# WHEN FUSION HURTS: DIAGNOSING AND REPAIRING RECOVERABLE FAILURES IN HETEROGENEOUS RETRIEVAL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Heterogeneous retrieval systems combine channels such as lexical search, dense embeddings, visual retrieval, metadata search, and domain-specific indexes. Fusion is intended to preserve complementary evidence, but it can also suppress a relevant item that one specialist channel had already retrieved. We study this recoverable failure mode, which we call *fusion hurt*: at least one component ranks a relevant item within the target cutoff, but the fused ranking does not. Classical remedies such as tuned RRF, score calibration, quotas, and tree-based learning-to-rank repair some failures, but our analysis shows that they remain limited because they impose global or weakly query-adaptive aggregation policies over uncalibrated, source-specific evidence. We therefore use fusion hurt not only as an evaluation subset, but as supervision for a lightweight post-fusion listwise reranker. The reranker is trained over candidate-union rank, score, source, fusion, and disagreement features, with a query-level loss weight for fusion-hurt training cases; at inference time it uses no qrel-derived feature or fusion-hurt label. Across controlled and public heterogeneous retrieval settings, FH weighting improves the recoverable failure slice over the same plain reranker by +2.8 Recall@10 on VisRAG and +3.1 Recall@10 on MMDocIR, while BEIR text-hybrid validation shows a +3.1 Recall@10 and +5.7 nDCG@10 gain over raw hybrid overall. The experiments also expose where classical fusion remains insufficient and identify boundary cases where strong individual retrievers or tuned fusion already repair much of the failure.

## 1 INTRODUCTION

Modern retrieval systems often combine heterogeneous channels: lexical retrieval, dense embeddings, visual page retrievers, late-interaction models, metadata search, or domain-specific sources. Fusion is meant to preserve complementary evidence, but a fixed aggregation rule can also suppress a relevant item that one component already retrieved. We study this recoverable aggregation failure. Existing responses usually treat the problem as one of better average aggregation: tune source weights, normalize scores, reserve quotas for sources, or train a generic learning-to-rank model over retrieval features. These are necessary baselines, but they do not name the specific failure in which a minority channel is correct and the fusion policy demotes its evidence below the user-facing cutoff. That distinction is important because it suggests a different training signal: identify the recoverable failures induced by a chosen fusion policy, then teach the repair model to recognize their retrieval traces. For a query $q$, component rankings $R_s(q)$, fused ranking $F(q)$, relevant set $\mathcal{Y}_q$, and cutoff $k$, define

$$\mathrm{FH}(q) = \mathbb{1}\left[\min_{s,d\in\mathcal{Y}_q} \mathrm{rank}_{R_s(q)}(d) \le k \wedge \min_{d\in\mathcal{Y}_q} \mathrm{rank}_{F(q)}(d) > k\right]. \tag{1}$$

We call this event *fusion hurt*: at least one channel had a relevant item within the target cutoff, but fusion pushed all relevant items below it. The definition is deliberately tied to a specified fusion policy and cutoff; it asks whether that policy lost evidence already present in its inputs.

Our method turns this diagnostic into supervision. We train a lightweight post-fusion listwise reranker over the candidate union, using ordinary rank/score/source/disagreement features, and upweight training queries where the baseline fusion rule caused fusion hurt. At inference time, the



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.

reranker receives no fusion-hurt label and no qrel-derived feature; the weighting changes only the training objective. Thus the method is an adaptation layer for a retrieval configuration, not a new universal retriever.

The idea is channel-agnostic. Multimodal retrieval is one natural instance, but the same failure can arise whenever $N$ retrievers with different evidence views are fused. We evaluate across controlled document-page retrieval, public visual-document retrieval, image/text diagnostics, and text-hybrid BEIR-style retrieval. The main empirical question is not whether a reranker beats raw fusion in aggregate, but whether fusion-hurt supervision improves the intended failure slice beyond the same plain listwise reranker. Before introducing the reranker, we first analyze why standard fusion remedies are not enough: global weights, calibrated score sums, quotas, and conventional LTR each repair part of the problem, but none directly optimizes for query-local cases where a minority source is the only one carrying relevant evidence.

Our contributions are:

- We formalize fusion hurt as a recoverable failure of heterogeneous rank aggregation and use it to derive targeted supervision.
- We analyze where classical fusion and LTR baselines fail, including source-rescue slices and rescue/regression patterns.
- We introduce a lightweight post-fusion listwise reranker whose objective upweights queries where fusion suppressed a specialist hit, with no inference-cost increase over the same plain reranker.
- We establish when the signal helps across heterogeneous retrieval settings, and where strong single-source or tuned-fusion baselines already address most recoverable failures.

## 2 Related Work

Rank aggregation methods such as Borda-style voting, distance-based unsupervised aggregation, domain-specific expertise models, and RRF combine ranked lists without requiring calibrated scores (Klementiev et al., 2008; 2009; Cormack et al., 2009); recent analyses continue to study when fusion rules help or fail (Tyomkin & Kurland, 2024). These methods learn or impose a global aggregation policy, whereas we identify query-level cases where that policy suppresses a relevant item already recovered by a component and use those cases as training signal. Heterogeneous retrieval benchmarks and systems, including sparse/dense BEIR retrieval (Thakur et al., 2021), hybrid text embeddings (Chen et al., 2024), visual document retrievers and benchmarks (Yu et al., 2025; Faysse et al., 2025; Dong et al., 2025; Osmulski et al., 2025), and multimodal representation or fusion systems (Sun et al., 2025; Zeng et al., 2025; Ma et al., 2024; Jiang et al., 2025; Wang et al., 2024; Lin et al., 2025; Uzan et al., 2026; Samuel et al., 2025; Nguyen et al., 2025; Yang et al., 2023) improve components or global fusion policies. Generic rerankers such as RankZephyr, RankRAG, and RagVL improve candidate ordering through listwise or instruction-tuned content reranking (Pradeep et al., 2023; Yu et al., 2024; Chen et al., 2025). Our contribution is complementary: after heterogeneous first-stage retrieval has produced evidence, we ask whether fusion preserved specialist hits and whether its recoverable failures can supervise a lightweight repair model.

## 3 Problem and Challenges

Fusion hurt is not simply low recall. It is a more specific failure: a relevant item is present in the candidate evidence supplied to the fusion stage, but the aggregation policy demotes it below the user-facing cutoff. This distinction matters because the right intervention is different. If every retriever misses the relevant item, the system needs better representation, parsing, OCR, metadata, or indexing. If one retriever already found it, the system needs better aggregation or post-fusion repair.

Three practical challenges make this failure hard for classical fusion to eliminate. First, heterogeneous channels expose incomparable scores: a BM25 score, dense similarity, and visual late-interaction score do not have the same scale or reliability across queries. Second, global source weights are too coarse. They can learn that one channel is usually strong, but they cannot express

| Query | Rescue source | Source rank | Equal-fusion rank | Tuned WRRF rank | Plain CE | FH-CE |
|---|---|---|---|---|---|---|
| mmdocir:0004:000 | ColQwen2 image | 8 | 43 | 12 | 5/5 hits | 5/5 hits |
| mmdocir:0144:001 | BM25 caption | 6 | 11 | 27 | 0/5 hits | 0/5 hits |
| mmdocir:0119:002 | ColQwen2 image | 8 | 35 | 12 | 0/5 hits | 4/5 hits |

Table 1: Concrete MMDocIR fusion-hurt examples from the held-out test split. The relevant page is inside the target cutoff for one specialist source, but equal fusion ranks it outside the top 10. The rows illustrate three outcomes: generic CE repair, a hard case that remains missed, and a case where FH-weighting changes the learned decision on the same candidate evidence.

that a usually weaker channel is decisive for a particular query. Third, generic learned ranking can improve the candidate union while still allocating capacity away from the recoverable failure subset, because ordinary listwise training optimizes average ranking quality rather than the subset where fusion provably lost available evidence.

The empirical audits show these challenges concretely. On VisRAG (Yu et al., 2025), 283 of 426 fusion-hurt test queries are BM25-rescued and 167 are dense-text-rescued; score-scaled fusion repairs many BM25-rescued cases (0.8375 Recall@10) but is much weaker on dense-text-rescued cases (0.4790). The dense-image-rescued slice is a different failure mode: it contains 39 queries, and FH weighting does not improve over plain CE there, suggesting that rank/score/source features are not always enough without visual-content evidence. On MMDocIR, dev-tuned WRRF learns a heavy ColQwen2 weight and repairs much of the equal-fusion failure, but the FH-weighted reranker still improves the held-out FH slice over the same plain reranker. On ViDoRe, the opposite boundary appears: a strong ColPali branch and tuned ColPali-heavy fusion repair enough of the slice that explicit FH weighting is not beneficial. Table 1 gives three held-out MMDocIR examples that make these outcomes concrete. Here CE denotes the listwise cross-entropy reranker used in the controlled comparisons.

These observations motivate the method. The goal is not to replace first-stage retrievers or to claim that one learned reranker dominates every fusion policy. The goal is to turn a measurable aggregation failure into a targeted training signal, while preserving a clean distinction between training-time supervision and inference-time retrieval traces.

## 4 METHOD

We use *component retriever* and *source* interchangeably for each first-stage retrieval channel that contributes a ranked list to the fusion policy. For each query, we form $C_q$ as the union of top-ranked items from these component retrievers. For each candidate $d \in C_q$, features include source-presence indicators, component ranks, reciprocal-rank transforms, raw and per-query normalized scores when available, baseline-fusion score/rank, and disagreement features such as best rank, rank spread, and missing-source indicators. Some runs add simple embedding-interaction or query-shape features, but the reranker is not a cross-encoder and uses no content cross-attention.

Let $y_{q,d} \in \{0, 1\}$ indicate whether candidate $d$ is relevant for query $q$. The model scores candidates with a small MLP $s_\theta(q, d)$ and optimizes a listwise softmax objective over each query's candidate set:

$$L_{\mathrm{CE}}(q) = -\sum_{d \in C_q} \frac{y_{q,d}}{\sum_{d' \in C_q} y_{q,d'}} \log \frac{\exp(s_\theta(q,d))}{\sum_{d' \in C_q} \exp(s_\theta(q,d'))}. \tag{2}$$

Fusion-hurt-aware training uses a query weight $w_{\mathrm{FH}}(q)$ to multiply this loss:

$$w_{\mathrm{FH}}(q) = 1 + \alpha_{\mathrm{FH}} \mathbb{1}[\mathrm{FH}(q)], \qquad L(q) = w_{\mathrm{FH}}(q) L_{\mathrm{CE}}(q). \tag{3}$$

Here $w_{\mathrm{FH}}(q)$ is the fusion-hurt query weight and $\alpha_{\mathrm{FH}} \geq 0$ controls the extra weight assigned to fusion-hurt queries; the controlled VisRAG runs use $\alpha_{\mathrm{FH}} = 0.5$. Queries with no relevant item in the candidate union cannot supply a positive listwise signal and are excluded from supervised reranker updates. Figure 1 shows the training/inference separation.

**Feature construction.** The reranker is intentionally restricted to features available after first-stage retrieval. For each source we encode whether the candidate appeared in that source, its rank when

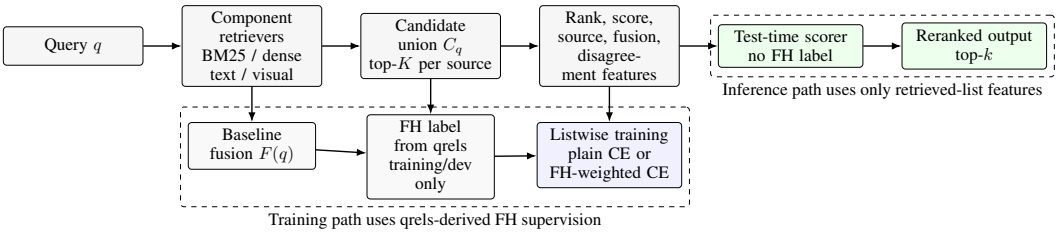


Figure 1: Fusion-hurt-aware post-fusion reranking architecture. The FH label is derived from qrels and the baseline fused ranking on training/development data only. At test time, the reranker receives the same candidate-union features but no FH label.

| Benchmark | Channels | Eval queries | FH queries |
|---|---|---|---|
| VisRAG | BM25, dense text, dense image | 2,300 | 426 |
| MMDocIR | BM25 captions, dense captions, ColQwen2 image | 331 | 26 |
| ViDoRe | BM25, ColPali | 223 | 19 |
| BEIR | BM25, dense, raw hybrid | 12,937 | 218 |
| AToMiC | BM25 metadata, CLIP text-to-image | 152 | 11 |

Table 2: Evaluation settings and fusion-hurt subset sizes. AToMiC is used only to audit whether fusion hurt appears in an image/text setting.

present, reciprocal rank, raw score, and per-query normalized score when scores are available. Score normalization is computed within each query and source over the retrieved candidates: min–max features use $(x - \min x)/(\max x - \min x)$ with zero for missing candidates or zero-range score lists, and runs with standardized score features use the corresponding within-query/source z-score. We also include aggregate signals: best component rank, mean observed rank, rank spread across sources, number of sources that retrieved the candidate, baseline fusion score, baseline fusion rank, and missing-source indicators. These features let the model learn patterns such as "a candidate is highly ranked by one specialist but absent elsewhere" or "a candidate has moderate support across many channels." They do not include qrel-derived indicators, component oracle labels, or the fusion-hurt label at inference time. This distinction matters because fusion hurt is easy to identify after seeing qrels, but the deployed model must infer when to trust specialist evidence from retrieval traces alone.

**Training protocol.** The listwise objective is computed over the candidate union for each query, so all candidates for the same query compete in one softmax. When a query has multiple relevant candidates in the pool, the positive mass in Equation 2 is distributed uniformly over those positives. For the plain CE baseline and the FH-weighted model, architecture, optimizer, candidate pools, features, and evaluation pipeline are held fixed. The FH-weighted model changes only the scalar query weight in Equation 3. This makes the plain CE comparison stricter than comparing against raw fusion: it asks whether the proposed label changes the learning signal after accounting for the generic benefit of supervised listwise reranking. Tree-LTR baselines use the same retrieval-derived feature families where applicable, but optimize standard LambdaRank or ranking losses rather than the listwise neural objective.

## 5 Experimental Setup

Table 2 summarizes the five heterogeneous retrieval settings. VisRAG is the controlled benchmark with fixed component outputs and full ablations; MMDocIR and ViDoRe provide public visual-document validation; BEIR tests the channel-agnostic claim in text-hybrid retrieval; and AToMiC is diagnostic only because its FH slice is too small for supervised reranker conclusions.

Baselines include equal and dev-tuned WRRF, score-scaled and quota fusion, conventional LightGBM/XGBoost LTR controls, plain listwise CE reranking, and FH-weighted CE reranking. Dev-tuned WRRF selects source weights on development splits using all-query Recall@10 and then evaluates once on held-out test splits. We report Recall@10 and nDCG@10 overall and on FH slices.

| Benchmark | Method | All R@10 | All nDCG@10 | FH R@10 | FH nDCG@10 |
|---|---|---|---|---|---|
| VisRAG | Hybrid | 0.4387 | 0.2209 | 0.0000 | 0.0000 |
| VisRAG | Plain CE | 0.6370 | **0.3659** | 0.8394 | 0.4082 |
| VisRAG | FH-CE | **0.6375** | 0.3646 | **0.8676** | **0.4292** |
| MMDocIR | Hybrid | 0.7221 | 0.6014 | 0.0000 | 0.0000 |
| MMDocIR | Plain CE | **0.7867** | 0.6848 | 0.8462 | 0.4268 |
| MMDocIR | FH-CE | 0.7855 | **0.6869** | **0.8769** | **0.5014** |
| ViDoRe | Hybrid | 0.8072 | 0.6373 | 0.0000 | 0.0000 |
| ViDoRe | Plain CE | **0.8583** | **0.6944** | **0.6947** | 0.3017 |
| ViDoRe | FH-CE | 0.8547 | 0.6916 | 0.6737 | **0.3145** |
| BEIR | Hybrid | 0.8388 | 0.7858 | 0.0000 | 0.0000 |
| BEIR | FH-CE | **0.8693** | **0.8427** | **0.7104** | **0.4831** |

Table 3: Cross-benchmark Recall@10 and rank-sensitive nDCG@10. Hybrid is equal WRRF for VisRAG, MMDocIR, and ViDoRe, and raw hybrid for BEIR. **Bold** marks the best non-oracle value within a benchmark block and underline marks the second-best value when applicable. The hybrid FH values are zero by definition because the FH slice contains queries where baseline fusion misses within the target cutoff. For BEIR, FH-CE denotes the no-OCI listwise CE repair model used in the text-hybrid validation; dense/prefix proxy and OCI reranker rows are omitted to keep the comparison focused on the boost over hybrid. For MMDocIR and ViDoRe, nDCG@10 is computed from the archived best relevant rank, so it is a rank-sensitive best-hit measure rather than full multi-positive nDCG.

For VisRAG and BEIR, nDCG@10 follows the standard binary/graded relevance evaluation available for those runs. For MMDocIR and ViDoRe reranker summaries, the archived per-query artifacts expose the best relevant rank rather than full ranked relevance lists, so the reported nDCG@10 is the corresponding best-relevant discounted gain; it is rank-sensitive but should be read as a companion to Recall@10 rather than a full multi-positive nDCG audit. Main paired comparisons use bootstrap confidence intervals, paired sign-flip randomization, and exact McNemar tests on query-level hits.

The comparison ladder separates three questions that are often conflated in aggregate retrieval results: whether stronger fusion can repair the failure, whether generic supervised LTR explains the effect, and whether FH supervision improves the intended slice beyond the same plain listwise reranker. For the datasets with both plain CE and FH-CE rows, only the last comparison isolates the proposed supervision signal; Appendix A gives the full baseline capability summary.

## 6 Results

The results are organized around three trends. First, FH-aware training improves the intended recoverable-failure slice in the strongest controlled setting and in MMDocIR. Second, the gains are not explained by classical fusion or tree-LTR controls, which repair some source-specific failures but leave substantial recoverable headroom. Third, public benchmarks expose useful boundary conditions: BEIR supports the channel-agnostic framing at larger scale, while ViDoRe shows that explicit FH weighting is not always better than a strong plain reranker or tuned fusion.

### 6.1 Trend 1: Cross-Benchmark Impact

Table 3 is the main result table. For VisRAG, MMDocIR, and ViDoRe, it reports the same comparison ladder: baseline hybrid fusion, the same candidate union reranked with a plain listwise CE model, and the same reranker trained with fusion-hurt weighting. The key comparison for those datasets is therefore plain CE versus FH-CE: FH weighting asks whether explicit failure-slice supervision improves the intended slice beyond generic listwise reranking. BEIR is reported separately within the same table as large-scale text-hybrid repair evidence: Hybrid versus the no-OCI CE repair model, labeled FH-CE for consistency with the failure-slice repair framing. VisRAG is the controlled ablation, MMDocIR repeats the same target-slice direction in a public visual-document setting, BEIR shows the effect at larger text-hybrid scale, and ViDoRe is a boundary case where plain CE is better than FH-weighted CE by Recall@10. AToMiC remains diagnostic only because its usable fusion-hurt subset contains 11 queries.

| Method | Seeds | All R@10 | FH R@10 | Non-FH R@10 |
|---|---|---|---|---|
| Best component on FH | oracle | – | 1.0000 | – |
| Tuned WRRF | 1 | 0.4857 | 0.4085 | 0.5032 |
| Score-scaled fusion | 1 | 0.4965 | 0.6479 | 0.4621 |
| Quota fusion | 1 | 0.4878 | 0.6221 | 0.4573 |
| XGBoost rank-only LTR | 5 | 0.4983 | 0.5249 | 0.4923 |
| Plain listwise MLP | 5 | 0.6370 | 0.8394 | 0.5910 |
| FH-weighted CE MLP | 5 | 0.6375 | 0.8676 | 0.5852 |

Table 4: Controlled VisRAG Recall@10 ablation. The FH subset contains 426 test queries; the non-FH subset contains 1,874. The oracle row is defined only on the FH subset and represents the recoverable ceiling implied by the FH definition. Appendix C reports the full table, including LightGBM controls, top-$k$ variants, latency, and source-rescue slices.

### 6.2 Trend 2: Classical Baselines Leave Recoverable Headroom

On VisRAG, the controlled comparison is FH-weighted CE versus the same listwise MLP trained with plain CE. FH-weighted CE improves fusion-hurt Recall@10 by +0.0282, with 95% paired bootstrap CI [+0.0192, +0.0376], paired sign-flip $p = 0.0001$, and exact McNemar $p = 1.12 \times 10^{-9}$. All-query Recall@10 changes by only +0.0004. The strongest conventional tree-LTR control, XGBoost rank-only, reaches 0.5249 FH Recall@10, far below the FH-weighted listwise model at 0.8676; the paired FH comparison gives CI [+0.3202, +0.3648] and exact McNemar $p = 1.62 \times 10^{-175}$. Thus the effect is not explained by generic tree-based LTR over rank/score/source features.

Table 4 gives the controlled VisRAG ablation. The non-neural fusion baselines are useful and cheap, but none reaches the listwise rerankers on the recoverable failure subset. Score-scaled fusion is the strongest non-neural fusion baseline on fusion hurt at 0.6479 Recall@10. Plain CE reaches 0.8394, and FH-weighted CE reaches 0.8676 with the same inference architecture. The non-FH drop from 0.5910 to 0.5852 shows the tradeoff directly; the method reallocates ranking capacity toward the subset where fusion provably suppressed an available relevant item.

The public benchmarks are mixed because they stress different parts of the claim. MMDocIR repeats the target-slice effect: FH-weighted CE improves the held-out FH slice over plain CE from 0.8462 to 0.8769 Recall@10 and from 0.4268 to 0.5014 nDCG@10, while all-query Recall@10 remains flat. ViDoRe confirms that fusion hurt is present but is a boundary case for FH weighting: plain CE is slightly stronger than FH-weighted CE on the 19-query FH test slice by Recall@10, although FH nDCG@10 rises from 0.3017 to 0.3145. BEIR gives the largest public validation of the channel-agnostic claim: on lexical, dense, and hybrid text retrieval channels, FH-CE improves over Hybrid from 0.8388 to 0.8693 Recall@10 and from 0.7858 to 0.8427 nDCG@10 overall; on the 218-query FH slice it reaches 0.7104 Recall@10 and 0.4831 nDCG@10 from a zero Hybrid baseline. Full per-benchmark tables, source-rescue analysis, rescue/regression counts, robustness checks, and latency details are in Appendix C.

### 6.3 Trend 3: Where Repair Works and Where It Does Not

The oracle row in Table 4 emphasizes the recoverable nature of the subset: every fusion-hurt query has at least one component retriever that already placed a relevant item within the target cutoff. The remaining question is how much of that recoverable headroom can be closed by practical aggregation or reranking methods without harming the rest of the distribution. The controlled comparison shows a specific pattern. Score-scaled fusion and quota fusion recover many failures, but remain substantially below a learned listwise reranker. Conventional tree-LTR baselines improve only modestly, despite seeing rank, score, and source features. The FH-weighted MLP then changes the error allocation of the same reranker rather than simply improving every slice.

Table 5 breaks the VisRAG fusion-hurt subset down by which component retriever could have rescued the query. The slices overlap because multiple component retrievers can retrieve a relevant item within the cutoff for the same query. Most recoverable failures are lexical- or dense-text-rescued, but the dense-image-rescued slice is harder for the learned reranker. Fusion-hurt weighting improves over plain CE on the BM25-rescued, dense-text-rescued, and multi-retriever-rescued slices, while

| FH source-rescue slice | Queries | Score-scaled fusion | Plain CE | FH-weighted CE |
|---|---|---|---|---|
| BM25 rescued | 283 | 0.8375 | 0.8989 | 0.9180 |
| Dense-text rescued | 167 | 0.4790 | 0.8311 | 0.8754 |
| Dense-image rescued | 39 | 0.1795 | 0.6462 | 0.6462 |
| Multi-retriever rescued | 62 | 0.7742 | 0.9645 | 0.9774 |

Table 5: VisRAG fusion-hurt taxonomy by rescuing component retriever. Rows overlap when more than one component retrieves a relevant item within the cutoff. Values are Recall@10 on the corresponding FH sub-slice.

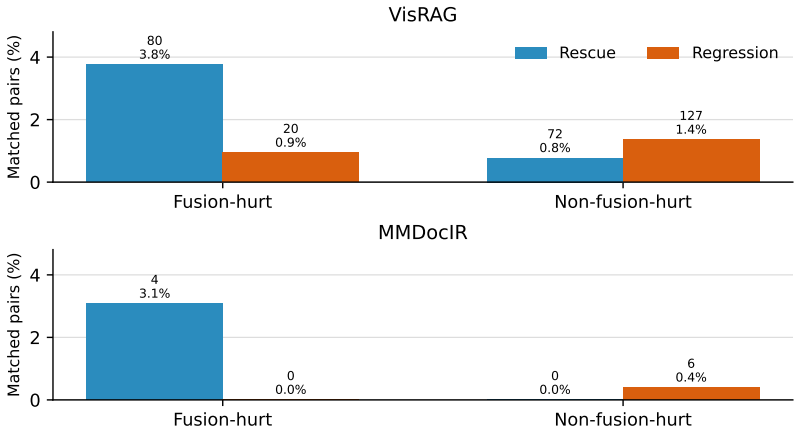


Figure 2: FH-weighted CE shifted errors toward the intended subset on both VisRAG and MMDocIR. On VisRAG fusion-hurt queries it produced 80 rescues and 20 regressions over five seeds, or 3.8% versus 0.9% of matched seed-query pairs; on MMDocIR the corresponding values were 4 rescues and 0 regressions, or 3.1% versus 0.0%. On non-fusion-hurt queries the tradeoff was smaller and reversed: VisRAG showed 72 rescues and 127 regressions, or 0.8% versus 1.4%, and MMDocIR showed 0 rescues and 6 regressions, or 0.0% versus 0.4%. A rescue is a query-seed pair where FH-weighted CE hit at Recall@10 and plain CE missed; a regression is the reverse.

the dense-image-rescued slice is unchanged in this controlled run. This suggests that the current feature-only reranker is best at detecting when rank disagreement reflects text-channel evidence that fusion undervalued; additional visual-content features may be needed to close the image-rescued gap.

Figure 2 shows the corresponding rescue/regression pattern over matched query-seed pairs, normalized within each subset, for VisRAG and MMDocIR. On the VisRAG fusion-hurt subset, FH-weighted CE rescues 80 seed-query pairs that plain CE misses and regresses 20 that plain CE hits, corresponding to 3.8% rescues and 0.9% regressions. On non-fusion-hurt queries the direction reverses mildly, with 72 rescues and 127 regressions, corresponding to 0.8% rescues and 1.4% regressions. MMDocIR shows the same directional pattern: 4 rescues versus 0 regressions on fusion-hurt seed-query pairs, or 3.1% versus 0.0%, compared with 0.0% rescues and 0.4% regressions on non-fusion-hurt pairs. Thus, the flat aggregate metric hides a targeted reliability change: the training signal shifts errors away from the recoverable fusion-failure subset.

**Robustness and cost.** The VisRAG target-subset effect persists across candidate-pool depths and cutoffs. Across top-50, top-100, and top-200 candidate pools, FH-weighted CE improves fusion-hurt Recall@10 over plain CE by 2.4–3.2 points with essentially unchanged overall Recall@10. When the fusion-hurt definition and evaluation cutoff are varied, the target-subset gain remains positive at Recall@5, Recall@10, and Recall@20. The reranker is more expensive than rank-only fusion but remains lightweight in the controlled CPU harness. FH weighting has the same inference cost as plain CE because it changes only training. Appendix C reports the CPU cost table.

| Method | All R@10 | FH R@10 | Non-FH R@10 |
|---|---|---|---|
| Equal WRRF | 0.7221 | 0.0000 | 0.7836 |
| ColQwen2 image only | 0.7795 | 0.8846 | 0.7705 |
| Dev-tuned WRRF | 0.7795 | 0.7692 | 0.7803 |
| Plain CE reranker | 0.7867 | 0.8462 | 0.7816 |
| FH-weighted CE reranker | 0.7855 | 0.8769 | 0.7777 |

Table 6: MMDocIR internal test Recall@10 over 331 queries and five seeds. The held-out FH subset contains 26 queries.

| Method | All R@10 | FH R@10 | Non-FH R@10 |
|---|---|---|---|
| BM25-only | 0.6996 | 0.3684 | 0.7304 |
| ColPali-only | 0.8520 | 0.6842 | 0.8676 |
| Equal WRRF | 0.8072 | 0.0000 | 0.8824 |
| Dev-tuned WRRF | 0.8475 | 0.6316 | 0.8676 |
| Plain CE reranker | 0.8583 | 0.6947 | 0.8735 |
| FH-weighted CE reranker | 0.8547 | 0.6737 | 0.8716 |

Table 7: Pooled ViDoRe internal test Recall@10 over 223 queries. The FH subset contains 19 queries, so public supervised/classical comparisons should be interpreted cautiously.

### 6.4 Public Visual-Document Validation

MMDocIR tests the same learning signal on a public page-level benchmark when the retrieval setup matches VisRAG. Across all 1,658 queries, equal WRRF over BM25 captions, dense caption retrieval, and ColQwen2 image retrieval induces 150 fusion-hurt queries (9.0%). The component behavior is heterogeneous: ColQwen2 is the strongest individual source at 0.7666 Recall@10, BM25-caption reaches 0.6634, and dense-caption reaches 0.5452. Equal WRRF improves over the weaker text branch but still suppresses specialist hits, reaching 0.7256 overall and 0.0000 on the fusion-hurt subset by construction.

Table 6 reports the held-out reranking result over five seeds. We first tuned WRRF weights on the internal development split using all-query Recall@10. The selected model assigns weights 0.05, 0.05, and 0.90 to BM25-caption, dense-caption, and ColQwen2-image retrieval. This tuned classical baseline is strong: it improves test Recall@10 over equal WRRF from 0.7221 to 0.7795, with paired sign-flip $p = 0.0001$, and repairs 0.7692 of the fusion-hurt subset. The result is an important control: much of the equal-fusion failure can be repaired by learning to trust the strongest visual retriever more heavily.

The learned rerankers remain competitive with this tuned fusion control. Plain CE reaches 0.7867 overall Recall@10 and 0.8462 on fusion-hurt. Fusion-hurt weighting further improves the fusion-hurt subset to 0.8769, a +0.0308 gain over plain CE, and improves FH nDCG@10 from 0.4268 to 0.5014. The seed-query paired sign-flip test gives $p = 0.0401$ and exact McNemar gives $p = 0.0391$; at query level, the confidence interval crosses zero because the held-out fusion-hurt subset contains 26 queries. Overall Recall@10 changes only from 0.7867 to 0.7855, while non-fusion-hurt Recall@10 drops from 0.7816 to 0.7777. Thus MMDocIR supports public targeted repair and a realistic boundary condition: fusion-hurt-aware learning improves the intended subset relative to the same reranker, but it does not dominate every tuned classical or single-source visual baseline.

ViDoRe provides a second public visual-document check. Across seven English or verified-English ViDoRe datasets, fusion hurt appears in 91 of 1,114 queries (8.2%). The failure is bidirectional: ColPali is the best rescuing source for 67 fusion-hurt queries and BM25 for 24. Thus, fusion hurt is not specific to VisRAG or to a single component being globally weak.

Table 7 shows that equal WRRF performs well overall but, by definition, misses the held-out fusion-hurt subset. Dev-tuned WRRF selects BM25 weight 0.20 and ColPali weight 0.80 using pooled internal dev all-query Recall@10. It repairs much of the equal-WRRF failure subset, but ColPali-only is also very strong. Plain CE reranking is the best point estimate, while FH-weighted CE is slightly worse on the FH slice. The ViDoRe conclusion is therefore narrower than the VisRAG

conclusion: ViDoRe validates that fusion hurt exists publicly and that a lightweight repair layer is competitive with strong ColPali-heavy fusion, but it does not establish a public gain from fusion-hurt weighting.

### 6.5 BEIR: Large Text-Hybrid Validation

The BEIR text-hybrid run validates the same failure-repair pattern in a larger heterogeneous-channel setting. Across 12,937 development queries from FEVER, HotpotQA, FiQA, and NFCorpus, the raw-hybrid ranking induces 218 fusion-hurt queries. Compared with Hybrid, FH-CE improves all-query Recall@10 from 0.8388 to 0.8693 and all-query nDCG@10 from 0.7858 to 0.8427. On the FH slice, Hybrid is 0.0000 by definition, while FH-CE reaches 0.7104 Recall@10 and 0.4831 nDCG@10. The result shows that recoverable fusion failures are not tied to visual retrieval channels: the same target-slice repair pattern appears over lexical, dense, and hybrid text retrieval channels. BEIR is therefore important for the framing of the paper: it is not a multimodal benchmark in this experiment, but it is a heterogeneous retrieval setting where sparse lexical evidence, dense semantic evidence, and a raw hybrid ranking disagree about which candidates should be promoted. Appendix C reports the BEIR metric table.

## 7 Conclusions

The results support a diagnostic-and-repair view of fusion: fixed fusion rules can be strong overall while hiding a recoverable failure subset, and a learned reranker over the candidate union can rebalance heterogeneous first-stage evidence without changing retrievers. The controlled VisRAG ablation tests whether FH supervision changes what the reranker learns beyond generic listwise training; MMDocIR shows the same target-slice direction publicly, and BEIR shows larger-scale raw-hybrid repair in text-hybrid retrieval. ViDoRe and AToMiC narrow the claim: when one source is already dominant or the FH subset is very small, plain reranking, tuned fusion, or a single strong component can repair much of the failure. Operationally, fusion hurt should be treated as an offline monitoring and training signal for deciding whether to tune fusion, add a repair reranker, or improve the missing component.

Fusion hurt identifies cases where a component found a relevant item but fusion suppressed it below the target cutoff. We turned that diagnostic into supervision for a lightweight post-fusion reranker across heterogeneous retrieval settings where this failure mode can be measured, audited, and often repaired without changing first-stage retrievers.

**Scope and future work.** Fusion hurt is relative to a fusion rule and cutoff and should be recomputed when aggregation changes. The method applies when relevant candidates are already in the candidate union; first-stage misses require stronger retrieval, parsing, OCR, metadata, or indexing. Future work should tune $\alpha_{\mathrm{FH}}$ adaptively, add richer visual/content features for image-rescued failures, and evaluate larger public FH slices; Appendix B gives details.

**Reproducibility statement.** All runs use fixed splits, fixed component outputs, deterministic FH labels, and five reranker seeds unless stated otherwise; supplementary materials include configs, commands, splits, per-query hits, and scripts.

**Ethics statement.** No new user dataset is introduced; main risks are evaluation bias and overclaiming, mitigated through subset sizes, paired tests, limitations, and boundary-case reporting.

## AI Use Statement

AI-assisted tools (OpenAI Codex) were used to revise author-written text and to assist with implementing the authors' methods in code and experiment scripts. The research ideas, methodology, experimental design, and analysis were developed by the authors. AI-assisted code and textual modifications were reviewed and verified by the authors. The authors take full responsibility for the final paper, implementation, and reported results.

## REFERENCES


Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024. URL https://arxiv.org/abs/2402.03216.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Xuhui Jiang, and Jian Guo. VLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8140–8158. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.432. URL https://aclanthology.org/2025.findings-emnlp.432/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1576. URL https://aclanthology.org/2025.emnlp-main.1576/.

Manuel Faysse, Hugo 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.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025. URL https://mlanthology.org/iclr/2025/jiang2025iclr-vlm2vec/.

Alexandre Klementiev, Dan Roth, and Kevin Small. Unsupervised rank aggregation with distance-based models. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 472–479. Association for Computing Machinery, 2008. doi: 10.1145/1390156.1390216.

Alexandre Klementiev, Dan Roth, Kevin Small, and Ivan Titov. Unsupervised rank aggregation with domain-specific expertise. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp. 1101–1106, 2009.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6d5d6afa9957cfc9142ba60e78a467e9-Abstract-Conference.html.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.373. URL https://aclanthology.org/2024.emnlp-main.373/.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1568. URL https://aclanthology.org/2025.emnlp-main.1568/.

Radek Osmulski, Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. MIRACL-VISION: A large, multilingual, visual document retrieval benchmark. *arXiv preprint arXiv:2505.11651*, 2025. URL https://arxiv.org/abs/2505.11651.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*, 2023. URL https://arxiv.org/abs/2312.02724.

Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Tanner Spendlove, Seun Eisape, Arun Reddy, Alexander Martin, Andrew Yates, Matthew Wiesner, Kenton Murray, and Reno Kriz. MMMORRF: Multimodal multilingual modularized reciprocal rank fusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

Hao Sun, Yingyan Hou, Jiayan Guo, Bo Wang, Chunyu Yang, Jinsong Ni, and Yan Zhang. Unveil: Unified visual-textual integration and distillation for multi-modal document retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23935–23945. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1166. URL https://aclanthology.org/2025.acl-long.1166/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Daniel Tyomkin and Oren Kurland. Analyzing fusion methods using the condorcet rule. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Omri Uzan, Asaf Yehudai, Roi Pony, Eyal Shnarch, and Ariel Gera. Guided query refinement: Multimodal hybrid retrieval with test-time optimization. In *International Conference on Learning Representations*, 2026.

Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. Unified embeddings for multimodal retrieval via frozen LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1537–1547. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.105. URL https://aclanthology.org/2024.findings-eacl.105/.

Xinyu Yang et al. AToMiC: An image/text retrieval test collection to support multimedia content creation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

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://mlanthology.org/iclr/2025/yu2025iclr-visrag/.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.02485.

Delong Zeng, Yuexiang Xie, Yaliang Li, and Ying Shen. Enhancing multimodal retrieval via complementary information extraction and alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22092–22105. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1073. URL https://aclanthology.org/2025.acl-long.1073/.

## A Experimental Setup Details

**Benchmarks and splits.** VisRAG is the controlled benchmark because all component outputs, qrels, fusion labels, and learned reranker features are fixed before the ablation. The split contains 16,100 train, 4,600 development, and 2,300 test queries; 426 test queries are fusion-hurt cases under equal WRRF at Recall@10. The component retrievers are BM25, dense text retrieval, and dense

image retrieval. MMDocIR is used as public page-level validation with BM25 over released VLM captions, dense caption retrieval, and ColQwen2 image retrieval. The full MMDocIR run contains 1,658 queries and 150 equal-WRRF fusion-hurt queries; the held-out internal test split used for reranker reporting contains 331 queries and 26 FH queries. ViDoRe pools seven English or verified-English ViDoRe V3 datasets using BM25 and ColPali. The pooled internal split contains 223 test queries and 19 FH test queries. BEIR uses FEVER, HotpotQA, FiQA, and NFCorpus development queries with BM25, dense retrieval, and raw hybrid rankings; the validation set contains 12,937 queries and 218 FH queries. AToMiC is a diagnostic-only image/text setting with BM25 metadata, CLIP text-to-image retrieval, and equal WRRF over 152 queries. It contains 11 FH queries, so we do not draw supervised reranker conclusions from it.

**Fusion and reranking baselines.** Equal WRRF is the default fusion policy for VisRAG, MMDocIR, ViDoRe, and AToMiC. Dev-tuned WRRF searches source weights on development splits using all-query Recall@10 and evaluates once on held-out test. Score-scaled fusion min-max normalizes component scores per query before combining them. Quota fusion allocates top-ranked output slots across sources to preserve specialist candidates. The conventional LTR controls use LightGBM LambdaRank and XGBoost rankers over rank-only or rank/score/source features. The neural reranker is a small MLP trained over the same post-retrieval candidate union. Plain CE and FH-weighted CE have the same inference-time features and architecture; the only difference is the query-level loss weight in Equation 3.

| Method family | Learns? | Uses scores? | Uses source identity? | FH supervision? |
|---|---|---|---|---|
| Equal WRRF / Borda | – | – | ✓ | – |
| Dev-tuned WRRF | ✓ | – | ✓ | – |
| Score-scaled fusion | – | ✓ | ✓ | – |
| Quota fusion | – | optional | ✓ | – |
| LightGBM / XGBoost LTR | ✓ | optional | optional | – |
| Plain listwise MLP | ✓ | ✓ | ✓ | – |
| FH-weighted listwise MLP | ✓ | ✓ | ✓ | ✓ |

Table 8: Baseline capability summary. The FH-weighted model is the only method that uses qrels-derived fusion-hurt labels, and it uses them only during training.

**Evaluation and inference constraints.** All FH labels are derived from the specified baseline fusion policy and qrels on the corresponding training/development data. No FH label, qrel-derived indicator, or oracle component-hit feature is available to the reranker at inference time. We report Recall@10 for all datasets because the failure definition is cutoff-based. We also report nDCG@10 to capture whether repaired queries move relevant items earlier inside the top ten. For MMDocIR and ViDoRe reranker summaries, nDCG@10 is reconstructed from the archived best relevant rank because the retained per-query artifacts do not include full ranked relevance lists; this is exact for single-positive queries and serves as a rank-sensitive best-hit proxy for multi-positive queries. Paired tests are computed over query-level or seed-query-level hit indicators as stated in the main text.

**Concrete MMDocIR examples.** Table 1 uses the MMDocIR query identifiers exported by the candidate-generation pipeline. The three rows correspond respectively to relevant pages 9, 0, and 38 in documents `52b3137455e7ca4df65021a200aef724`, `D18-1334`, and `disciplined-agile-business-analysis-160218012713_95`. The source-rank column is the rank assigned by the rescuing component retriever, equal-fusion rank is the baseline WRRF rank, tuned WRRF rank is the development-selected WRRF rank, and the CE columns count hits across the five reranker seeds.

**Aggregate retrieval-trace features.** For candidate $d$ and sources $s \in S$, let $r_s(d)$ be the source rank when present and a large sentinel rank when absent. The aggregate feature set includes $\min_s r_s(d)$, the mean of observed source ranks, the spread between worst and best observed ranks, the number of sources retrieving $d$, the baseline fusion score and rank, and one missingness indicator per source. These features summarize cross-source agreement and disagreement without exposing qrel-derived labels or the fusion-hurt indicator at inference time.

## B Additional Scope Details

**Candidate-pool coverage.** The approach assumes that at least one relevant item is present in the candidate union. This is appropriate for diagnosing aggregation failures, but it does not address first-stage recall failures. If all component retrievers miss the relevant item, neither plain CE nor FH-weighted CE can recover it. In that regime, improving encoders, document parsing, OCR, image representation, metadata extraction, or indexing policy is the right intervention. The paper's oracle rows and FH definition are intended to make this boundary explicit: the recoverable ceiling is conditional on component recall.

**Small public FH slices.** The public visual-document benchmarks are valuable because they test whether the phenomenon recurs outside the controlled VisRAG setup, but their supervised FH test slices are small. MMDocIR has 26 held-out FH queries and ViDoRe has 19. Those numbers are enough to expose directionality and boundary conditions, but not enough to support broad claims about every visual-document domain. This is why the paper treats VisRAG as the controlled ablation, MMDocIR as public targeted validation, ViDoRe as a boundary case, and BEIR as larger heterogeneous-channel evidence. The evidence bundle is intentionally multi-part rather than relying on a single public multimodal subset.

**Model class and feature limits.** The reranker is lightweight and feature-based. That choice keeps the method cheap and isolates the effect of fusion-hurt supervision, but it also limits what the model can learn. For example, the VisRAG dense-image-rescued slice does not improve under FH weighting in the current controlled run. A stronger model with visual-content features or a cross-encoder over query and page content might repair more of that slice, but it would also change the cost profile and make it harder to attribute gains specifically to the fusion-hurt objective. The current design should therefore be read as a conservative repair layer rather than an upper bound on what fusion-hurt-aware training could achieve.

## C Detailed Results

### C.1 VisRAG Controlled Ablation

Table 9 gives the full VisRAG comparison. The component oracle row is defined only on the FH subset: by definition, every FH query has at least one component retriever with a relevant item within the target cutoff. It is therefore a recoverable ceiling, not a deployable method.

| Method | Seeds | All R@10 | FH R@10 | Non-FH R@10 | CPU latency |
|---|---|---|---|---|---|
| Best component on FH | oracle | – | 1.0000 | – | – |
| Tuned WRRF | 1 | 0.4857 | 0.4085 | 0.5032 | 0.20 ms |
| Score-scaled fusion | 1 | 0.4965 | 0.6479 | 0.4621 | 0.36 ms |
| Quota fusion | 1 | 0.4878 | 0.6221 | 0.4573 | 0.37 ms |
| LightGBM LambdaRank, rank-only | 5 | 0.4958 | 0.5014 | 0.4946 | not profiled |
| XGBoost ranker, rank-only | 5 | 0.4983 | 0.5249 | 0.4923 | not profiled |
| LightGBM LambdaRank, rank/score/source | 5 | 0.4966 | 0.5080 | 0.4940 | not profiled |
| XGBoost ranker, rank/score/source | 5 | 0.4983 | 0.5249 | 0.4923 | not profiled |
| Plain listwise MLP | 5 | 0.6370 | 0.8394 | 0.5910 | 6.57 ms |
| Top-$k$ only MLP | 5 | 0.6352 | 0.8376 | 0.5892 | 6.57 ms$^\dagger$ |
| FH-weighted CE MLP | 5 | 0.6375 | 0.8676 | 0.5852 | 6.55 ms |
| FH-weighted + top-$k$ MLP | 5 | 0.6364 | 0.8700 | 0.5834 | 6.55 ms$^\dagger$ |

Table 9: VisRAG Recall@10. FH denotes the fusion-hurt test subset of 426 queries; the non-FH subset has 1,874 queries. Latency is CPU mean end-to-end cost for learned rerankers and fusion-only cost for fusion methods. $^\dagger$Top-$k$ variants use the same inference architecture as the profiled CE variant; only the training loss differs.

The strongest classical fusion baseline on the VisRAG FH subset is score-scaled fusion at 0.6479 Recall@10. The strongest conventional tree-LTR baseline is XGBoost rank-only at 0.5249. The

plain listwise MLP already improves substantially over both, reaching 0.8394 on the FH subset. FH-weighted CE improves that same reranker to 0.8676, while all-query Recall@10 remains effectively unchanged. This is the cleanest controlled evidence that the fusion-hurt label changes the learned repair behavior beyond generic listwise reranking.

Table 10 breaks the FH subset down by which component retriever could have rescued the query. Rows overlap when multiple components retrieve a relevant item within the cutoff. Most recoverable failures are lexical- or dense-text-rescued, but dense-image-rescued failures are harder for the learned reranker.

| FH source-rescue slice | Queries | Score-scaled fusion | Plain CE | FH-weighted CE |
|---|---|---|---|---|
| BM25 rescued | 283 | 0.8375 | 0.8989 | 0.9180 |
| Dense-text rescued | 167 | 0.4790 | 0.8311 | 0.8754 |
| Dense-image rescued | 39 | 0.1795 | 0.6462 | 0.6462 |
| Multi-retriever rescued | 62 | 0.7742 | 0.9645 | 0.9774 |

Table 10: VisRAG fusion-hurt taxonomy by rescuing component retriever. Values are Recall@10 on the corresponding FH sub-slice.

**Robustness.** The VisRAG target-subset effect persists across candidate-pool depths and cutoffs. Across top-50, top-100, and top-200 candidate pools, FH-weighted CE improves fusion-hurt Recall@10 over plain CE by 2.4–3.2 points with essentially unchanged overall Recall@10. When the fusion-hurt definition and evaluation cutoff are varied, the target-subset gain remains positive at Recall@5, Recall@10, and Recall@20.

### C.2 Public Visual-Document Results

MMDocIR tests the learning signal on a public page-level benchmark with the same three-channel structure as VisRAG. Across all 1,658 queries, equal WRRF over BM25 captions, dense caption retrieval, and ColQwen2 image retrieval induces 150 FH queries. ColQwen2 is the strongest individual source at 0.7666 Recall@10, BM25-caption reaches 0.6634, and dense-caption reaches 0.5452. The selected tuned-WRRF model assigns weights 0.05, 0.05, and 0.90 to BM25-caption, dense-caption, and ColQwen2-image retrieval. This control is intentionally strong: it repairs much of the equal-fusion failure by trusting the strongest visual retriever.

| Method | All R@10 | FH R@10 | Non-FH R@10 |
|---|---|---|---|
| Equal WRRF | 0.7221 | 0.0000 | 0.7836 |
| ColQwen2 image only | 0.7795 | 0.8846 | 0.7705 |
| Dev-tuned WRRF | 0.7795 | 0.7692 | 0.7803 |
| Plain CE reranker | 0.7867 | 0.8462 | 0.7816 |
| FH-weighted CE reranker | 0.7855 | 0.8769 | 0.7777 |

Table 11: MMDocIR internal test Recall@10 over 331 queries and five seeds. The held-out FH subset contains 26 queries.

FH-weighted CE improves the MMDocIR FH subset over plain CE by +0.0308 Recall@10 and +0.0746 nDCG@10. The seed-query paired sign-flip test gives $p = 0.0401$ and exact McNemar gives $p = 0.0391$; at query level, the confidence interval crosses zero because the held-out FH subset contains only 26 queries. The result supports targeted public repair, while also showing that tuned classical fusion and a strong visual branch are serious baselines.

ViDoRe provides a second public visual-document check. Across seven English or verified-English datasets, fusion hurt appears in 91 of 1,114 queries. The rescuing source is not one-sided: ColPali is the best rescuing source for 67 FH queries and BM25 for 24.

ViDoRe validates that fusion hurt recurs outside VisRAG, but it does not support a positive FH-weighting claim. Plain CE is the best point estimate, and its direct query-level difference over dev-tuned WRRF is not significant: +0.0108 all-query Recall@10, CI [-0.0045, +0.0287], $p = 0.2197$.

| Method | All R@10 | FH R@10 | Non-FH R@10 |
|---|---|---|---|
| BM25-only | 0.6996 | 0.3684 | 0.7304 |
| ColPali-only | 0.8520 | 0.6842 | 0.8676 |
| Equal WRRF | 0.8072 | 0.0000 | 0.8824 |
| Dev-tuned WRRF | 0.8475 | 0.6316 | 0.8676 |
| Plain CE reranker | 0.8583 | 0.6947 | 0.8735 |
| FH-weighted CE reranker | 0.8547 | 0.6737 | 0.8716 |

Table 12: Pooled ViDoRe internal test Recall@10 over 223 queries. The FH subset contains 19 queries, so supervised comparisons should be interpreted cautiously.

This is a useful boundary condition: when a strong ColPali branch and tuned fusion already repair much of the equal-fusion failure, explicit FH weighting need not help.

### C.3 BEIR Text-Hybrid Details

BEIR is included to test the channel-agnostic claim with lexical, dense, and hybrid text retrieval channels. Hybrid reaches 0.8388 Recall@10 and 0.7858 nDCG@10. FH-CE improves all-query Recall@10 to 0.8693 and all-query nDCG@10 to 0.8427. On the 218-query FH slice, raw hybrid is 0.0000 by definition; the listwise CE model reaches 0.7104 Recall@10 and 0.4831 nDCG@10. The result is important because it removes any dependence on visual retrieval: fusion hurt is a property of heterogeneous channels and their aggregation policy, not a property of modality.

### C.4 Cost Details

| Method family | CPU scoring cost | CPU end-to-end cost |
|---|---|---|
| Borda / WRRF rank fusion | $< 0.23$ ms/query | $< 0.23$ ms/query |
| Score-scaled / quota fusion | $< 0.38$ ms/query | $< 0.38$ ms/query |
| Listwise MLP reranker | 2.3 ms/query | 6.6 ms/query |
| FH-weighted MLP reranker | 2.3 ms/query | 6.6 ms/query |

Table 13: CPU inference cost in the controlled setup. FH weighting changes training only, so its inference path is identical to the plain MLP reranker. Conventional tree-LTR baselines were trained and evaluated for accuracy, but were not profiled in this latency harness.

The reranker is more expensive than rank-only fusion but remains lightweight in the measured CPU harness. The central cost point is that FH weighting has no additional inference cost over plain CE, because it changes only the training objective. This matters for deployment: if a system already accepts a lightweight post-fusion reranker, the FH-aware variant changes supervision rather than serving-time architecture.