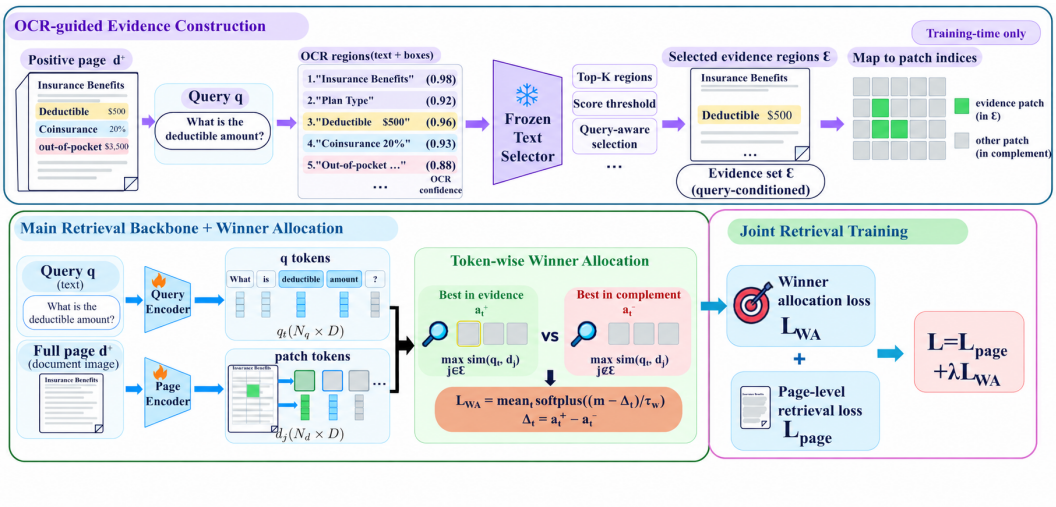

Figure 2: **RegionMax-WA training pipeline.** OCR text and boxes select query-conditioned candidate evidence and map it to full-page patch indices (top). Each content token compares its best match inside this set with its best match in the complement; Winner Allocation supplements page-ranking loss (bottom). The diagram depicts the margin condition; Eq. 4 specifies the smoothed, content-token-averaged loss used in training. OCR and the auxiliary loss are absent at inference.

3.1 OCR-Guided Evidence Construction

The visual encoder represents a page as a regular patch grid, which is effective for retrieval but does not mark semantic units such as table cells or field–value pairs. Page labels cannot fill this gap: they say which page is relevant, not where its supporting content lies. OCR supplies both pieces needed for a lightweight proposal. A transcription can be compared directly with the query, and its box locates that text on the page without asking the retriever to change its visual representation.

Concretely, for a positive page d^+, OCR returns regions (T_i, B_i, c_i) containing text, a box, and confidence. A frozen text encoder g scores each transcription against the query by $\cos(g(q), g(T_i))$. After confidence and similarity filtering, we retain up to K regions. Let P_j be the image footprint of page patch j and $\mathcal{I}_K(q, d^+)$ the selected region indices. Their boxes define the evidence set

$$E(q, d^+) = \bigcup_{i \in \mathcal{I}_K(q,d^+)} \left\{ j : \frac{|P_j \cap B_i|}{|P_j|} \geq \eta \right\}. \tag{2}$$

The union allows nearby text regions to jointly represent a field name and its value. In the experiments, at most three regions are retained, and their mapped patches cover about 7% of a page (Table 3).

The construction turns a page-level training pair into a structured, query-dependent evidence proposal. OCR determines both *which content* is plausible and *where it lies* in the original patch space; Winner Allocation can therefore supervise visual patches without changing the retriever's representation. The resulting set is the spatial target for the next stage, rather than a new retrieval unit. Appendix A gives the exact filtering and mapping rules.

3.2 Token-Wise Winner Allocation

A plausible evidence region is not enough if training compares it only with other selected regions: at retrieval time, an unselected patch anywhere on the page may still win. A second issue is aggregation. Even with page-wide competitors, a strong evidence match for one query term can hide a misplaced match for another when token scores are combined before supervision. Winner Allocation addresses both issues by asking, separately for each content term, whether its best evidence patch defeats its best competitor on the rest of the page.

To express this competition, let $\mathcal{T}_c(q)$ contain the query's content tokens after excluding control tokens, punctuation, and function-like terms. For $t \in \mathcal{T}_c(q)$, define

$$\Delta_t = \underbrace{\max_{j \in E} \mathbf{q}_t^\top \mathbf{d}_j}_{\text{best evidence match}} - \underbrace{\max_{j \notin E} \mathbf{q}_t^\top \mathbf{d}_j}_{\text{strongest competing match}} . \tag{3}$$

The complement contains every retained page patch outside E, including content not selected by OCR; it is not assumed to be irrelevant. We encourage a positive margin m with

$$\mathcal{L}_{\text{WA}} = \frac{1}{|\mathcal{T}_c(q)|} \sum_{t \in \mathcal{T}_c(q)} \text{softplus}\left(\frac{m - \Delta_t}{\tau_w} \right). \tag{4}$$

Proposition 1 (Winner-location surrogate). Suppose E and its complement are nonempty, $m \geq 0$, and $\tau_w > 0$. Without a tie across the two sets, $\Delta_t > 0$ if and only if token t's full-page MaxSim winner lies in E. Moreover, writing $\ell_t = \text{softplus}((m - \Delta_t)/\tau_w)$, we have

$$\mathbf{1}[\Delta_t \leq 0] \leq \frac{\ell_t}{\text{softplus}(m/\tau_w)}, \qquad \frac{1}{|\mathcal{T}_c|} \sum_{t \in \mathcal{T}_c} \mathbf{1}[\Delta_t \leq 0] \leq \frac{\mathcal{L}_{\text{WA}}}{\text{softplus}(m/\tau_w)}. \tag{5}$$

Indeed, if $\Delta_t \leq 0$, monotonicity of softplus gives $\ell_t \geq \text{softplus}(m/\tau_w)$; otherwise the indicator is zero. Thus Winner Allocation smoothly upper-bounds the rate of full-page winners outside the *proposed* evidence, including ties. A margin on $\sum_t \Delta_t$ does not yield this per-token guarantee: a large positive term can compensate for a negative one. The recorded A3 aggregate control also uses a sum-scale margin rather than a token-normalized margin, so its measured gap to A4 should not be attributed exclusively to token-wise reduction (Appendix B).

3.3 Joint Training and Inference

Localizing token matches is useful only if the model still ranks relevant pages above irrelevant ones. We therefore keep the original retrieval task as the primary objective and add Winner Allocation on positive pages where OCR yields usable evidence. This design also separates training from deployment: the evidence proposal improves the visual retriever without becoming a serving dependency.

Let $\mathcal{L}_{\text{page}}$ be full in-batch cross-entropy on the page scores in Eq. 1, and let $\overline{\mathcal{L}}_{\text{WA}}$ average Eq. 4 over positive pairs with valid evidence. The joint objective is

$$\mathcal{L} = \mathcal{L}_{\text{page}} + \lambda \, \overline{\mathcal{L}}_{\text{WA}}. \tag{6}$$

The two terms have complementary roles: the first learns *which page* to rank, while the second encourages content-token matches to land inside the proposed evidence set. A pair without usable evidence still contributes to page learning. At inference, pages are encoded into query-independent full-page vectors and scored by Eq. 1; OCR, evidence selection, and the auxiliary loss disappear. The index *format* and scoring operator stay the same, although a newly fine-tuned document encoder requires re-encoding the collection. Appendix A gives the minibatch reduction and Appendix E lists implementation settings.

4 Experiments

The experiments follow the method's logic. We first ask whether RegionMax-WA improves page retrieval, then compare evidence proposals and local objectives, and finally test whether the learned winners agree with independently annotated evidence. We also report uncertainty and serving cost to distinguish a training benefit from an architectural change at inference.

4.1 Dataset

Training uses 118,695 English query–positive-page pairs from the corpus used by AGREE (Cui et al., 2026). We evaluate on four out-of-domain ViDoRe V2 collections (Macé et al., 2025): ESG Reports Human (52 queries, 1,538 pages), ESG Reports (228, 1,538), Economics Reports (232, 452), and Biomedical Reports (640, 1,016). We report nDCG@5 on each collection and their unweighted macro average. No evaluation query enters training-time evidence selection or checkpoint tuning; the evidence audit on evaluation pages below is descriptive. Appendix E gives the collection and training-set details.

4.2 Implementation Details

We study PaliGemma-3B and Qwen2.5-VL-3B visual retrievers. For each backbone, all matched variants use the same training pairs, page loss, optimization schedule, and seeds $\{42, 123, 3407\}$; only the local supervision changes. OCR uses English PP-OCRv4, and a frozen `bge-base-en-v1.5` encoder selects up to three regions. The OCR-confidence, selector-similarity, and box-to-patch overlap thresholds are 0.5, 0.4, and 0.5, respectively. Winner Allocation uses $m = 0.1$, $\tau_w = 0.05$, and $\lambda = 0.2$. Training runs on four RTX 4090 GPUs; complete model identifiers, token filters, optimizer settings, and reductions are in Appendix E.

4.3 Comparison with Existing Methods

Table 1 groups systems by backbone family and reports all four collections. The comparator results are taken from AGREE Table 2 (Cui et al., 2026); A0 and A4 are our matched runs.

Table 1: **ViDoRe V2 retrieval (nDCG@5 on a 0–100 scale; higher is better).** Published comparison rows are reproduced from AGREE Table 2 (Cui et al., 2026) and are contextual rather than same-protocol runs; our matched A0/A4 rows are three-seed means. Per-collection deviations appear in Appendix Table 8.

Backbone family	Method	ESG-H	ESG	Econ.	Biomed.	Macro avg.
Other models	BiCLIP	22.74	11.38	22.83	22.39	19.84
	SigLIP (zero-shot)	23.07	28.19	16.34	24.30	22.98
	BiSigLIP	45.83	34.92	28.23	29.05	34.51
	DSE	61.74	55.15	52.80	55.29	56.25
PaliGemma-3B	ColPali	63.67	54.05	53.30	57.56	57.14
	AGREE-Pali	65.28	58.87	51.33	55.51	57.75
	A0 Page only (ours)	60.10	59.20	56.30	55.52	57.78
	A4 RegionMax-WA	**62.55**	**61.45**	**58.20**	**57.76**	**59.99**
Qwen2.5-VL-3B	ColQwen2.5	63.70	57.12	52.90	60.64	58.59
	AGREE-Qwen2.5	70.73	58.43	56.50	60.49	61.54
	A0 Page only (ours)	63.60	62.80	59.90	58.54	61.21
	A4 RegionMax-WA	**67.95**	**67.10**	**63.95**	**62.68**	**65.42**

A4 improves its matched page-only baseline by 2.21 macro nDCG@5 points on PaliGemma and 4.21 on Qwen2.5-VL. Every evaluated collection improves under both backbones, showing that the training signal transfers across distinct document domains.

On PaliGemma, the per-collection gains are 2.45, 2.25, 1.90, and 2.24 points on ESG-H, ESG, Economics, and Biomedical, respectively; the corresponding Qwen2.5-VL gains are 4.35, 4.30, 4.05, and 4.14 points. We next compare the local supervision choices behind these gains.

4.4 Ablation Study

Under a fixed within-backbone retrieval setup, we compare local supervision targets (Table 2), then test evidence corruption, sensitivity, and local grounding. A0 is trained with the same initialization and page objective as A4, but without an auxiliary local loss.

Table 2: **Matched objective ablation (macro nDCG@5).** Within each backbone, all rows share the training and evaluation protocol. Mean±standard deviation over three seeds; ΔA0 is relative to page-only training. Full collection results appear in Appendix Table 8.

Variant	Local supervision	Pali-3B	ΔA0	Qwen-3B	ΔA0
A0 Page only	None	57.78 ± 0.11	–	61.21 ± 0.14	–
A1 Region–Text	OCR text alignment	58.49 ± 0.10	+0.71	62.41 ± 0.13	+1.20
A2 RegionRank	Selected-region ranking	59.22 ± 0.09	+1.44	64.16 ± 0.15	+2.95
A3 Complement-Aggregate	Full complement; aggregate	59.50 ± 0.07	+1.72	64.47 ± 0.12	+3.26
A4 RegionMax-WA	**Full complement; per token**	$\mathbf{59.99 \pm 0.09}$	**+2.21**	$\mathbf{65.42 \pm 0.14}$	**+4.21**

Which competition should be supervised? A1 aligns OCR text and visual regions, A2 ranks query-conditioned OCR regions, and A3 compares selected evidence with the full-page complement after aggregating token scores. A4 compares the same evidence and complement per token. Macro nDCG@5 increases along this sequence on both backbones. A4 exceeds A3 by 0.49/0.95 points, with paired 95% bootstrap intervals [0.15, 0.80]/[0.50, 1.40] (Table 6). Because A3 applies its margin to a token *sum* while A4 applies it to individual token margins, the measured difference includes a change in loss scale; it is not a reduction-only ablation. The reported size-matched A4–A2 comparison retains 0.67/1.08 points of the original 0.77/1.26-point gains (Appendix C).

Does the evidence proposal matter? Area-matched random and query-shuffled masks lower macro nDCG@5 from 59.99 to 59.27/59.34 on PaliGemma and from 65.42 to 64.24/64.38 on Qwen2.5-VL (Table 4). These controls still outperform A0, so the data support an added benefit from query-corresponding evidence without ruling out a broader regularization effect. Multiple OCR regions also beat one, consistent with labels and values spanning neighboring boxes. Table 3 makes the proposal's scale explicit: across collections, the selected boxes map to only 6.7–8.0% of page patches, leaving a substantial complement for the winner comparison.

Table 3: **Scale and availability of OCR evidence.** Region counts are per-page means; patch coverage and valid-WA rates are percentages. Evaluation rows are a post-hoc proposal audit, not training or tuning data. Coverage measures proposal size rather than evidence accuracy.

Split	OCR regions	Selected	Patch coverage (%)	Valid WA (%)
Training	48	2.4	7.2	95.9
ESG-H	51	2.5	7.5	96.3
ESG	49	2.4	7.1	95.9
Economics	43	2.3	6.7	95.4
Biomedical	57	2.6	8.0	96.1

This audit clarifies the supervision available to the retriever. The selector does not force three boxes onto every page: only regions passing the text and confidence filters enter the proposal, giving 2.3–2.6 selected regions per evaluation page on average. Their small patch footprint leaves most of the page in the competing set of Eq. 3; the objective therefore tests whether proposed evidence beats alternatives across the page, rather than merely ranking selected regions against one another. Valid local comparisons are available for 95.4–96.3% of audited evaluation pairs. These figures describe proposal availability and size; the independent human-box test below addresses spatial agreement.

How sensitive are the two main settings? At fixed $K = 3$, $\lambda = 0.1$ and 0.4 trail the default 0.2 by 0.54/0.77 and 0.24/0.32 macro points on PaliGemma/Qwen2.5-VL (Table 4). At fixed $\lambda = 0.2$, $K = 5$ nearly matches $K = 3$ (59.92/65.30 versus 59.99/65.42), while $K = 1$ is lower. The default leads this sweep on both backbones, while the neighboring settings retain substantial gains over page-only training.

Table 4: **Evidence and sensitivity controls (macro nDCG@5).** Mean±standard deviation over seeds 42, 123, and 3407. Unless varied, $\lambda = 0.2$ and OCR Top-$K = 3$. Random masks match evidence area; shuffled masks break query correspondence.

Test	Variant	PaliGemma-3B	Qwen2.5-VL-3B
Evidence set	Area-matched random	59.27 ± 0.10	64.24 ± 0.14
	Query-shuffled	59.34 ± 0.10	64.38 ± 0.14
OCR Top-K	$K = 1$	59.78 ± 0.09	65.08 ± 0.14
	$K = 5$	59.92 ± 0.10	65.30 ± 0.15
Token set	All query tokens	59.80 ± 0.10	65.14 ± 0.14
Loss weight	$\lambda = 0.1$	59.45 ± 0.10	64.65 ± 0.15
	$\lambda = 0.4$	59.75 ± 0.11	65.10 ± 0.16
Complete method	$\boldsymbol{\lambda = 0.2,\ K = 3}$	$\mathbf{59.99 \pm 0.09}$	$\mathbf{65.42 \pm 0.14}$

Do local matches agree with human evidence? We evaluate token winners against independently annotated boxes on 250 query–positive-page pairs (Table 5).

Table 5: **Independent human-box grounding (higher is better).** All objectives use the same 250 annotated query–positive-page pairs per backbone. Winner Hit is the share of content-token full-page MaxSim winners inside human evidence; Human Margin is the best similarity inside minus the best outside. Values are point estimates.

Backbone	Objective	Winner Hit (%)	Human Margin
PaliGemma-3B	A0 Page only	47.51	0.021
	A2 RegionRank	51.26	0.029
	A3 Complement-Aggregate	52.37	0.036
	A4 RegionMax-WA	**54.48**	**0.052**
Qwen2.5-VL-3B	A0 Page only	49.13	0.024
	A2 RegionRank	52.89	0.033
	A3 Complement-Aggregate	54.28	0.042
	A4 RegionMax-WA	**57.34**	**0.061**

The pairs span all four collections (63/63/62/62). Two annotators label each pair independently and a third adjudicates disagreements; the same pairs are used for both backbones and all compared objectives. Appendix Figure 4 illustrates token winners on example pages.

Relative to A3, A4 raises Human Winner Hit by 2.11 and 3.06 percentage points and Human Margin by 0.016 and 0.019, respectively. A4 is highest on both spatial measures and backbones, and retrieval and winner-location agreement move in the same direction. The A3–A4 comparison uses the same evidence and competitor sets, but differs in token reduction and effective margin scale; the human-box results are descriptive point estimates, not a measure of token-level semantic correctness. Appendix H describes the annotation protocol.

Consistency and uncertainty. A4 improves over matched A0 on all four collections for both backbones; the complete A0–A4 trajectories are in Appendix Figure 3. Table 6 complements three-seed standard deviations with paired uncertainty over evaluation queries. Its intervals exclude zero for A4–A0, A4–A2, and A4–A3. These are distinct uncertainty summaries: query resampling does not replace analysis of variation across training runs. Appendix G describes the paired comparison.

Training and serving cost. The OCR and selection pipeline runs only during training and can be cached. Training takes approximately 16 hours for PaliGemma-3B and 24 hours for Qwen2.5-VL-3B on four RTX 4090 GPUs. At serving time, index size and peak inference memory match A0; measured query latency is 1.00–$1.01\times$ and throughput 0.99–$1.00\times$ the matched baseline (Appendix Table 12). The gains therefore require no additional inference stage.

Table 6: **Paired retrieval uncertainty.** Macro nDCG@5 differences with 95% paired-bootstrap intervals over evaluation queries. A4–A2 contrasts region ranking with full-page token supervision; A4–A3 contrasts aggregate and per-token losses, with different effective margin scales.

Comparison	Pali effect	Pali 95% CI	Qwen effect	Qwen 95% CI
A4−A0	+2.21	$[+1.6, +2.8]$	+4.21	$[+3.4, +5.0]$
A4−A2	+0.77	$[+0.4, +1.1]$	+1.26	$[+0.8, +1.7]$
A4−A3	+0.49	$[+0.15, +0.80]$	+0.95	$[+0.50, +1.40]$

Together, these results follow the method's computation: query-conditioned OCR evidence identifies candidate patches, full-page comparison exposes the competitors seen by MaxSim, and a separate margin for each token improves both retrieval and agreement with independently annotated evidence.

5 Discussion

The central finding. The results support the value of connecting candidate evidence to the retrieval decision itself. RegionMax-WA compares each content token against the full-page complement before aggregating token scores. It improves retrieval over the tested page-only and local-objective controls, while independently annotated boxes show better spatial agreement of its winners. Corrupted-evidence controls retain part of the gain, so the improvement plausibly combines query-specific evidence with a more general training effect. The present A3 comparison also changes effective margin scale, which limits a reduction-only interpretation.

Practical significance. RegionMax-WA improves the training signal rather than expanding the serving pipeline. OCR supplies a structured evidence interface during training, but search uses the same index format and MaxSim computation. Fine-tuning changes document vectors, so a collection must be encoded again with the new model; once rebuilt, index size and query-time operations remain comparable to A0. This separation lets local supervision alter model behavior without imposing an OCR dependency at query time.

Consistency across settings. The two backbones differ in page-only retrieval levels, yet the same training-side intervention improves every evaluated collection. Paired retrieval intervals and human-box point estimates describe complementary outcomes: improved page ranking and more winners inside annotated evidence.

6 Limitations

The evidence constructor uses English OCR and a shared, query-dependent union of selected boxes. OCR can miss nontextual or implicit support, and a winner inside the union is not necessarily aligned with the particular box semantically associated with its token. Human-box overlap measures spatial agreement, not causal faithfulness. Evaluation covers two 3B backbones and four ViDoRe V2 collections; broader models, scripts, and domains remain future work.

7 Conclusion

Visual late-interaction retrieval is trained with page labels but scored through token-wise patch decisions. RegionMax-WA uses OCR to propose candidate evidence, encourages full-page MaxSim winners to fall inside it, and retains the page-ranking objective. Across the tested backbones and collections, retrieval improves over matched page-only training, and independent annotations show greater spatial agreement of token winners. The approach changes training supervision while preserving the serving index format and MaxSim operator.

REFERENCES

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 993–1003, 2021.

Wanqing Cui, Wei Huang, Yazhi Guo, Yibo Hu, Meiguang Jin, Junfeng Ma, and Keping Bi. Attention grounded enhancement for visual document retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 223–233, 2026. doi: 10.1145/3805712.3809532.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30971–31005, 2025. doi: 10.18653/v1/2025.emnlp-main.1576.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025.

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*, pp. 4083–4091, 2022.

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. doi: 10.1145/3397271.3401075.

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 *European Conference on Computer Vision*, pp. 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*, pp. 18893–18912, 2023.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. RegionRAG: Region-level retrieval-augmented generation for visual document understanding. *arXiv preprint arXiv:2510.27261*, 2025.

Quentin Macé, Antônio Loison, and Manuel Faysse. ViDoRe benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025.

Ahmed Masry, Megh Thakkar, Patrice Bechard, Sathwik Tejaswi Madhusudhan, Rabiul Awal, Shambhavi Mishra, Akshay Kalkunte Suresh, Srivatsava Daruru, Enamul Hoque, Spandana Gella, Torsten Scholak, and Sai Rajeswar. ColMate: Contrastive late interaction and masked text for multimodal document retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2071–2080, 2025. doi: 10.18653/v1/2025.emnlp-industry.145.

Thong Nguyen, Yibin Lei, Jia-Huei Ju, and Andrew Yates. SERVAL: Surprisingly effective zero-shot visual document retrieval powered by large vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30807–30822, 2025. doi: 10.18653/v1/2025.emnlp-main.1568.

Jingfen Qiao, Jia-Huei Ju, Xinyu Ma, Evangelos Kanoulas, and Andrew Yates. Reproducibility, replicability, and insights into visual document retrieval with late interaction. *arXiv preprint arXiv:2505.07730*, 2025.

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. doi: 10.18653/v1/2022.naacl-main.272.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. REAL-MM-RAG: A real-world multi-modal retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 31660–31683, 2025. doi: 10.18653/v1/2025.acl-long.1528.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

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*, pp. 2579–2591, 2021. doi: 10.18653/v1/2021.acl-long.201.

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 and Data Mining*, pp. 1192–1200, 2020.

Hao Yang, Yifan Ji, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Zulong Chen, Shuo Wang, Yu Gu, and Ge Yu. ReAlign: Optimizing the visual document retriever with reasoning-guided fine-grained alignment. *arXiv preprint arXiv:2604.07419*, 2026.

AI USE STATEMENT

Generative AI tools were used for English translation, language editing, LaTeX formatting, conceptual figure generation, visual inspection of the typeset manuscript, and non-decisive code and analysis assistance. Experimental measurements, data labels, model outputs, and statistical conclusions came from the authors' experiments; AI did not generate or replace them. The authors take responsibility for all scientific claims, equations, results, and citations.

A EVIDENCE CONSTRUCTION AND TRAINING DETAILS

OCR proposals and selection. For a positive page d^+, OCR returns regions $\mathcal{R}(d^+) = \{R_i = (B_i, T_i, c_i)\}_{i=1}^{M}$, where B_i is a box, T_i its transcription, and c_i the OCR confidence. A frozen text encoder g scores each region by $u_i = \cos(g(q), g(T_i))$. We first discard regions below the OCR-confidence threshold and then retain up to K highest-scoring regions that also pass the selector threshold:

$$\mathcal{R}_{\text{valid}} = \{i : c_i \geq \theta_{\text{ocr}}\}, \qquad \mathcal{I}^+(q, d^+) = \{i \in \text{TopK}_{k \in \mathcal{R}_{\text{valid}}}(u_k) : u_i \geq \theta_s\}. \tag{7}$$

The same frozen OCR proposals and selector outputs are cached for all matched local-objective variants.

Box-to-patch mapping. Let P_j be the image footprint of full-page vector $\mathbf{d}_j$. With $\text{IoA}(P_j, B_i) = |P_j \cap B_i|/|P_j|$, the evidence indices are

$$\mathcal{P}(R_i) = \{j : \text{IoA}(P_j, B_i) \geq \eta\}, \qquad E(q, d^+) = \bigcup_{i \in \mathcal{I}^+} \mathcal{P}(R_i). \tag{8}$$

The competitor indices are $C(q, d^+) = \{1, \ldots, N_d\} \setminus E(q, d^+)$. Both sets refer to the original page vectors. We omit the local loss if E, C, or the content-token set is empty.

Page objective and reduction. For a minibatch $\mathcal{B} = \{(q_i, d_i^+)\}_{i=1}^B$ of query–positive-page pairs, the page-level term is full in-batch cross-entropy:

$$\mathcal{L}_{\text{page}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{\text{page}}(q_i, d_i^+)/\tau_p)}{\sum_{j=1}^{B} \exp(S_{\text{page}}(q_i, d_j^+)/\tau_p)}. \tag{9}$$

Winner Allocation is averaged over content tokens within each valid positive pair, then over valid pairs in the minibatch. A pair without valid local evidence still contributes to $\mathcal{L}_{\text{page}}$.

B Diagnostic Local Objectives

The main method defines RegionMax-WA; we specify the diagnostic alternatives here to separate representation alignment, competitor coverage, and supervision granularity.

Region–Text. A1 aligns a visual region with its OCR transcription. It tests generic local semantic alignment, but its text target differs from the query–visual geometry used by the retriever.

RegionRank. For OCR region R_i, let $\mathcal{P}(R_i)$ be the overlapping full-page patches. RegionRank uses

$$S_{\text{reg}}(q, R_i) = \sum_{t \in \mathcal{T}_c(q)} \max_{j \in \mathcal{P}(R_i)} \mathbf{q}_t^\top \mathbf{d}_j. \tag{10}$$

It ranks selected OCR evidence above competing OCR regions. This objective uses query–visual similarity, but excludes patches outside the proposals and applies its margin only after aggregating over query tokens.

Complement-Aggregate. This control expands competition to the full-page complement while retaining aggregation:

$$S_E(q) = \sum_{t \in \mathcal{T}_c(q)} \max_{j \in E} \mathbf{q}_t^\top \mathbf{d}_j, \tag{11}$$

$$S_C(q) = \sum_{t \in \mathcal{T}_c(q)} \max_{j \in C} \mathbf{q}_t^\top \mathbf{d}_j. \tag{12}$$

Its recorded local margin requires $S_E(q) > S_C(q) + m$. A large positive margin from one query token can compensate for a negative margin from another. Moreover, this sum-scale margin is not numerically equivalent to the per-token margin in Eq. 4. RegionMax-WA applies the evidence-over-complement penalty before token aggregation. A normalized aggregate-margin control would be needed to isolate the reduction choice alone.

For comparison, a *hypothetical* normalized aggregate loss would be

$$\mathcal{L}_{\text{agg,mean}} = \text{softplus}\left(\frac{m - \frac{1}{|\mathcal{T}_c|}\sum_{t \in \mathcal{T}_c} \Delta_t}{\tau_w}\right) \le \mathcal{L}_{\text{WA}}, \tag{13}$$

where the inequality follows from convexity of softplus. This is an analytical comparison, not a reported experiment; even this normalized aggregate loss can be small while some individual Δ_t are negative.

Table 7: **Positioning against the closest supervision paradigms.** Methods differ both in where local evidence comes from and in the decision their objective supervises. RegionMax-WA compares each token's evidence winner with the complete non-evidence portion of the serving patch grid before token aggregation.

Paradigm	Local signal	Training comparison	Optimized decision
Page-only	Page relevance label	Other pages	Page ordering
AGREE	Teacher-attention patch target	Full map or salient/complement patches	Query-averaged map alignment
ColMate	OCR-spatial pretraining	OCR/text and visual tokens	TopKSim training; MaxSim serving
RegionRAG	Gold or pseudo region target	Inside/outside regions	Region-level contrast
ReAlign	Region-grounded descriptions	Candidate-document rankings	Ranking-distribution alignment
RegionMax-WA	OCR boxes + frozen selector	**Full-page complement**	**Per-token MaxSim winner**

C Evidence and Robustness Controls

Table 8: **Collection-wise matched objective ablation (nDCG@5).** Values are mean±standard deviation over seeds $\{42, 123, 3407\}$; ΔA0 is the macro-average change from the matched page-only baseline. This is the full collection breakdown of Table 2.

Variant	ESG-H	ESG	Econ.	Biomed.	Macro avg.	ΔA0
PaliGemma-3B						
A0 Page only	60.10 ± 0.13	59.20 ± 0.12	56.30 ± 0.15	55.52 ± 0.14	57.78 ± 0.11	–
A1 Region–Text	60.90 ± 0.12	59.90 ± 0.11	56.90 ± 0.14	56.26 ± 0.13	58.49 ± 0.10	+0.71
A2 RegionRank	61.70 ± 0.10	60.60 ± 0.09	57.60 ± 0.12	56.98 ± 0.11	59.22 ± 0.09	+1.44
A3 Complement-Aggregate	62.00 ± 0.09	60.90 ± 0.08	57.85 ± 0.10	57.25 ± 0.09	59.50 ± 0.07	+1.72
A4 RegionMax-WA	$\mathbf{62.55 \pm 0.10}$	$\mathbf{61.45 \pm 0.09}$	$\mathbf{58.20 \pm 0.11}$	$\mathbf{57.76 \pm 0.10}$	$\mathbf{59.99 \pm 0.09}$	**+2.21**
Qwen2.5-VL-3B						
A0 Page only	63.60 ± 0.16	62.80 ± 0.15	59.90 ± 0.18	58.54 ± 0.17	61.21 ± 0.14	–
A1 Region–Text	64.85 ± 0.15	64.00 ± 0.14	61.00 ± 0.16	59.79 ± 0.15	62.41 ± 0.13	+1.20
A2 RegionRank	66.60 ± 0.17	65.90 ± 0.16	62.80 ± 0.18	61.34 ± 0.17	64.16 ± 0.15	+2.95
A3 Complement-Aggregate	66.95 ± 0.14	66.20 ± 0.13	63.05 ± 0.15	61.68 ± 0.14	64.47 ± 0.12	+3.26
A4 RegionMax-WA	$\mathbf{67.95 \pm 0.15}$	$\mathbf{67.10 \pm 0.14}$	$\mathbf{63.95 \pm 0.16}$	$\mathbf{62.68 \pm 0.15}$	$\mathbf{65.42 \pm 0.14}$	**+4.21**

Table 9: **Design and candidate-size controls.** Values are macro nDCG@5 unless marked as differences; ± denotes standard deviation across seeds 42, 123, and 3407.

Study	Variant	PaliGemma	Qwen2.5-VL
Evidence	Top-1 WA	59.78 ± 0.09	65.08 ± 0.14
	Set-valued WA	$\mathbf{59.99 \pm 0.09}$	$\mathbf{65.42 \pm 0.14}$
Tokens	WA-AllToken	59.80 ± 0.10	65.14 ± 0.14
	WA-ContentMask	$\mathbf{59.99 \pm 0.09}$	$\mathbf{65.42 \pm 0.14}$
Size	Overall A4−A2	+0.77	+1.26
	Size-matched A4−A2	**+0.67**	**+1.08**

Set-valued evidence improves over Top-1 selection on both backbones. Content-token masking also provides a small consistent gain. The reported size-matched comparison retains 87.0% and 85.7% of the A4-over-A2 gains. Its exact candidate-sampling procedure is needed to assess how fully it controls competitor-set size.

Table 10: **Evidence-specific negative controls on ViDoRe V2.** Macro nDCG@5 mean±standard deviation over three seeds. Random evidence is area- or patch-count-matched; shuffled evidence preserves evidence-set statistics but breaks query correspondence.

Method	PaliGemma-3B	Qwen2.5-VL-3B
A2 RegionRank	59.22 ± 0.09	64.16 ± 0.15
WA-Random	59.27 ± 0.10	64.24 ± 0.14
WA-Shuffled	59.34 ± 0.10	64.38 ± 0.14
A3 Complement-Aggregate	59.50 ± 0.07	64.47 ± 0.12
A4 RegionMax-WA	$\mathbf{59.99 \pm 0.09}$	$\mathbf{65.42 \pm 0.14}$

Neither negative control reproduces A4: query-corresponding evidence yields an additional 0.65–0.72 points on PaliGemma and 1.04–1.18 on Qwen2.5-VL. Both corrupted-evidence variants nevertheless remain above A0, so these comparisons do not exclude a generic local-training contribution.

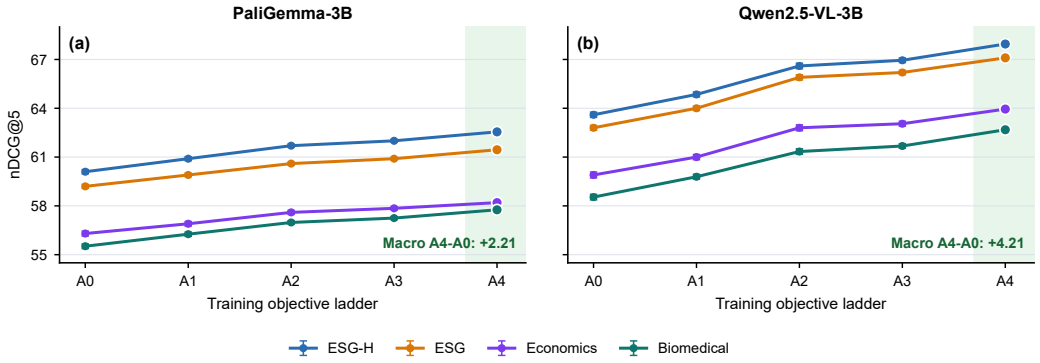

Figure 3: **Collection-wise objective progression.** Mean nDCG@5 with standard-deviation bars over three seeds for each ViDoRe V2 collection and both backbones. A0 denotes page-only training, A1–A2 region-level alternatives, A3 aggregate full-complement supervision, and A4 token-wise Winner Allocation.

D Independent Winner Grounding

Table 5 in the main paper reports the complete A0/A2/A3/A4 grounding trajectory. The examples below provide spatial illustrations of the matched A3–A4 comparison; they were not selected to estimate the prevalence of winner relocation.

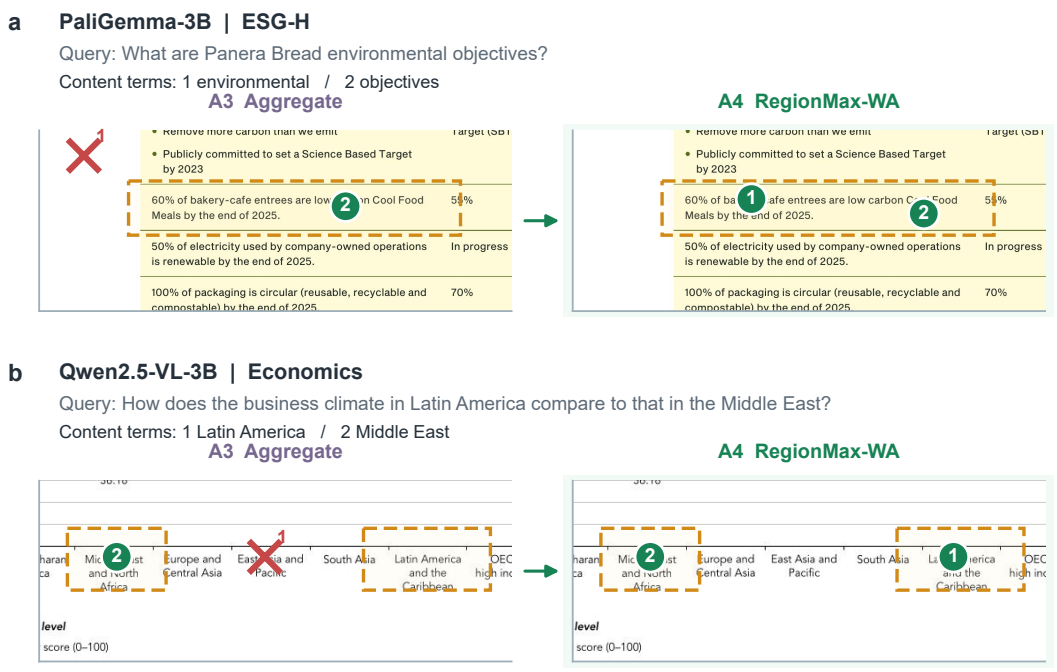

Figure 4: **Where do content-token matches land?** Read each row from A3 (left) to A4 (right) on the same evaluation page. Orange dashed boxes mark independent human evidence; numbered green circles show winners inside it, and red crosses show winners outside it. Table 5 quantifies this behavior over 250 pairs per backbone.

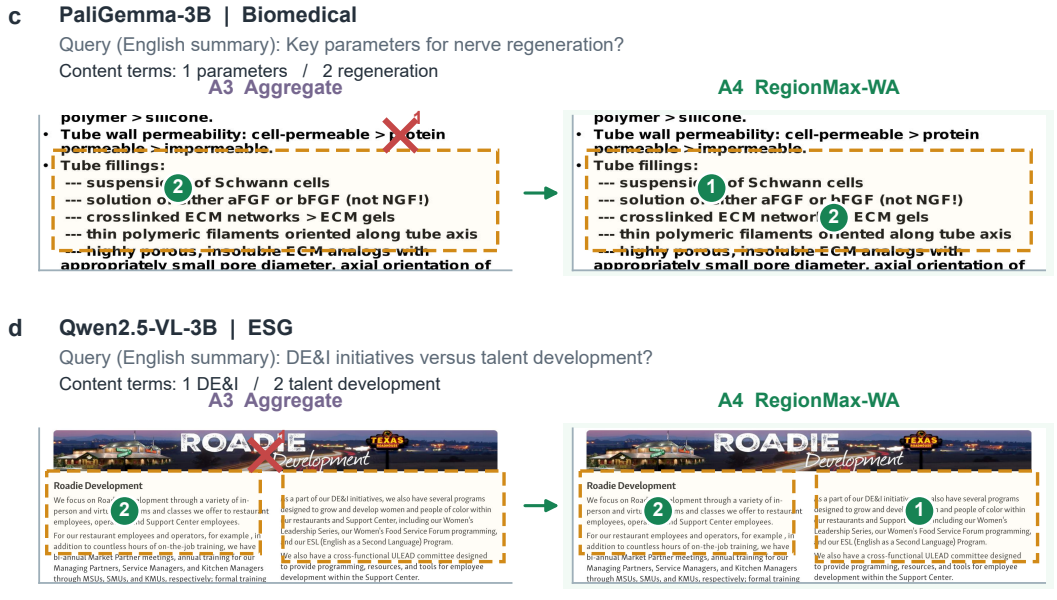

Figure 5: **Additional winner-location examples.** Original evaluation-page crops from Biomedical (PaliGemma-3B) and ESG (Qwen2.5-VL-3B), with English query summaries and vector overlays. Orange dashed rectangles denote human evidence; numerals mark displayed content terms under A3 and A4. These cases do not estimate frequency; the 250-pair summary is in Table 5.

E Data and Implementation Protocol

Table 11: **ViDoRe V2 evaluation collections.** Statistics follow the published benchmark protocol used by AGREE.

Collection	Queries	Pages	Mean relevant pages
ESG Reports Human	52	1,538	2.46
ESG Reports	228	1,538	3.89
Economics Reports	232	452	15.64
Biomedical Reports	640	1,016	3.22

Training data. We use the 118,695 English query–page pairs in `vidore/colpali_train_set`. The corpus combines DocVQA, InfoVQA, TAT-DQA, arXivQA, and web-scraped PDFs. Each query has one labeled positive page. The data are shuffled with a fixed seed, and candidate negatives are the other pages in the minibatch; no additional hard-negative mining is used.

Evidence construction. PaddleOCR PP-OCRv4 runs in English mode (`en`); OCR regions below confidence 0.5 are removed (Du et al., 2020). The frozen selector uses `BAAI/bge-base-en-v1.5` with custom mean pooling (Xiao et al., 2023). It retains the top three regions with cosine similarity at least 0.4. A page patch enters the evidence set when its intersection over patch area with an unpadded selected OCR box is at least 0.5. We set the Winner Allocation margin to $m = 0.1$ and its smoothness temperature to $\tau_w = 0.05$.

Content-token mask. The mask excludes padding and model-control tokens (<pad>, <bos>, <eos>, <unk>, <image>, <image_pad>, <|vision_start|>, and <|vision_end|>), pure-punctuation tokens, and the following case-normalized function words: articles/determiners {a, an, the, this, that, these, those}; auxiliaries/copulas {is, are, was, were, be, been, being, do, does, did, have, has, had}; prepositions/conjunctions {of, to, in, on, for, with, at, by, from, and, or, as}; and WH/modal-like terms {what, which, who, whom, whose, where, when, why, how, can, could, would, should, may, might, will}. For the example query "What is the deductible amount?", the retained content terms are "deductible" and "amount".

Backbones and optimization. The PaliGemma run initializes from `vidore/colpaligemma-3b-pt-448-base`; the Qwen run initializes from `vidore/colqwen2.5-base`. Their page representations contain 1,024 and at most 768 visual patches, respectively. Every variant is trained on four NVIDIA RTX 4090 GPUs for three epochs using per-device batch size 32, effective batch size 128, learning rate 5×10^{-4}, bfloat16 computation, gradient checkpointing, and FlashAttention-2. Optimization uses AdamW with linear decay, zero weight decay, gradient clipping at 1.0, and 100/500 warm-up steps for PaliGemma/Qwen2.5-VL. LoRA is applied with rank 32, scaling 32, dropout 0.1, Gaussian initialization, and no bias. We use the final checkpoint and batch size 1 for retrieval evaluation.

Executed objectives and reductions. The page objective is full in-batch cross-entropy with fixed temperature $\tau_p = 0.05$. The joint-loss weight is $\lambda = 0.2$. Winner Allocation first averages over content tokens, then over valid samples in the minibatch. A sample with $m_{\text{valid}} = 0$ has zero local loss and is optimized only by the page loss. In the training data, empty evidence, empty complement, and empty content-token sets occur for 4.0%, 0.08%, and 0% of examples, respectively; 95.9% of examples satisfy $m_{\text{valid}} = 1$ (rounded).

Controlled variables. For a given backbone, A0–A4 share the same training pairs, data order, in-batch candidates, patch budget, number of updates, and evaluation code. A0 uses only the page objective; each local variant adds its specified auxiliary objective. Every variant is repeated with seeds 42, 123, and 3407. OCR, region selection, and patch mappings are cached so that A1–A4 receive the same proposal inputs. The human-box annotations are excluded from all training and tuning; the evaluation-page evidence audit in Table 3 is descriptive only.

Compute and deployment. Training takes approximately 16 hours for PaliGemma-3B and 24 hours for Qwen2.5-VL-3B on four RTX 4090 GPUs. OCR, selector inference, and box-to-patch mapping occur only during training and can be cached. The deployed model stores the same *format* of full-page vectors and uses the same MaxSim scorer as its corresponding baseline. Updating the document encoder requires rebuilding its corpus embeddings.

F Evidence-Construction Statistics

Table 3 in the main paper summarizes OCR-region counts, selected-region counts, patch coverage, and valid-WA rates. The corresponding empty-evidence rates are 4.0% in training and 3.6%, 4.0%, 4.5%, and 3.8% on ESG-H, ESG, Economics, and Biomedical, respectively. The selector typically retains fewer than the maximum of three regions because candidates must also satisfy both OCR-confidence and selector-score thresholds. A pair without selected evidence or eligible content tokens is omitted from the local loss but remains in page-level training. The evaluation-page audit is descriptive and was not used to fit the selector or tune a checkpoint.

G Paired Retrieval Uncertainty

Table 6 in the main paper gives the intervals. Paired resampling keeps each method comparison on the same evaluation queries and recomputes the macro nDCG@5 difference, complementing the three-seed variation reported in Tables 1 and 2. The intervals assess retrieval differences, not uncertainty in the separate human-box grounding point estimates.

H Human-Annotation Protocol

The grounding set contains 250 query–positive-page pairs from the ViDoRe V2 test split: 63 from ESG Reports Human, 63 from ESG Reports, 62 from Economics Reports, and 62 from Biomedical Reports. The same pairs are used for both backbones and for A0, A2, A3, and A4. Multiple disjoint evidence boxes are allowed. Two research assistants with machine-learning and document-understanding backgrounds independently label each pair; a third research-team member resolves disagreements. Annotators receive brief written guidance and example-based training, are compensated according to the project's internal standard, and provide informed consent before participating. An inter-annotator agreement statistic was not computed, so no agreement value is reported. Winner Hit and Human Margin are computed after annotation and do not affect training or model selection.

I Serving Efficiency

Table 12: **Serving cost relative to the matched A0 baseline.** Ranges summarize the two backbones. Values near one confirm that training-time supervision does not materially change deployment.

Metric	A0	A4 RegionMax-WA
Index size	1.000×	1.000×
Query latency	1.000×	1.00–1.01×
Throughput	1.000×	0.99–1.00×
Peak inference memory	1.000×	1.000×

J Reproducibility and Ethics

The matched training protocol, exact model identifiers, evidence filters, optimization settings, and seeds are specified above. Main Table 1 reports three-seed means, with collection-wise deviations in Appendix Table 8; main Tables 2 and 4 show mean±standard deviation. Paired retrieval intervals and human-box point estimates appear in main Tables 6 and 5. Per-seed evaluation outputs, per-pair grounding records, and implementation artifacts are not included in this review draft. The reported

sensitivity sweeps leave the default settings unchanged, and no ViDoRe V2 test query is used to select a checkpoint.

The 250 human-box pairs are used only for aggregate evaluation, not for training. Annotators participate with informed consent and are compensated; no identifying annotator information is reported. Institutional-review applicability should be checked against the authors' local requirements before submission.