READ, THEN RETRIEVE: READER FEEDBACK FOR BUDGETED MULTIMODAL DOCUMENT SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal document retrieval typically applies an expensive reader to a fixed set of candidate pages, leaving relevant evidence outside that set inaccessible to reranking. We propose Reader Feedback, a two-batch acquisition method that uses intermediate reader judgments to select subsequent pages under a fixed reading budget. A frozen 7B reader first evaluates 16 visual candidates. A lightweight selector combines their relevance scores with semantic similarity, document membership, and uncovered query terms to choose 16 additional pages from visual and text retrieval pools. The same reader and a shared ranking head then order all 32 pages. Across three ViDoRe V3 development corpora, the fraction of queries with every annotated relevant page acquired increases from 55.34% to 62.43% over visual top-32 acquisition. On a separate HR corpus evaluated after method freezing, this coverage increases from 49.69% to 62.58%, while nDCG@10 rises from 64.97 to 67.75. Matched controls show a 5.97-point coverage gain over uniform feedback on HR, but the nDCG advantage over static hybrid retrieval is 0.13 points with an interval spanning zero. These results identify evidence acquisition as a useful role for reader feedback while distinguishing coverage gains from improvements in final ranking.

1 INTRODUCTION

Retrieving evidence from visually rich documents requires matching questions to pages that combine text, tables, diagrams, and layout. Visual retrievers preserve these signals in page representations, while multimodal readers provide more detailed judgments of query–page relevance (Faysse et al., 2025; Yu et al., 2025; Liu et al., 2025). A common retrieval pipeline uses an efficient retriever to form a candidate set and an expensive reader to rerank it. This separation makes candidate acquisition a bottleneck: evidence excluded by the first stage remains unavailable to the reader, regardless of the quality of its relevance judgments.

This bottleneck is particularly relevant when support is distributed across several document pages. In an industrial query about the maximum hot-refueling pressure for the AV-8 aircraft, two annotated support pages appear at visual ranks 33 and 35, just beyond a 32-page reading budget (Figure 1). An earlier page already states the pressure limit, but the fixed candidate set omits these additional supporting pages. Text retrieval can expose them through their wording; the remaining challenge is to allocate the limited reader budget among visual and textual matches. Expanding the candidate set increases the number of expensive judgments, while selecting the entire set in advance leaves the reader's intermediate observations unused.

We investigate whether intermediate reader judgments can improve the selection of subsequent pages. A relevant page provides several signals for further acquisition: its content identifies semantically related passages, its document membership links it to other pages from the same source, and its text indicates which query terms remain uncovered. These signals can guide the remaining reads, but they can also reinforce an overly narrow interpretation of the query. Evaluating feedback therefore requires comparisons that separate its contribution from the complementary evidence already available through text retrieval.

We propose *Reader Feedback*, a two-batch acquisition method with a fixed budget of 32 page judgments. The first batch contains the 16 highest-ranked visual pages. A shallow supervised selector uses their reader scores to compute semantic support, same-document support, and uncovered query

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.

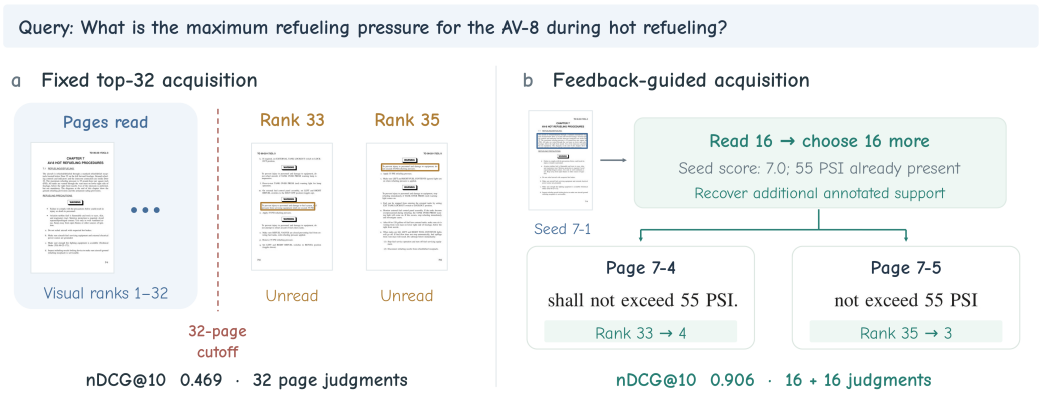

Figure 1: Reader feedback recovers additional support under a fixed reading budget. In industrial query 97, an initial seed already states the 55 PSI limit. Two further annotated support pages lie at visual ranks 33 and 35: fixed top-32 acquisition leaves them unread, whereas feedback acquisition reads both and places them at final ranks 4 and 3. Thumbnails and magnified snippets are original PDF crops. Section 6 examines this coverage gain alongside a failure case.

terms for unread candidates. It combines these signals with retrieval features to select 16 additional pages from a flat visual, dense-text, and lexical pool. The same frozen 7B reader evaluates the second batch, and a shared score head ranks the union. This design integrates the reader into candidate selection while preserving its weights, prompt, and page-input format.

Our evaluation distinguishes evidence acquisition from final ordering. We measure relevant-page recall and the fraction of queries for which every annotated relevant page is acquired, alongside nDCG@10. Static hybrid retrieval controls for access to complementary text proposals; a separately trained selector with uniform seed confidence controls for the reader's differentiated judgments. All methods share the reader and final ranking head. We develop the method on three ViDoRe V3 corpora and evaluate a fixed configuration on a separate HR corpus, with complete predictions saved before its relevance labels are obtained.

Across the three development corpora, Reader Feedback raises all-page coverage from 55.34% to 62.43% over visual top-32 acquisition. On HR, it improves coverage from 56.60% to 62.58% over uniform feedback, with a document-family bootstrap interval of $[+0.95, +7.06]$ points. The ranking gains are more limited: HR nDCG reaches 67.75, compared with 64.97 for visual acquisition and 67.61 for static hybrid retrieval; the latter difference is not statistically resolved. The results support using reader judgments to assemble a more complete annotated evidence set, while showing that this benefit need not translate into a comparable improvement in the top-ranked results.

Our contributions are threefold:

- We introduce a two-batch acquisition design that uses relevance judgments from a frozen multimodal reader to allocate the remainder of a fixed page-reading budget.
- We isolate the effects of text proposals, differentiated reader feedback, and final score correction through matched controls and separate coverage and ranking measurements.
- We characterize the gains and failure modes across three development corpora and one frozen target corpus, including evidence displacement and the limited benefit of query decomposition and signed feedback.

2 Related Work

Visual and hybrid document retrieval. ColPali encodes document-page images for retrieval, while VisRAG uses visual document inputs throughout retrieval-augmented generation (Faysse et al., 2025; Yu et al., 2025). Qwen3-VL retrieval models extend multimodal representations to embeddings and reranking (Li et al., 2026). Hybrid approaches combine visual and textual evidence:

VisDoM integrates both modalities for multi-document question answering, and PREMIR indexes questions generated from document content to connect user queries to pages (Suri et al., 2024; Choi et al., 2025). Our method uses visual representations and extracted page text as proposal sources. Its acquisition decision additionally conditions on relevance judgments from pages already inspected. Static hybrid fusion provides a control for the contribution of text–image complementarity alone.

Iterative evidence gathering. MoLoRAG studies logic-aware multimodal retrieval, SimpleDoc combines dual-cue retrieval with iterative refinement, and MAB-DQA allocates attention across query aspects through a bandit formulation (Wu et al., 2025; Jain et al., 2025; Xiang et al., 2026). These approaches motivate adapting retrieval as evidence becomes available. We study this principle in a bounded page-retrieval setting: a single feedback step redistributes a fixed number of frozen-reader judgments. Query-aspect allocation further motivates our decomposition and coverage ablations. These experiments evaluate specific acquisition components under a shared reader; they do not reproduce the complete question-answering pipelines of the cited systems.

Structure and multimodal ranking. LAD-RAG exploits document layout for dynamic retrieval, and EviProp propagates relevance over chunk–page graphs (Sourati et al., 2026; Zhang et al., 2026). Reader Feedback operates on a flat candidate pool, using document membership as a source-consistency signal. Its adaptive selection is driven by reader judgments collected during inference. LamRA provides the multimodal relevance reader (Liu et al., 2025); we use the later author-released Qwen2.5-VL-7B checkpoint. Standard LambdaMART models implement both the selector and the shared score correction (Burges, 2010; Ke et al., 2017). The resulting comparison isolates how page judgments are allocated with the same underlying models.

3 Two-Batch Evidence Acquisition

3.1 Setting and flat candidate pool

Let q denote a question and $\mathcal{P}$ a page corpus. A visual retriever induces an initial ordering $r_v(q, p)$ with one-based ranks, and a multimodal reader returns a relevance score $s(q, p)$ from the question and page image. Given a budget $B = 32$ distinct page judgments, we acquire a set $C(q)$ in two batches of $b = 16$ pages and rank the resulting set. Training and evaluation use page-relevance labels; the task is page retrieval.

We index automatically extracted PDF blocks with a small text encoder and BM25. The extraction pipeline uses native PDF text and falls back to the benchmark's automatic markdown for pages without native text. Dense page similarity is the maximum nonnegative query–block cosine, and lexical page similarity is the maximum block BM25 score. BM25 inputs prepend the source-document identifier to each block; dense inputs contain block text alone. The candidate pool is

$$U(q) = \mathrm{Top}_{128}(r_v) \cup \mathrm{Top}_{128}(r_d) \cup \mathrm{Top}_{128}(r_l), \tag{1}$$

where r_d and r_l denote dense and lexical rankings. Industrial uses its available top-100 visual cache, with 128 proposals retained from each text component. Pages without extracted text remain eligible through visual retrieval; their raw text similarities and anchor-support features are zero, while corpus-wide rank features remain defined. All candidates are individual pages in a flat index.

3.2 Read first, then condition acquisition on feedback

The first batch $S(q) = \mathrm{Top}_{16}(r_v)$ is identical for every method. Each page is evaluated once using the unchanged reader prompt. The reader score is the logit difference between its next-token `Yes` and `No` outputs. We map this score to a bounded weight

$$c_a = \sigma(s(q, a)/2)\,, \qquad a \in S(q). \tag{2}$$

The fixed sigmoid transform supplies bounded weights rather than calibrated relevance probabilities. Seed pages without indexed text receive visual judgments but supply no text anchors. The following maxima therefore range over text-bearing seeds, with an empty maximum defined as zero.

For every text-bearing seed page a, let z_a be its block embedding with the highest question similarity. For an unread page p, three features condition selection on the observed scores. First, semantic

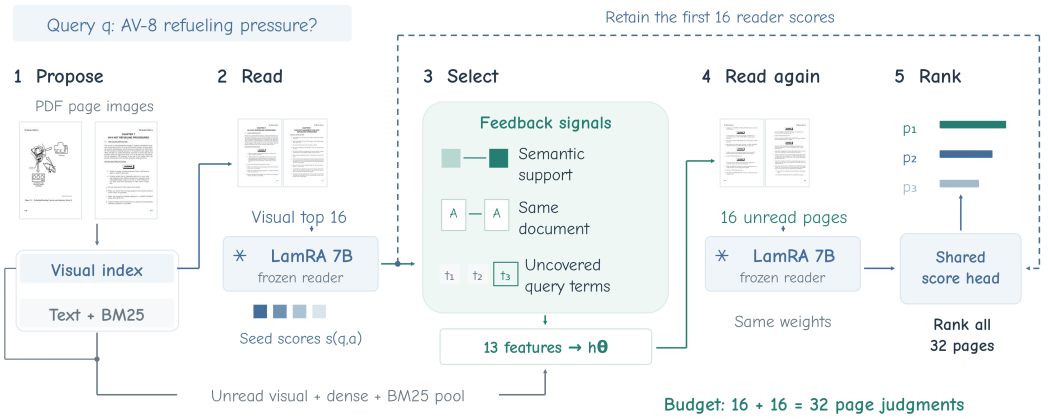

Figure 2: Reader-guided acquisition with two batches. Visual top-16 pages are judged by the frozen reader; their scores condition selection from unread visual, dense-text, and BM25 proposals. The same reader judges 16 selected pages. The dashed route retains the first batch's scores for the shared final head, which ranks all 32 pages. Page images come from the corpus; feature icons and ranking bars are schematic. All stages condition on the original query.

support transfers confidence to a page whose blocks resemble a promising anchor:

$$f_{\text{sem}}(p) = \max_{a \in S(q),\, z \in p} c_a \left[\cos(z, z_a)\right]_+. \tag{3}$$

Second, source consistency uses the strongest seed confidence from the same document:

$$f_{\text{doc}}(p) = \max_{a \in S(q):\, d(a) = d(p)} c_a, \tag{4}$$

where $d(p)$ denotes the source document of page p. Document membership defines the matching relation; document identifiers are not learned as categorical features.

Third, an uncovered-term feature rewards query words that have not appeared on high-confidence seed pages. Let T_q be the query terms retained by the lexical index, w_t their normalized inverse-document-frequency weights, and I_{pt} indicate whether a term occurs in page p's indexed text. Then

$$o_t = \max_{a \in S(q)} c_a I_{at}, \qquad f_{\text{gap}}(p) = \sum_{t \in T_q} w_t I_{pt}(1 - o_t). \tag{5}$$

This feature measures query-term coverage missing from confident seeds. It guides acquisition through lexical evidence; the subsequent image judgment evaluates the selected page's relevance.

These three feedback features join nine retrieval features and an unweighted anchor-similarity feature, yielding 13 inputs. The retrieval features encode visual rank, visual top-32 membership, dense and lexical scores and ranks, query-term coverage, and text availability (Appendix A). A shallow LambdaMART model h_θ scores each unread page in $U(q) \setminus S(q)$. Its highest-scoring 16 distinct pages form the second batch $A(q)$, giving $C(q) = S(q) \cup A(q)$ and exactly 32 acquired pages.

3.3 Training and Final Ranking

We fit the selector with standard graded page-relevance labels, grouping candidates by query and assigning equal total weight to each training corpus. It uses 96 trees, at most seven leaves per tree, depth three, and a learning rate of 0.05. No target-domain labels, answer strings, query categories, or gold source-document identifiers enter acquisition. Corpus-provided page-to-document membership is available for all methods.

Both batches are evaluated with the same question-and-image reader. Every primary comparison shares a source-trained score head with four inputs: the raw reader score, within-set reader rank, relative initial rank among acquired pages, and standardized reader score. A 64-tree LambdaMART

Table 1: Evaluation scope. Documents are counted by the benchmark's source-document identifiers. Text-bearing pages have at least one extracted block in the reused index. Only English queries are evaluated.

Corpus	Queries	Pages	Documents	Text-bearing	Role
Industrial	283	5,244	27	5,039	Development
Finance	309	2,942	6	2,926	Development
Pharma	364	2,313	52	2,313	Development
HR	318	1,110	14	1,099	Frozen evaluation

residual is added to the standardized score. The head is trained on fixed visual top-32 sets and held constant across acquisition policies. Direct reader-score ranking provides an additional comparison without this correction.

The method retains the 7B reader's architecture and prompt and adds a text index, per-query CPU features, and a learned selector. Comparisons match the number of page judgments and their visual-token cap. Total computation can still differ through indexing, feature construction, and the synchronization required between the two batches.

4 Experimental Protocol

Data and separation of development from evaluation. We use English queries from four ViDoRe V3 corpora: industrial, finance, pharma, and HR (Loison et al., 2026). Table 1 summarizes their sizes. Industrial, finance, and pharma inform method development. For each reported source result, both learned components are fitted on the other two corpora. These corpora were repeatedly inspected during design, so the source results measure development performance with held-out parameter fitting, rather than independent generalization.

HR provides a separate evaluation after method freezing. We fix the acquisition policies, training settings, and ranking mode using source results, fit the selector on all three source corpora, and use the all-source score head. HR pages and questions are available for indexing, while relevance labels are withheld until complete candidate and final rankings have been saved for every method. The reported HR results use this fixed configuration without target-label tuning.

Models and inputs. Industrial uses cached EVIE-4.5B retrieval rankings (Wang & Wen, 2026). Finance, pharma, and HR use Qwen3-VL-Embedding-2B (Li et al., 2026); each corpus has identical initial rankings across acquisition methods. The dense text index uses `all-MiniLM-L6-v2` with at most 128 tokens per block (Reimers & Gurevych, 2019; Wang et al., 2020). The reader is the author-released `code-kunkun/LamRA-Rank-Qwen2.5VL-7b` checkpoint based on Qwen2.5-VL (Liu et al., 2025; Bai et al., 2025). It receives the original question and page image under the released binary relevance prompt, with a cap of 1,280 visual tokens per page. Its weights and prompt remain frozen. Experiments use existing benchmark annotations without additional human annotation or evaluation.

Matched controls. **Visual top-32** reads the first 32 visual candidates. **Static hybrid** keeps the same first 16, then selects 16 unread pages by equal-weight reciprocal-rank fusion of the visual, dense-text, and lexical lists, using offset 60 (Cormack et al., 2009). **Uniform feedback** uses the same pool, architecture, and training protocol as our selector but sets all text-anchor confidences to one during both training and prediction. It retains semantic and document information while removing the reader's differentiated judgments. **Reader feedback** uses the actual seed scores. All four methods make exactly 32 distinct page judgments, then use the same score head. The static method is a fusion baseline, not a reproduction of the full VisDoM system.

Metrics and uncertainty. Our ranking metric is nDCG@10 with the benchmark's linear graded gains. We additionally report unweighted relevant-page recall within the 32 acquired pages and All@32, the fraction of queries whose acquired set includes every positively judged page. These are annotation-based coverage metrics: relevance sets may be incomplete, and All does not establish that an answer can be composed correctly. Oracle nDCG sorts acquired pages by their known grades and

Table 2: Source development results. The first four numeric columns are nDCG@10; Recall and All measure coverage in the 32 acquired pages. Values are multiplied by 100. Each source corpus is excluded from both fitted models used to evaluate it; these corpora nevertheless informed method development. All rows share the same reader and final score head.

Acquisition	Ind.	Fin.	Pharma	Mean	Recall	All
Visual top-32	60.58	68.64	70.22	66.48	77.37	55.34
Static hybrid	59.84	70.88	70.91	67.21	78.89	56.74
Uniform feedback	60.15	70.08	70.78	67.01	79.42	59.09
Reader feedback (ours)	60.56	70.13	71.40	67.36	81.23	62.43

Table 3: Frozen evaluation on 318 HR queries. nDCG uses the shared source-trained head; Reader only omits that head. All methods were fixed and their complete predictions saved before target relevance labels were acquired. Values are multiplied by 100.

Acquisition	nDCG	Reader only	Recall	All	Oracle
Visual top-32	64.97	64.34	75.42	49.69	84.11
Static hybrid	67.61	67.03	82.60	59.12	89.98
Uniform feedback	66.74	65.89	80.66	56.60	87.85
Reader feedback (ours)	67.75	66.99	83.82	62.58	89.88

bounds performance for a fixed acquired set; it is never used for prediction. All reported scores are multiplied by 100.

Source means weight the three corpora equally. Paired query bootstrap intervals use 10,000 draws for source means and 5,000 for HR; HR also uses 5,000 connected-document-family resamples to account for queries sharing source documents. These intervals describe uncertainty within the observed domains. Finance's relevant-document graph has only one connected component, so a source macro family interval is not estimable. Intervals are descriptive, with no adjustment for repeated source development or multiple comparisons. An independent `pytrec_eval` implementation reproduces each query's final nDCG to numerical precision.

5 Results

5.1 Source evidence: acquisition improves more than ordering

Reader Feedback increases source macro nDCG from 66.48 to 67.36 over visual top-32 acquisition (Table 2), a paired difference of $+0.89$ points with a query-bootstrap interval of $[+0.41, +1.37]$. Finance and pharma account for the improvement; industrial remains similar at 60.56 versus 60.58.

Static hybrid acquisition reaches 67.21 nDCG. The feedback difference is $+0.15$ points with interval $[-0.43, +0.69]$; relative to uniform feedback, it is $+0.36$ with interval $[-0.06, +0.77]$. Both intervals span zero, leaving the incremental ranking benefit over these controls unresolved.

Candidate recall rises from 77.37 to 81.23 against visual acquisition, and $\mathrm{All@32}$ rises from 55.34 to 62.43. Static hybrid reaches 78.89 recall and 56.74 all-page coverage; uniform feedback reaches 79.42 and 59.09. Against static hybrid, feedback's all-page coverage difference is $+5.69$ points, with descriptive interval $[+3.80, +7.57]$; against uniform feedback it is $+3.35$ $[+1.99, +4.72]$. The largest change is thus in the completeness of the acquired set. A relevant page can be recovered yet remain below rank ten, or enter the top ten while displacing another relevant page. Consequently, coverage gains need not appear at the same magnitude in nDCG. Appendix B reports further coverage intervals.

5.2 Frozen evaluation on HR

On the frozen HR evaluation, Reader Feedback reaches 67.75 nDCG, compared with 64.97 for visual top-32, 67.61 for static hybrid, and 66.74 for uniform feedback (Table 3). The gain over visual acquisition is $+2.77$ points, with query interval $[+1.61, +4.00]$ and document-family inter-

Table 4: Source macro nDCG@10 under different readouts. The oracle sorts acquired pages by benchmark relevance and is an evaluation bound, not an implemented method. Coverage gains leave a substantial ranking gap.

Acquisition	Reader only	Shared head	Candidate oracle
Visual top-32	65.82	66.48	84.60
Static hybrid	66.69	67.21	85.83
Uniform feedback	66.52	67.01	85.89
Reader feedback (ours)	66.77	67.36	87.09

val $[+1.09, +6.28]$. Family resampling uses six connected groups, limiting the precision of the dependence-aware estimate.

The nDCG difference from static hybrid is $+0.13$ points, with query interval $[-0.86, +1.04]$ and family interval $[-2.07, +0.76]$. Direct reader-score ranking slightly favors static hybrid, 67.03 versus 66.99. Against uniform feedback, the difference is $+1.01$, with both intervals again including zero. Final-ranking performance is therefore not clearly separated from either control.

Candidate evidence gives a more specific result. Reader feedback reaches 83.82 recall and 62.58 all-page coverage. Uniform feedback reaches 80.66 and 56.60; the respective differences are $+3.15$ and $+5.97$ points. Their family intervals are $[+2.47, +5.18]$ and $[+0.95, +7.06]$. Thus, differentiated seed scores improve evidence coverage over the same learned policy trained with uniform confidence in this target corpus. Static hybrid remains a stronger alternative: feedback adds 1.22 recall points and 3.46 all-page points, with family intervals that include or touch zero. Most of the improvement over visual-only acquisition comes from access to complementary text proposals, while feedback supplies a more limited additional contribution.

The HR results locate feedback's clearest contribution in evidence coverage relative to uniform confidence. Static hybrid retrieval remains competitive, particularly for final ranking. HR was evaluated once after method freezing and is reported separately from the source development results.

6 Analysis

6.1 The final reader remains a bottleneck

Direct reader scores yield source macro nDCG values of 65.82 for visual acquisition and 66.77 for feedback acquisition. The shared head raises these to 66.48 and 67.36, respectively (Table 4). The gain over visual acquisition thus persists without score correction. Feedback's candidate oracle reaches 87.09, substantially above its realized score of 67.36. This upper bound indicates that much of the relevance available in the acquired set is not reflected in its final ordering.

Using a common ranking head separates acquisition gains from score correction. Reporting coverage alongside nDCG further distinguishes access to relevant pages from their positions in the final list. The two measurements reveal different outcomes: feedback substantially improves candidate completeness on the source corpora, while its incremental ranking gain over hybrid retrieval remains small. This distinction matters for applications that consume an evidence set as well as those that return a short ranked list.

6.2 Completeness and top-ranked relevance are different targets

Coverage and ranking need not agree even under an ideal reader. Recall counts every positively judged page equally, while nDCG weights relevance grades and discounts position. Recovering several low-grade pages while losing a higher-grade page can increase recall yet lower oracle nDCG. All-page coverage adds another distinction: a single missing positive page makes a query incomplete, even when its most strongly judged pages have already been acquired. It describes completeness relative to the annotation set, rather than the utility of the first few results.

HR illustrates this difference even under ideal ordering. Feedback attains higher recall and all-page coverage than static hybrid, yet its candidate oracle is slightly lower: 89.88 versus 89.98. The

Table 5: Additional modules evaluated on source development corpora. The final column counts extra query-generation calls; all rows use 32 page judgments from the same 7B reader. The shuffled row randomizes only the four added signed features' seed confidences, retaining the original feedback features. None of these extensions enters the frozen primary method.

Source-only variant	nDCG	Recall	All	3B calls
Reader feedback	67.36	81.23	62.43	0
+ Facet proposals only	67.36	81.23	62.43	1
+ Facet features	67.28	81.14	62.01	1
+ Greedy facet coverage	67.42	81.26	62.12	1
Greedy, uniform feedback	67.15	79.52	59.25	1
+ Signed feedback	67.38	81.29	62.33	0
Signed-feature shuffle	67.46	81.14	62.05	0

acquired sets therefore favor different relevance objectives. Their coverage comparison measures recovery of annotated pages; it cannot establish an advantage in graded top-ten relevance or downstream answer sufficiency.

The fixed budget also creates a displacement cost. Adding pages to a candidate set cannot lower its oracle, since the previous ideal ordering remains available. Our compared sets have equal size and are generally not nested: each newly selected page can replace an initial visual candidate. Feedback replaces an average of 7.70 industrial, 6.89 finance, and 9.03 pharma pages relative to visual top-32. The recovery and loss counts below quantify the resulting tradeoff.

6.3 Query decomposition and signed-feedback ablations

Inspired by query-aspect allocation, we test a text-only Qwen2.5-VL-3B call that decomposes each original question into up to four short information needs. These subqueries add dense proposals and four facet-affinity features; a further variant greedily rewards relevance plus uncovered facet coverage. The decomposition prompt receives neither document content nor benchmark answers. Adding facet proposals alone changes none of the selected pages. Facet features reduce source nDCG from 67.36 to 67.28; greedy coverage reaches 67.42, only 0.05 above the primary. Its paired interval against the primary spans $[-0.19, +0.36]$. All variants retain 32 page judgments, but decomposition adds model execution and generated tokens. We therefore exclude it from the frozen method.

Four additional features encode low-confidence seeds: negative semantic support, its difference from positive support, signed same-document confidence, and maximum negative same-document confidence. This variant reaches 67.38 nDCG. Shuffling seed confidences only for these four features yields 67.46, with the original 13 features unchanged. The real-minus-shuffled difference is -0.07, with interval $[-0.34, +0.23]$ (Table 5). These results provide no evidence of an incremental benefit from the proposed signed signal, so the frozen method retains the original 13 features.

6.4 Evidence recovery and displacement

Industrial query 97 illustrates recovery of additional support (Figure 1). A seed from the AV-8 procedure already states the 55 PSI limit and receives a reader logit difference of 7.0, giving the two unread support pages a document-support weight of 0.971. Feedback acquires both and places them at final ranks 4 and 3, raising nDCG from 0.469 to 0.906. Static hybrid acquires one and reaches 0.704; uniform feedback misses both and remains at 0.469. The trace links the observed reader signal to the subsequent acquisition decision, although it does not isolate a single feature's causal contribution. The recovered pages expand annotated support for a limit already present in the initial evidence.

A complementary failure occurs in industrial query 58, which asks how piping material selection balances safety and durability in live-fire training. A relevant page is at rank 3 after fixed visual reranking but falls outside the acquired set under feedback, and nDCG drops from 0.195 to zero. Static hybrid and uniform feedback also fail on this question. Replacing the second half of visual candidates can therefore discard useful evidence. This tradeoff is visible across the industrial corpus: feedback acquires 73 new relevant query–page pairs but loses 63 that were in visual top-32.

Finance gains 122 and loses 9; pharma gains 182 and loses 33. The selector has no guarantee of retaining every useful visual candidate, and high-confidence seed similarity can reinforce a narrow interpretation of the question. Both cases were selected after source evaluation and are explanatory examples, not independent evidence of average performance.

7 LIMITATIONS

The evaluation covers English page retrieval in four corpora from one benchmark. Three corpora informed repeated method development, and HR supplies a single independent target after freezing. The learned components use standard tree-based ranking on existing representations; the study evaluates acquisition with a fixed reader rather than introducing a new foundation model. Results establish matched comparisons within this setting and do not support a state-of-the-art claim against systems evaluated with different queries, candidate pools, or readers.

The matched budget counts page judgments with the same visual-token cap; it does not equalize native-text extraction, index storage, CPU feature computation, actual image token counts, or the synchronization latency of two stages. Native text is incomplete on some pages, and the method has no dedicated OCR or table parser. Document membership is assumed available. Industrial uses a different initial retriever and a shorter visual proposal cache than the other corpora, preventing clean attribution of cross-domain differences to document content alone.

Finally, benchmark page relevance is not evidence sufficiency for downstream question answering. We measure neither generated-answer correctness nor robustness to document edits, multilingual queries, adversarial pages, or evolving corpora. Improved acquisition can leave the final reader's ordering errors unresolved and can sacrifice initially promising pages. These limitations constrain the method to the evaluated retrieval setting.

8 CONCLUSION

We presented Reader Feedback, a two-batch method that uses a frozen multimodal reader's relevance judgments to select subsequent pages under a fixed reading budget. Semantic support, document consistency, and uncovered query terms guide a lightweight selector over visual and text proposals. Experiments show improved annotated evidence coverage, including over uniform feedback on a frozen target corpus, while final nDCG remains close to a strong static hybrid baseline. The findings establish the value and limits of using reader feedback for acquisition: a more complete evidence set is a measurable benefit, but its effective ordering remains a separate challenge.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An Overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010. URL https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/.

Yejin Choi, Jaewoo Park, Janghan Yoon, Saejin Kim, Jaehyun Jeon, and Youngjae Yu. Zero-shot multimodal document retrieval via cross-modal question generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26068–26083, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1324. URL https://aclanthology.org/2025.emnlp-main.1324/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009. URL https://research.google/pubs/reciprocal-rank-fusion-outperforms-condorcet-and-individual-rank-learning-methods/.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. ColPali: Efficient Document Retrieval with Vision Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99e9e141aafc314f76b0ca3dd66898b3-Abstract-Conference.html.

Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Shengyu Dai, Zhenwen Shao, Qingyun Wu, and Huazheng Wang. SimpleDoc: Multi-Modal Document Understanding with Dual-Cue Page Retrieval and Iterative Refinement. *arXiv preprint arXiv:2506.14035*, 2025. URL https://arxiv.org/abs/2506.14035.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4015–4025, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_LamRA_Large_Multimodal_Model_as_Your_Advanced_Retrieval_Assistant_CVPR_2025_paper.html.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel de Souza P. Moreira, Bo Liu, Manuel Faysse, Celine Hudelot, and Gautier Viaud. ViDoRe v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16570–16600, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.755. URL https://aclanthology.org/2026.acl-long.755/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Zhivar Sourati, Zheng Wang, Marianne Menglin Liu, Yazhe Hu, Mengqing Guo, Sujeeth Bharadwaj, Kyu J. Han, Tao Sheng, Sujith Ravi, Morteza Dehghani, and Dan Roth. LAD-RAG: Layout-aware dynamic RAG for visually-rich document understanding. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15945–15968, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.724. URL https://aclanthology.org/2026.acl-long.724/.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2412.10704*, 2024. URL https://arxiv.org/abs/2412.10704.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zifei Wang and Wei Wen. EVIE: High-Performance Multilingual Visual Document Retrieval with Matryoshka Embeddings and Token Compression. Author software and model release, 2026. URL https://github.com/Tencent/EVIE.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. MoLoRAG: Bootstrapping document understanding via multi-modal logic-aware retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14024–14045, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.708. URL https://aclanthology.org/2025.emnlp-main.708/.

Yixin Xiang, Yunshan Ma, Xiaoyu Du, Yibing Chen, Yanxin Zhang, and Jinhui Tang. MAB-DQA: Addressing query aspect importance in document question answering with multi-armed bandits. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22973–22992, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1053. URL https://aclanthology.org/2026.acl-long.1053/.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3640a1997a4c9571cea9db2c82e1fc35-Abstract-Conference.html.

Hongwei Zhang, Xiaoman Wang, Zehui Ling, Ruicheng Zhu, Yue Zhang, Pinlong Cai, Fuke Shen, Botian Shi, Tongquan Wei, and Guohang Yan. EviProp: Seeded Relevance Diffusion on Chunk-Page Graphs for Long Multimodal Document Retrieval. *arXiv preprint arXiv:2606.08979*, 2026. URL https://arxiv.org/abs/2606.08979.

AI USE

Generative AI tools were used primarily to improve the readability and language of the manuscript.

A EXACT FEATURE DEFINITIONS AND TRAINING

The selector has 13 features, listed in Table 6. Dense scores are maximum nonnegative block cosines. Lexical scores are maximum block BM25 values. All ranks are one-based; the reciprocal-rank feature is $61/(60+r)$. A page absent from the available visual ranking receives visual reciprocal rank zero. Dense and lexical ranks cover all indexed pages, with deterministic tie breaking. Their rank features remain defined when a page's raw text score is zero.

Both learned components use LightGBM's standard LambdaMART implementation. The selector has 96 trees and truncation level 16. The final head has 64 trees and truncation level 10. Both use learning rate 0.05, maximum depth 3, at most 7 leaves, minimum 64 examples per leaf, L_2 regularization 1, seed 2027, four CPU threads, and deterministic column-wise construction. Gains equal integer relevance grades. There is no target-domain early stopping. Each training corpus

Table 6: Selector inputs. Only features 10–12 depend on actual seed-reader scores. Uniform feedback recomputes these inputs with every available seed confidence set to one.

	Feature	Definition
1	Visual rank	$61/(60 + r_v)$, or zero if unavailable.
2	Visual top-32	Membership in the original top 32.
3	Dense similarity	Maximum nonnegative query–block cosine.
4	Dense standardized score	Feature 3 standardized over corpus pages.
5	Relative lexical score	Page BM25 divided by the maximum page BM25 for the query.
6	Dense rank	$61/(60 + r_d)$.
7	Lexical rank	$61/(60 + r_l)$.
8	Query-term coverage	Sum of normalized query-term IDF weights present in the indexed page text.
9	Text availability	Whether at least one extracted block is available.
10	Semantic feedback	Confidence-weighted maximum similarity to seed anchor blocks.
11	Document feedback	Maximum confidence of a text-bearing seed from the same document.
12	Uncovered terms	IDF-weighted term coverage discounted by confidence-weighted seed coverage.
13	Anchor similarity	Maximum similarity to any seed anchor, without confidence weights.

receives equal total sample weight; within a corpus, every candidate row has equal weight. Query groups define the ranking objective. The final head uses the standardized reader logit as its initial score and adds its predicted residual at inference.

The four head features are s_p, the descending reader-rank value linearly mapped from 1 to 0 across 32 pages, the analogous relative initial-rank value, and $(s_p - \bar{s})/\max(10^{-6}, \mathrm{std}(s))$. Initial rank is computed among the selected pages. Ties use initial retrieval order and then page identifier. Training examples for the head come only from fixed visual top-32 sets, ensuring the same head can be applied without fitting separate models to the policies being compared.

Reader prompt. The verbatim author-released instruction is: *I will provide you with a query and a candidate. Please evaluate whether the candidate meets the requirements of the query. If it does, respond with 'Yes'; if it doesn't, responed with 'No'.* The spelling is retained for exact prompt reproduction. The original question and candidate image follow the author template, and scores are next-token logit differences. Questions are checked against a 128-token limit. The prompt contains no appended native text.

B Coverage and Uncertainty

For positively judged pages G_q, acquired set C_q, and relevance grade g_{qp},

$$\mathrm{Recall@32}(q) = |G_q \cap C_q|/|G_q|, \tag{6}$$

$$\mathrm{All@32}(q) = \mathbf{1}[G_q \subseteq C_q], \tag{7}$$

$$\mathrm{nDCG}\ @10(q) = \frac{\sum_{i=1}^{10} g_{q,p_i}/\log_2(i+1)}{\sum_{i=1}^{10} g_{q,p_i^*}/\log_2(i+1)}. \tag{8}$$

Here p_i^* is the ideal ordering over the full corpus. Candidate-oracle nDCG uses the ideal ordering restricted to C_q in the numerator and the full-corpus denominator. Queries are scored individually and then averaged. This distinguishes a change in candidate availability from a change in ranking quality.

Table 7: Reader feedback minus each source control, in percentage points. Query-paired bootstrap with equal corpus weight and 10,000 draws. These descriptive intervals do not remove source model-selection effects or within-document dependence.

Comparison baseline	Recall difference [95% interval]	All difference [95% interval]
Visual top-32	$+3.86$ $[+3.01, +4.73]$	$+7.10$ $[+5.21, +9.05]$
Static hybrid	$+2.34$ $[+1.43, +3.24]$	$+5.69$ $[+3.80, +7.57]$
Uniform feedback	$+1.81$ $[+1.19, +2.43]$	$+3.35$ $[+1.99, +4.72]$

Table 8: Frozen reader feedback minus each HR control. Differences and 95% intervals are in score points; 5,000 paired resamples. Family resampling has only six connected groups. No multiplicity correction is applied.

Baseline	Metric	Difference	Query interval	Family interval
Visual top-32	nDCG	$+2.77$	$[+1.61, +4.00]$	$[+1.09, +6.28]$
Visual top-32	Recall	$+8.39$	$[+6.39, +10.39]$	$[+6.29, +16.32]$
Visual top-32	All	$+12.89$	$[+8.81, +16.98]$	$[+9.82, +22.76]$
Static hybrid	nDCG	$+0.13$	$[-0.86, +1.04]$	$[-2.07, +0.76]$
Static hybrid	Recall	$+1.22$	$[-0.67, +2.93]$	$[-0.88, +5.45]$
Static hybrid	All	$+3.46$	$[+0.00, +6.92]$	$[+0.00, +10.73]$
Uniform feedback	nDCG	$+1.01$	$[-0.01, +2.09]$	$[-0.09, +2.32]$
Uniform feedback	Recall	$+3.15$	$[+1.69, +4.62]$	$[+2.47, +5.18]$
Uniform feedback	All	$+5.97$	$[+3.14, +9.12]$	$[+0.95, +7.06]$

C Frozen HR Evaluation Details

The frozen target uses 318 English questions, 1,110 pages, and 14 source-document identifiers. Its complete corpus rankings are saved for all four methods before relevance acquisition. Each method is charged for 10,176 page judgments. Recorded decoder input lengths sum to 13,743,584 tokens for each method on this corpus, including visual, question, and prompt tokens; this numerical match is observed for HR, not assumed for arbitrary page sets. All primary methods have zero additional query-generation calls. The shared initial HR retrieval pass takes 106.6 seconds including this run's retrieval computation; cached extraction is not charged as a fresh parse.

For family intervals, source identifiers are normalized to lowercase with underscores replaced by spaces; stand-alone years are mapped to a common year symbol. Families joined by a query's positively judged source documents are merged into one connected component. This yields six groups for HR. Query resampling draws paired per-query differences; family resampling draws groups with replacement and divides the summed differences by the corresponding query count. Both use 5,000 draws and seed 2027. Neither scheme estimates uncertainty across a population of unseen domains.

The prefix analysis uses direct reader scores because the final head was trained for 32 inputs. At 16 pages every method is identical. At 24 pages static hybrid is slightly better in nDCG, while reader feedback has higher recall and all-page coverage. At 32 pages the raw-reader ordering remains essentially tied between those two methods. This supports the distinction between the quality of the acquired evidence set and the accuracy of the reader's final ordering; it does not establish a speed or compute advantage over static hybrid retrieval.

D Additional Module Implementations

Query facets. The text-only decomposition call uses Qwen2.5-VL-3B-Instruct with greedy decoding and at most 192 generated tokens. The prompt asks for up to four short information needs, preserving entities, dates, and units without answering the question. Output is parsed as either a list or numbered/bulleted lines. When no valid decomposition is obtained, the original question is retained as one facet. Across 956 source queries, the run yields 3,259 facets, including 132 fallback queries, 129,991 input tokens, and 59,567 output tokens. Batched generation takes 145 seconds on

Table 9: HR acquisition prefixes. The 24-page rows use eight entries from the already selected second batch; they are not separately tuned policies. Reader nDCG excludes the final score head.

Pages	Acquisition	Reader nDCG	Recall	All
16	Visual top-32	61.30	65.09	40.88
16	Static hybrid	61.30	65.09	40.88
16	Uniform feedback	61.30	65.09	40.88
16	Reader feedback (ours)	61.30	65.09	40.88
24	Visual top-32	62.46	70.49	45.28
24	Static hybrid	66.12	77.64	53.14
24	Uniform feedback	65.58	76.52	51.26
24	Reader feedback (ours)	65.86	78.62	54.40
32	Visual top-32	64.34	75.42	49.69
32	Static hybrid	67.03	82.60	59.12
32	Uniform feedback	65.89	80.66	56.60
32	Reader feedback (ours)	66.99	83.82	62.58

the experimental GPU; this is aggregate batched throughput, not single-query latency. The same small text encoder embeds facets.

Each facet proposes its top 64 dense pages. The expanded union adds these proposals to the original flat pool. Four new features are maximum and mean facet affinity and maximum and mean affinity weighted by each facet's missing-coverage factor. A 17-feature selector is fitted with the same tree settings. In the proposal-only ablation, the 13-feature policy scores the expanded pool. No selected candidate changes in that ablation.

Greedy facet coverage. Let $v_{pj} \in [0, 1]$ be page p's nonnegative similarity to facet j, and let $k_j = \max_{a \in S} c_a v_{aj}$ be initial observed affinity. Convert the learned selector score to quality $u_p = 61/(60 + r_h(p))$. This is a reciprocal rank, not a relevance probability. At each step, choose the unselected page maximizing

$$\frac{u_p}{16} + \frac{1}{J} \sum_{j=1}^{J} \max(u_p v_{pj} - k_j, 0), \tag{9}$$

then update $k_j \leftarrow \max(k_j, u_p v_{pj})$ and repeat until 16 pages are selected. This relevance-plus-coverage ablation gives the normalized terms equal weight without a weight sweep. Its uniform-feedback counterpart sets every c_a to one in both training features and initial facet coverage.

Signed feedback. Four added inputs use $1 - c_a$ for negative semantic support, positive minus negative support, mean signed confidence within a source document, and maximum negative confidence within a source document. The shuffle control permutes seed confidences within each query only when constructing these new inputs, preserving the original 13 features. Both variants use the original flat pool and the same 96-tree settings. The shuffle is a diagnostic ablation confined to source evaluation.

E Resource Accounting and Reproducibility

Each primary method reads 30,592 page–query pairs across the source corpora and 10,176 on HR: exactly 32 distinct pages per query with a 1,280-visual-token cap. Experiments reuse scores only for identical questions, pages, models, prompts, and preprocessing. Physical recomputation is therefore lower than the sum of per-method budgets. The primary method makes no extra generative call; its text encoder, index, CPU selector, and initial retriever remain additional costs.

Recorded artifacts include acquired sets, reader scores, full rankings, fitted trees, source-only selection, frozen-target predictions, and metric checks. Extraction and embedding caches predate these experiments; reading cached files does not measure their preparation cost. Reader inference uses one RTX 5090 with 32 GB memory, and feature construction and fitting use four CPU threads.