SELFRANK: CLOSED-FORM RERANKING FOR LOW-LATENCY MULTIMODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reranking with a multimodal language model orders visual document retrieval accurately but slowly. Some existing methods require relevance labels to recover accuracy after compression and ignore retriever scores. We propose SelfRank, a reranker built on a closed form rule that quantifies and restores the missing relevance information. We derive the optimal operating point under a correlation objective and establish the exact condition under which reranker scores alone are insufficient to achieve it. SelfRank further restores the reranker with a single vector on its shallow state, fitted to the model's own uncompressed score by one centered ridge solve. Across 12 datasets from ViDoRe 2 and ViDoRe 3 with two retrievers and two backbones, SelfRank comes within 1.2 pp of a full cross encoder in NDCG@5 while running up to 48x faster, pushing the Pareto frontier of accuracy against latency in visual document reranking.

1 INTRODUCTION

Multimodal language model rerankers improve ranking quality in visual document retrieval, but their cost grows with the number of pages scored. Visual document retrieval supports question answering and generation over long, visually rich documents (Yu et al., 2025; Cho et al., 2024; Tanaka et al., 2025; Dong et al., 2025) in domains such as biomedical lectures, economics reports and ESG disclosures (Macé et al., 2025). These applications require low response latency, yet an uncompressed reranker processes each candidate page as a full image.

Some existing methods require relevance labels to recover accuracy after compression and use only reranker scores for ranking. They reduce inference cost by caching page prefixes offline (MacAvaney et al., 2020; Cao et al., 2020; Fan et al., 2026), stopping at intermediate layers (Xin et al., 2020; Schuster et al., 2022; Busolin et al., 2025; Campagnano et al., 2025), and subsampling or quantizing stored visual positions (Liu et al., 2026a; 2024; Sun et al., 2026). These limitations motivate calibrating the compressed reranker with the full model's scores and incorporating retriever scores into the final ranking.

We propose SelfRank, a reranker built on a closed form rule that quantifies and restores the missing relevance information. The two scores are normalized to zero mean and unit variance within each candidate list and ranked by their weighted combination. The operating point is the weight that maximizes the correlation of this combination with relevance, and it depends only on the correlation of each score with relevance and the correlation between the two scores. The same derivation yields the exact condition under which the retriever carries information about relevance beyond the reranker and should receive a positive weight.

SelfRank further restores the reranker with a single vector on its shallow state, fitted to the model's own uncompressed score by one centered ridge solve. That score is the margin between the two answer token logits and equals an inner product with the model's final hidden state, which makes a linear readout of an earlier state the natural object to fit. Centering the fit within each candidate list lets it learn how candidates of the same query differ from one another, the only information the ranking uses. Its ridge strength is chosen by the same centered error on held out queries, and the whole fit is therefore determined by the model's own scores while the backbone stays frozen.

We evaluate SelfRank on 12 datasets using two retrievers and two multimodal backbones. SelfRank comes within 1.2 pp of the full cross encoder in NDCG@5 while running up to 48x faster.

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.

Our contributions are the following.

- We quantify the relevance information missing from the reranker score, derive the optimal operating point under a correlation objective in closed form, and establish the exact condition under which reranker scores alone are insufficient to achieve it. (Section 3.1)
- We propose SelfRank, which fits a shallow linear readout to the model's scores through one centered ridge solve without gradient steps or relevance judgements. A bound relates agreement with that target to correlation with relevance. (Sections 3.2 and 3.3)
- Experiments on 12 datasets from ViDoRe 2 and ViDoRe 3 with two retrievers and two backbones, Qwen2.5-VL-7B and GLM-4.1V-9B, show that SelfRank comes within 1.2 pp of a full cross encoder in NDCG@5 while running up to 48x faster. (Section 4)

2 Related Work

Multimodal document reranking. Late interaction over image patches removed the parsing stage from document retrieval and set the standard for first stage recall (Khattab & Zaharia, 2020; Faysse et al., 2025), and the ViDoRe benchmarks made page level retrieval the common ground for comparing such systems (Faysse et al., 2025; Macé et al., 2025). A second stage built on a multimodal language model improves the order further. The prevailing form is pointwise. Each candidate is scored on its own and the resulting score replaces the retriever's (lightonai, 2024; Fan et al., 2026), and the strongest current reranker orders a list by the margin between its yes and no tokens alone (Li et al., 2026a). Hybrid systems combine two retrievers in representation space (Uzan et al., 2026), and test time methods feed the reranker score back into the query representation and retrieve again (Reddy et al., 2023; Sung et al., 2023). We instead characterize, for a fixed candidate set, the relevance information that the reranker discards and the retriever score still carries.

Efficient inference for rerankers. Query independent parts of the input are encoded once and cached, so a request runs only the query against a stored prefix (MacAvaney et al., 2020; Cao et al., 2020). The forward pass is halted early, either adaptively per instance or at a layer fixed in advance (Xin et al., 2020; Schuster et al., 2022; Busolin et al., 2025; Campagnano et al., 2025). What is stored is reduced by discarding visual positions and by quantizing the tensors that remain (Liu et al., 2026a; 2024). A listwise variant drops autoregressive decoding by scoring a whole candidate list in one pass (Sun et al., 2026), and recent multimodal rerankers combine caching, early exit and cache compression to report large savings on ViDoRe (Fan et al., 2026). We adopt these techniques unchanged and claim none of them. Some methods train the compressed model to recover accuracy. We fit a readout through a closed form solve on unlabelled queries.

First stage scores and layerwise readouts. Neural and first stage scores are combined by rank (Cormack et al., 2009) or by a weighted sum whose weight is tuned on labelled queries (Bruch et al., 2023; Wang et al., 2021; Leonhardt et al., 2024), or the first stage score is fed into the reranker as an input (Askari et al., 2023). In these works the weight is a hyperparameter searched anew for each collection, and when the first stage score adds information beyond the reranker is left unexplained. We instead derive the optimal operating point in closed form, together with the exact condition under which the reranker score alone is insufficient. Linear readouts of intermediate states have been fitted to reproduce the final layer for interpretation (Belrose et al., 2023), trained into the network by supervising shallow layers with deep ones (Zhang et al., 2019; Liu et al., 2020; 2026b), or trained contrastively on the layers of a frozen retriever (Li et al., 2026b). Each of them is obtained by gradient training and corrects a single change to the network. Our readout is a single vector solved in closed form against the frozen reranker's own uncompressed score, and it compensates for the reduced depth and the compressed input at once.

3 Method

A two stage retrieval system returns a set $\mathcal{C}_q$ of K candidate pages and retriever scores $\mathbf{b}_q \in \mathbb{R}^K$ for query q. A multimodal language model then scores these candidates to produce $\mathbf{s}_q \in \mathbb{R}^K$. We combine both scores to determine the final ranking. The retriever fixes the candidate set, so reranking changes only its order and requires no additional retrieval.

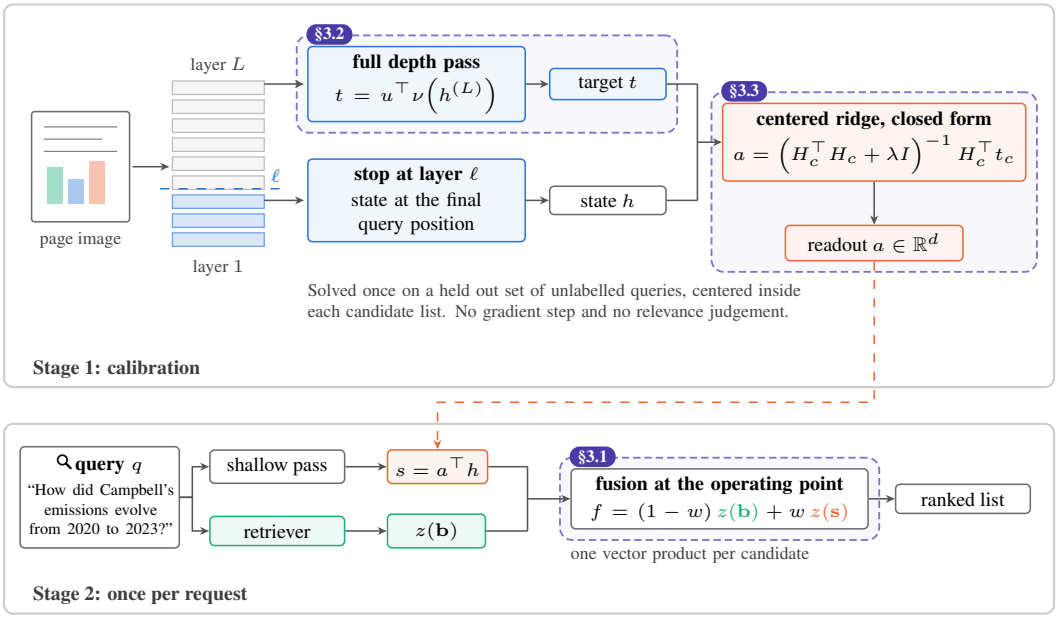

Figure 1: SelfRank. **Stage 1** runs once on unlabelled query page pairs. A full depth pass on the uncompressed page gives the target t, a pass stopped at layer ℓ gives the state h, and one centered ridge solve turns the pairs into the readout a. **Stage 2** runs once per request. A pass stopped at layer ℓ gives the state of each candidate, a scores it with one vector product, and that score is fused with the retriever's at the operating point.

Following prior work (MacAvaney et al., 2020; Cao et al., 2020; Fan et al., 2026), we encode the query independent part of the prompt, comprising the page image and a fixed instruction, once offline and store the first ℓ layers in int8. We keep a fraction γ of the n_{vis} visual positions (Liu et al., 2026a; 2024). At inference, only the query suffix runs over the stored prefix. We treat ℓ and γ as fixed serving settings rather than parameters selected by our method.

Figure 1 summarizes the method. Section 3.1 derives the fusion weight, Section 3.2 defines the self-supervised target, and Section 3.3 gives the centered ridge solution.

3.1 Fusion as Operating Point

We rank by a convex combination of the two scores, each standardized inside the candidate set because the two are on different scales that vary from query to query. The standardization is

$$z(\mathbf{v})_i \;=\; \frac{v_i - \mu(\mathbf{v})}{\sigma(\mathbf{v})}, \qquad \mathbf{v} \in \mathbb{R}^K, \tag{1}$$

where μ and σ are the mean and standard deviation over the K entries, and the ranking is by

$$f_{q,p} \;=\; (1-w)\, z(\mathbf{b}_q)_p \;+\; w\, z(\mathbf{s}_q)_p, \qquad w \in [0,1]. \tag{2}$$

At $w = 1$ the retriever score is discarded and ranking uses the reranker score alone, as in pointwise multimodal reranking. At $w = 0$ the reranker is ignored. The following result characterizes when the correlation objective favors an interior weight.

Proposition 1. *Write $\mathbf{y} \in \mathbb{R}^K$ for the standardized graded relevance of a candidate set and set*

$$c_b = \tfrac{1}{K}\langle z(\mathbf{b}), \mathbf{y}\rangle, \qquad c_s = \tfrac{1}{K}\langle z(\mathbf{s}), \mathbf{y}\rangle, \qquad \rho = \tfrac{1}{K}\langle z(\mathbf{b}), z(\mathbf{s})\rangle, \tag{3}$$

so that c_b and c_s are the correlations of the two scores with relevance and ρ is their correlation with each other. If $|\rho| < 1$, $c_b > \rho c_s$, and $c_s > \rho c_b$, the weight maximizing the correlation between $\mathbf{f}$ of Equation 2 and $\mathbf{y}$ over $w \in [0, 1]$ is

$$w^\star \;=\; \frac{c_s - \rho\, c_b}{(c_b + c_s)(1 - \rho)}, \tag{4}$$

which lies strictly inside $(0, 1)$.

Proof sketch. Writing $\alpha = 1 - w$ and $\beta = w$, the correlation between $\mathbf{f}$ and $\mathbf{y}$ is $(\alpha c_b + \beta c_s)/\sqrt{\alpha^2 + \beta^2 + 2\alpha\beta\rho}$, a linear form divided by the norm induced by the score correlation matrix Σ. This matrix has diagonal entries one and off diagonal entries ρ; $|\rho| < 1$ ensures that it is positive definite. By the Cauchy Schwarz inequality the correlation is maximized along $\Sigma^{-1}(c_b, c_s)^\top$, proportional to $(c_b - \rho c_s,\ c_s - \rho c_b)$. Both components are positive under the stated conditions, and rescaling them to $\alpha + \beta = 1$ gives Equation 4. Section A.1 gives the full proof. □

The condition $c_b > \rho c_s$ says that the retriever contributes information about relevance beyond its correlation with the reranker score. The difference $c_b - \rho c_s$ is the numerator of the retriever's partial correlation with relevance after accounting for the reranker. The corresponding condition $c_s > \rho c_b$ ensures that the reranker also receives positive weight. When both hold, ranking by either score alone gives a lower value of the correlation objective. These conditions concern the contribution of each score to relevance, not whether the two scores assign similar values to individual candidates. We read the three quantities of Equation 3 as averages over a corpus, each computed with an inner product weighted by the discount of the ranking metric, $\omega_i = 1/\log_2(1 + \text{rank}_i)$ under the retriever's order. The statement remains valid with this weighted inner product (Section A.1). We deploy $w = w^\star$, with every vector standardized and the three correlations read in this weighted inner product on another corpus of the same benchmark, so the corpus being ranked contributes no relevance judgement to its own weight. The fusion itself keeps the standardization of Equation 1.

3.2 Self-Supervised Target

The reranker score is a linear readout of the state at layer ℓ. We run the query suffix over the stored prefix to that layer and take the hidden state at the final suffix position, $h_{q,p} \in \mathbb{R}^d$, where d is the backbone width, so that

$$s_{q,p} \;=\; a^\top h_{q,p}, \qquad a \in \mathbb{R}^d, \tag{5}$$

whose cost at request time is one vector product per candidate, negligible beside the forward pass that produces $h_{q,p}$.

The target is the score the same backbone gives at full depth on the uncompressed page. Let L be the final layer, let $h^{(L)}_{q,p}$ be the state at the final position of that uncompressed forward, and write $o_{\texttt{yes}}$ and $o_{\texttt{no}}$ for the logits of the two answer tokens. The target is the decision margin

$$t_{q,p} \;=\; o_{\texttt{yes}}(q,p) \;-\; o_{\texttt{no}}(q,p) \;=\; u^\top \nu\big(h^{(L)}_{q,p}\big), \qquad u = u_{\texttt{yes}} - u_{\texttt{no}}, \tag{6}$$

where $u_{\texttt{yes}}$ and $u_{\texttt{no}}$ are the unembedding rows of the two answer tokens and ν is the final normalization the backbone applies before unembedding. The full depth model is used only during offline calibration.

Equation 6 motivates a linear readout but does not imply an exact linear map from the deployed state to the target. The final score is a linear function of the normalized final state; both the intervening layers and the final normalization can be nonlinear. The deployed state $h_{q,p}$ differs from $h^{(L)}_{q,p}$ because depth is truncated from L to ℓ and the stored prefix is subsampled and quantized. Writing W for a linear surrogate of the map from the deployed state to the uncompressed one on the subspace spanned by calibration features, and ν' for a local linear approximation to ν on that subspace, we obtain

$$t_{q,p} \;\approx\; u^\top \nu' W\, h_{q,p} \;=\; \big(W^\top \nu'^\top u\big)^\top h_{q,p}, \tag{7}$$

which has the form of Equation 5 with $a = W^\top \nu'^\top u$. This approximation combines the effects of depth truncation and cache compression in W. The readout estimates only the direction of that map relevant to the target margin, rather than the entire final state. Thus one fitted vector can compensate for both changes when the linear approximation holds on the calibration distribution. Any constant offset in a local approximation does not affect ranking and is removed by query centering in Section 3.3. The approximation is empirical and need not describe every possible shallow state.

3.3 Centered Ridge in Closed Form

Only the order within a candidate set matters, and Equation 2 is unchanged when a constant is added to every score for one query. We give the fit the same invariance by removing the query mean from states and targets. Let $J \in \mathbb{R}^{K\times K}$ be the matrix of all ones and

$$P \;=\; I - \tfrac{1}{K} J \tag{8}$$

the projector that removes the mean, with I the identity of the indicated size. Stacking the states of a query into $H_q \in \mathbb{R}^{K\times d}$ and its targets into $t_q \in \mathbb{R}^K$, the readout solves

$$\min_{a\in\mathbb{R}^d} \sum_{q\in\mathcal{Q}_{\mathrm{fit}}} \big\|P\big(H_q a - t_q\big)\big\|_2^2 \;+\; \lambda\|a\|_2^2, \tag{9}$$

over a fixed calibration set $\mathcal{Q}_{\mathrm{fit}}$, with $\lambda > 0$ the ridge strength. Since P is idempotent and symmetric, $H_q^\top P H_q = (PH_q)^\top (PH_q)$, so Equation 9 is an ordinary ridge problem on the centered data and carries no intercept. Collecting the centered states of every query into H_c and the centered targets into t_c, its solution is

$$a \;=\; \big(H_c^\top H_c + \lambda I\big)^{-1} H_c^\top t_c, \tag{10}$$

obtained by one linear solve. This fit requires no gradient step and leaves the backbone unchanged. Algorithm 1 states the two procedures in full. Although Equation 9 uses no relevance judgement, the following proposition relates agreement with its teacher target to the relevance correlation c_s in Proposition 1.

Proposition 2. *Fix a candidate set and let $c_t = \frac{1}{K}\langle z(\mathbf{t}), \mathbf{y}\rangle$ be the correlation of the target with relevance and $r = \frac{1}{K}\langle z(\mathbf{s}), z(\mathbf{t})\rangle$ the correlation of the model score with the target. Then*

$$c_s \;\geq\; r\, c_t \;-\; \sqrt{\big(1-r^2\big)\big(1-c_t^2\big)}, \tag{11}$$

and for $r, c_t \in (0,1)$ the right hand side is strictly increasing in r.

Proof sketch. The correlation matrix of $z(\mathbf{s})$, $z(\mathbf{t})$ and $\mathbf{y}$ is a Gram matrix and hence positive semidefinite, so its determinant $1 - r^2 - c_t^2 - c_s^2 + 2\, r\, c_t\, c_s$ is at least zero. Read as a quadratic in c_s, this confines c_s between the roots $r\, c_t \pm \sqrt{(1-r^2)(1-c_t^2)}$, and the lower root is Equation 11. Its derivative in r is $c_t + r\sqrt{(1-c_t^2)/(1-r^2)}$, which is positive on $(0,1)$. Section A.2 gives the full proof. □

The objective fits the teacher margin without reading a relevance judgement. Once the backbone is fixed, c_t is fixed as well, while agreement r is the term in Equation 11 that the readout can affect. For any candidate set with $r, c_t \in (0,1)$, increasing r raises the lower bound on c_s. At $\lambda \to 0$, the solution of Equation 9 maximizes squared correlation with the teacher over the centered calibration data within the linear family. This is a global property of the fit and does not guarantee an increase in r or c_s for every candidate set. Positive λ limits the size of the fitted coefficients and can reduce variance when calibration data are limited. We select λ by minimizing the centered squared error of Equation 9, without the penalty, on held out queries. This uses the target of Equation 6 without relevance annotation (Section B.3).

Algorithm 1 SelfRank.

Require: calibration queries $\mathcal{Q}_{\text{fit}}$, depth ℓ, ridge λ, weight w

function FIT($\mathcal{Q}_{\text{fit}}$) ▷ once, no relevance judgement
  for all $q \in \mathcal{Q}_{\text{fit}},\ p \in \mathcal{C}_q$ **do**
    $h_{q,p} \leftarrow \textsc{Shallow}(q, p, \ell)$
    $t_{q,p} \leftarrow o_{\texttt{yes}} - o_{\texttt{no}}$ **from** FORWARD($q, p,$ depth=$L,$ uncompressed)
  end for
  $H_c, t_c \leftarrow PH,\ Pt$ ▷ $P = I - \frac{1}{K}J$, per candidate set
  return $a \leftarrow (H_c^\top H_c + \lambda I)^{-1} H_c^\top t_c$
end function

function RANK(q) ▷ online, per request
  $\mathcal{C}_q,\ \mathbf{b}_q \leftarrow \textsc{Retrieve}(q)$
  for all $p \in \mathcal{C}_q$ **do**
    $s_{q,p} \leftarrow a^\top \textsc{Shallow}(q, p, \ell)$
  end for
  return $\textsc{ArgSort}\big((1-w)\, z(\mathbf{b}_q) + w\, z(\mathbf{s}_q)\big)$
end function

4 EVALUATION

Section 4.1 reports ranking quality, Section 4.2 the accuracy and latency tradeoff, Section 4.3 the validation of the two propositions, Section 4.4 the component ablation, and Section 4.5 the parameter sensitivity.

Datasets and models. We evaluate on ViDoRe 2 (Macé et al., 2025) and ViDoRe 3 (Loison et al., 2026), 1038 and 4730 queries over four and eight corpora, on one NVIDIA A100 with 40 GiB. ColNomic-7B (NomicAI, 2025) and ColPali-v1.3 (Faysse et al., 2025) each supply twenty candidates, and Qwen2.5-VL-7B (Bai et al., 2025) and GLM-4.1V-9B (Hong et al., 2025) each supply a readout, taken at layer 21 from one third of the visual positions.

Baselines. We compare against the retriever alone, against Guided Query Refinement (Uzan et al., 2026) under the retriever its published configuration uses, and against a full cross encoder, which serves as a full model reference, not a competing baseline. Two hybrid baselines, RRF (Cormack et al., 2009) and Score Aggregation (Bruch et al., 2023), fuse the retriever with the text retriever Linq-Embed (Choi et al., 2024), Qwen3-Reranker-0.6B (Zhang et al., 2025) reranks the page text and MonoQwen2-VL (lightonai, 2024) the page image, both trained on relevance judgements, and the weight of every fused baseline is picked by NDCG@5 on the same inner corpus that sets ours. The reference is the full Qwen 7B scoring every candidate image; both backbones are compared against it.

For each backbone, we fit four readouts using the 154 calibration queries sampled from the four ViDoRe 2 corpora under leave one corpus out. In each fold, queries from three corpora are used for fitting, and the resulting readout evaluates the remaining corpus. On ViDoRe 3, each corpus is evaluated with all four readouts, and we report the mean of the resulting metrics. We assign one fusion weight per corpus using the three correlations defined in Proposition 1, estimated on the next corpus in a fixed cyclic order. Weight estimation is performed separately within ViDoRe 2 and ViDoRe 3.

4.1 RANKING QUALITY

Table 1 and Table 2 give NDCG@5 by subset. On ViDoRe 2 with ColNomic candidates the retriever scores 60.9 and the full cross encoder 72.4, leaving 11.5 points of distance. GQR closes 2.9 of them, RRF, Score Aggregation and Qwen3-Reranker-0.6B reach 62.3, 62.2 and 61.6, and MonoQwen2-VL reaches 65.4. The frozen readout closes 8.1 with the Qwen backbone and 10.3 with the GLM backbone, which is 70 and 89 percent of what the cross encoder is worth.

Table 1: NDCG@5 (%) on all four subsets of ViDoRe 2. Within each block the deltas are taken against the retriever alone. The full cross encoder scores every candidate from its image and is reported as a reference, not as a competing baseline. GQR is listed only under the retriever its published configuration uses.

Retriever	Reranker	Avg		Biomed		Econ		ESG-H		ESG-F	
		val	Δ	val	Δ	val	Δ	val	Δ	val	Δ
ColNomic-7B	No reranking	60.9	–	63.9	–	55.1	–	70.7	–	53.8	–
	Full cross encoder	72.4	+11.5	71.7	+7.8	68.2	+13.1	83.1	+12.4	66.7	+12.9
	RRF	62.3	+1.4	63.1	−0.8	58.7	+3.6	68.2	−2.6	59.1	+5.3
	Score Aggregation	62.2	+1.3	63.6	−0.2	57.4	+2.2	69.9	−0.9	58.0	+4.2
	GQR	63.8	+2.9	66.3	+2.4	59.4	+4.2	69.3	−1.4	60.1	+6.3
	Qwen3-Reranker-0.6B	61.6	+0.7	64.3	+0.4	57.1	+2.0	68.3	−2.5	56.9	+3.1
	MonoQwen2-VL	65.4	+4.5	67.6	+3.7	59.3	+4.1	73.7	+2.9	61.0	+7.2
	Ours (Qwen2.5-VL-7B)	69.0	↑ 8.1	71.4	↑ 7.5	64.8	↑ 9.6	76.9	↑ 6.1	63.1	↑ 9.3
	Ours (GLM-4.1V-9B)	71.2	↑ 10.3	72.1	↑ 8.3	65.8	↑ 10.7	79.1	↑ 8.4	67.8	↑ 14.0
ColPali-v1.3	No reranking	50.2	–	56.2	–	46.3	–	49.7	–	48.8	–
	Full cross encoder	67.6	+17.3	68.8	+12.7	64.1	+17.9	70.0	+20.3	67.4	+18.5
	RRF	55.8	+5.5	59.9	+3.7	53.5	+7.2	56.8	+7.1	52.9	+4.0
	Score Aggregation	56.8	+6.6	60.2	+4.1	54.9	+8.6	57.2	+7.5	54.9	+6.1
	Qwen3-Reranker-0.6B	54.7	+4.5	56.5	+0.4	51.8	+5.6	57.8	+8.0	52.7	+3.8
	MonoQwen2-VL	62.6	+12.4	64.8	+8.6	56.1	+9.9	67.3	+17.6	62.3	+13.5
	Ours (Qwen2.5-VL-7B)	64.9	↑ 14.7	68.3	↑ 12.2	62.4	↑ 16.2	67.4	↑ 17.6	62.6	↑ 13.8
	Ours (GLM-4.1V-9B)	66.1	↑ 15.8	69.1	↑ 13.0	62.9	↑ 16.7	68.2	↑ 18.5	64.1	↑ 15.3

Table 2: NDCG@5 (%) on all eight public subsets of ViDoRe 3. Within each block the deltas are taken against the retriever alone. The full cross encoder scores every candidate from its image and is reported as a reference, not as a competing baseline. GQR is listed only under the retriever its published configuration uses.

Retriever	Reranker	Avg		cs		energy		fin en		fin fr		hr		indus		pharma		physics	
		val	Δ	val	Δ	val	Δ	val	Δ	val	Δ	val	Δ	val	Δ	val	Δ	val	Δ
ColNomic-7B	No reranking	52.8	–	72.4	–	61.9	–	43.4	–	42.1	–	54.1	–	43.7	–	57.8	–	47.1	–
	Full cross encoder	58.0	+5.2	76.6	+4.1	67.5	+5.7	52.1	+8.6	48.6	+6.6	59.4	+5.3	50.0	+6.3	61.9	+4.0	48.2	+1.1
	RRF	51.4	−1.4	70.6	−1.9	57.4	−4.4	41.1	−2.3	42.6	+0.5	51.2	−2.9	42.2	−1.5	58.9	+1.1	47.4	+0.3
	Score Aggregation	53.5	+0.7	72.4	+0.0	62.2	+0.3	44.2	+0.8	43.0	+0.9	54.8	+0.7	44.5	+0.8	59.6	+1.7	47.2	+0.1
	GQR	54.3	+1.5	73.4	+1.0	63.4	+1.5	45.6	+2.2	43.4	+1.3	55.2	+1.1	45.1	+1.5	59.9	+2.1	48.1	+1.0
	Qwen3-Reranker-0.6B	54.0	+1.2	68.8	−3.6	64.0	+2.1	45.7	+2.3	44.2	+2.2	56.5	+2.4	45.0	+1.3	60.2	+2.4	47.3	+0.2
	MonoQwen2-VL	54.5	+1.7	73.9	+1.5	63.1	+1.3	47.4	+4.0	43.8	+1.7	55.4	+1.3	45.9	+2.2	59.1	+1.3	47.7	+0.6
	Ours (Qwen2.5-VL-7B)	56.2	↑ 3.4	76.3	↑ 3.9	66.2	↑ 4.4	48.7	↑ 5.2	45.1	↑ 3.1	57.6	↑ 3.6	47.8	↑ 4.1	61.2	↑ 3.4	46.9	↓0.2
	Ours (GLM-4.1V-9B)	55.8	↑ 3.0	76.4	↑ 4.0	65.2	↑ 3.3	47.9	↑ 4.5	43.8	↑ 1.8	57.1	↑ 3.0	46.8	↑ 3.1	62.8	↑ 5.0	47.7	↑ 0.6
ColPali-v1.3	No reranking	38.8	–	57.3	–	43.9	–	27.3	–	23.6	–	40.1	–	30.1	–	48.0	–	39.8	–
	Full cross encoder	51.1	+12.3	73.5	+16.2	58.3	+14.4	39.8	+12.4	37.7	+14.1	53.6	+13.4	42.0	+11.9	57.5	+9.6	46.1	+6.2
	RRF	42.9	+4.2	60.2	+2.8	48.5	+4.5	29.5	+2.2	28.9	+5.3	45.6	+5.5	32.5	+2.4	54.2	+6.3	44.3	+4.4
	Score Aggregation	44.6	+5.8	64.5	+7.2	51.2	+7.2	31.3	+4.0	31.0	+7.4	45.1	+5.0	32.8	+2.7	55.9	+8.0	44.9	+5.0
	Qwen3-Reranker-0.6B	47.8	+9.0	68.0	+10.6	56.8	+12.9	35.7	+8.3	33.7	+10.2	49.8	+9.7	36.9	+6.8	54.6	+6.7	46.6	+6.8
	MonoQwen2-VL	46.7	+7.9	67.2	+9.8	54.6	+10.6	34.3	+7.0	32.0	+8.4	48.3	+8.2	38.7	+8.6	54.9	+7.0	43.7	+3.9
	Ours (Qwen2.5-VL-7B)	48.8	↑ 10.0	71.8	↑ 14.4	56.8	↑ 12.9	36.5	↑ 9.2	34.5	↑ 10.9	51.6	↑ 11.5	39.2	↑ 9.1	57.5	↑ 9.5	45.6	↑ 5.8
	Ours (GLM-4.1V-9B)	48.8	↑ 10.1	75.1	↑ 17.7	55.8	↑ 11.8	35.0	↑ 7.6	32.8	↑ 9.2	50.9	↑ 10.8	37.5	↑ 7.4	57.8	↑ 9.8	46.8	↑ 7.0

NDCG@5 is the primary metric throughout. Table 5 in Appendix B reports four further metrics for every arm, and SelfRank leads every external baseline under all of them. The readout improves on the retriever in every one of the four settings, by 8.1 and 10.3 points on ViDoRe 2 with ColNomic candidates, by 14.7 and 15.8 with ColPali, and by 3.4 and 3.0 on ViDoRe 3.

The same four vectors serve both retrievers and both benchmarks. They were fitted on ColNomic candidates over four ViDoRe 2 corpora, so the ColPali pools and every ViDoRe 3 corpus are outside the fit entirely, and the coefficients still close 85 and 91 percent of the ColPali distance on ViDoRe 2 and 82 percent of it on ViDoRe 3 with either backbone.

The primary comparison is against the full cross encoder reference. On ViDoRe 2 the GLM readout sits 1.2 points below the full cross encoder with ColNomic candidates and 1.5 points below it with ColPali candidates, so a vector fitted once on 154 queries comes within two points of the reference cross encoder, at a small fraction of its cost.

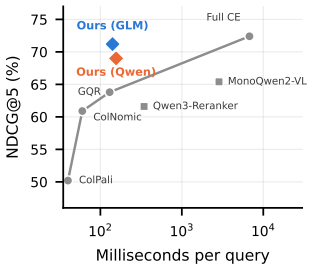

Figure 2: Pareto frontier.

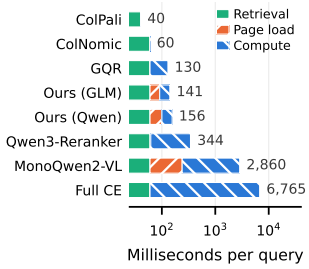

Figure 3: Latency breakdown.

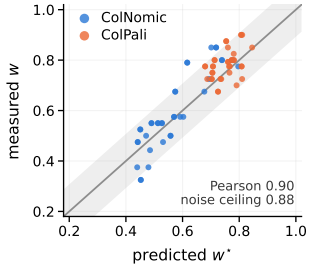

Figure 4: Predicted weight. The band is the 95% interval of the measured weight.

4.2 Accuracy and Latency Tradeoff

Figure 2 places accuracy against measured cost, and Table 4 in Appendix B gives the same numbers for both retrievers. Every latency is the mean over the whole query set and not over a sample. On ViDoRe 2 the retriever costs 60 ms, GQR costs 130 ms, Qwen3-Reranker-0.6B costs 344 ms, MonoQwen2-VL costs 2860 ms, our method costs 156 ms with the Qwen backbone and 141 ms with the GLM backbone, and the full cross encoder costs 6765 ms. The GLM configuration is therefore 48 times cheaper than the cross encoder it matches, and 11 ms more expensive than GQR while closing 65 more percentage points of the available distance.

The curve of Figure 2 is steep up to about 150 ms. The first 80 ms of reranking buys 10.3 points and the remaining 6.6 seconds buy 1.2 more. Both backbones sit at the corner, which is the part of the curve a deployment would choose from. On ViDoRe 3 the same ordering holds against the cross encoder, at 213 ms and 233 ms against 7534 ms, but the margin over GQR narrows to 1.6 points.

Figure 3 splits each cost into retrieval, reading what a method keeps for each page, and model computation. SelfRank computes for 50 ms with the GLM backbone and 56 ms with the Qwen backbone, against 282 ms for Qwen3-Reranker-0.6B over the page text, 2.6 s for MonoQwen2-VL over the page images and 6.7 s for the full cross encoder.

4.3 Theoretical Validation

We evaluate Proposition 1 on 72 subsets: the full cross encoder and both readouts on every corpus under both retrievers. Both $c_b - \rho c_s$ and $c_s - \rho c_b$ are at least 0.085, giving interior optima. We define the empirical optimum as the mean of weights within one NDCG@5 point of the best. Over 200 random half splits, Spearman–Brown correction gives reliabilities of 0.903 and 0.863 for the predicted and measured weights. Their geometric mean, 0.883, estimates a correlation ceiling under parallel measurements and independent errors. Shared annotations violate error independence here, so this ceiling does not apply. Figure 4 reports a correlation of 0.898, with the weaker retriever leaving more weight to the reranker. Here the correlations are read on each subset itself, which tests the proposition, while the deployed weight reads them on another corpus.

Proposition 2 ties the agreement Equation 9 maximizes without labels to c_s. We test the tie by holding the corpus fixed, which leaves c_t constant, and moving only the fit over every layer crossed with every regularization strength. Figure 5 shows r and c_s correlating from 0.894 to 0.994 over that family, with a median of 0.975 on the eight pairings of backbone and corpus.

4.4 Component Ablation

Table 3 reports component ablations for Qwen2.5-VL-7B on the calibration sample. Each row removes one component with the remaining settings fixed.

The hardest row is the logit lens, which reads the same cached state through the backbone's own unembedding and costs nothing to fit. The solve is worth 4.1 points over it. Consuming the model

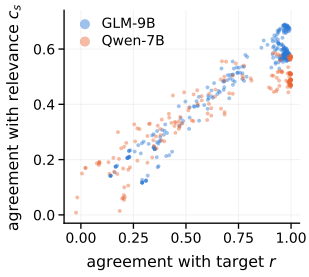

Figure 5: Objective alignment.

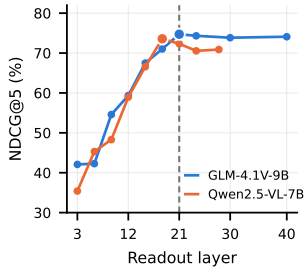

Figure 6: Accuracy by layer.

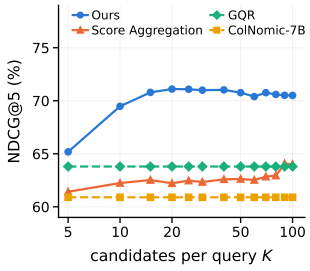

Figure 7: Pool depth.

Table 3: Component ablation on the 154 calibration queries under leave one corpus out, with $w = 0.85$ unless fusion is removed.

Component	Avg		Biomed		Econ		ESG-H		ESG-F	
	val	Δ	val	Δ	val	Δ	val	Δ	val	Δ
(−) Operating point	69.6	↓2.7	54.8	↓4.7	71.1	↓2.7	75.5	↓3.7	76.0	↓0.8
(−) Fitted readout	68.3	↓4.1	55.4	↓4.0	71.5	↓2.3	77.5	↓1.7	68.6	↓8.3
(−) List centering	71.7	↓0.6	59.1	↓0.4	73.0	↓0.9	79.0	↓0.2	75.9	↓0.9
SelfRank	**72.3**	–	**59.4**	–	**73.8**	–	**79.2**	–	**76.8**	–

score on its own costs 2.7 points, which is Proposition 1 measured. Centering inside the candidate list is worth 0.6 points, which is what Equation 9 buys over a pointwise regression with an intercept.

4.5 Parameter Sensitivity

Every input the method takes sits on a plateau. Figure 6 reads the state at each depth of the backbone and finds both peaking before their last layer. Stopping at an intermediate layer follows prior work (MacAvaney et al., 2020; Cao et al., 2020; Fan et al., 2026) and is not part of our contribution, and we fix $\ell = 21$ in advance because choosing the peak of this curve would read relevance judgements. Figure 7 scores pools from five to a hundred candidates with the coefficients frozen, which climbs to the deployed twenty and is flat over the fivefold range above it. The retriever's line there is flat by construction, since the top five of the top twenty is the top five of the top hundred. The ridge strength and the fusion weight behave the same way, moving ranking quality by less than 1.6 points across two decades of λ and by less than half a point across the weights nearest the best.

The fit is also cheap in the only resource it consumes. Drawing calibration queries at random and refitting, sixteen queries already reach 99.4 percent of the quality obtained from all 154 on GLM and 98.3 percent on Qwen.

5 Conclusion

Multimodal rerankers improve visual document retrieval but incur substantial inference costs, and recovering accuracy after compression can require relevance judgements. SelfRank combines retriever and reranker scores using a fusion weight derived in closed form from their correlations with relevance and with each other. The calibration target is the uncompressed model's answer token margin, computed by an inner product with its normalized final state. A shallow linear readout approximates this target through one centered ridge solve, using differences within candidate lists without gradient updates or relevance annotations. Experiments on ViDoRe 2 and ViDoRe 3 show that SelfRank comes within 1.2 pp of a full cross encoder in NDCG@5 while running up to 48x faster.

6 AI USE STATEMENT

In this work, we used generative AI tools to implement methods, to design or provide feedback on the methodology and experiments, and to assist in the writing of proofs. We have not used generative AI tools to generate synthetic data sets, to interpret results, to clean or reformat datasets, to formulate mathematical claims, or to propose or refine hypotheses, and assistance with translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or edit software code and to identify, search and summarize relevant literature. We have reviewed all AI-assisted work. The authors reviewed all generated code and reran every experiment reported in the paper, traced every reported number to the stored outputs of those runs, checked each proof line by line, verified every cited reference against the original publication, and revised all drafted text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arian Askari, Amin Abolghasemi, Gabriella Pasi, Wessel Kraaij, and Suzan Verberne. Injecting the bm25 score as text improves bert-based re-rankers. In *European Conference on Information Retrieval*, pp. 66–83. Springer, 2023.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Sebastian Bruch, Siyu Gai, and Amir Ingber. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, 42(1):1–35, 2023.

Francesco Busolin, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Salvatore Trani, and Alberto Veneri. Efficient re-ranking with cross-encoders via early exit. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2534–2544, 2025.

Cesare Campagnano, Antonio Mallia, Jack Pertschuk, and Fabrizio Silvestri. E2rank: Efficient and effective layer-wise reranking. In *European Conference on Information Retrieval*, pp. 417–426. Springer, 2025.

Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian. Deformer: Decomposing pre-trained transformers for faster question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4497, 2020.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy yong Sohn. Linq-embed-mistral technical report, 2024. URL https://arxiv.org/abs/2412.03223.

Gordon V Cormack, Charles LA 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*, pp. 758–759, 2009.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. Mm-docir: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30959–30993, 2025.

Yingqi Fan, Xuan Lu, Anhao Zhao, Junlong Tong, Ping Nie, Kai Zou, Yunpu Ma, Wei Zhang, and Xiaoyu Shen. minireranker: Efficient multimodal reranking through visual cache reuse and interaction sparsity. *arXiv preprint arXiv:2606.10759*, 2026.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Jurek Leonhardt, Henrik Müller, Koustav Rudra, Megha Khosla, Abhijit Anand, and Avishek Anand. Efficient neural ranking using forward indexes and lightweight encoders. *ACM Transactions on Information Systems*, 42(5):1–34, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

Weien Li, Rui Song, Zeyu Li, Haochen Liu, Gonghao Zhang, Difan Jiao, Zhenwei Tang, Bowei He, Haolun Wu, Xue Liu, et al. Miner: Mining multimodal internal representation for efficient retrieval. *arXiv preprint arXiv:2605.06460*, 2026b.

lightonai. Monoqwen-vision, the first visual document reranker, 2024. URL `https://lighton.ai/lighton-blogs/monoqwen-vision`.

Siyi Liu, Hanjun Yang, Chenchen Zhang, Xiaorong Zhu, Xinyu Zuo, Lisheng Duan, Haijin Liang, Jin Ma, Junfu Pu, and Yongqi Zhang. From saliency to discriminability: Rank-preserving visual token pruning for vlm rerankers. *arXiv preprint arXiv:2609.00667*, 2026a.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. Fastbert: a self-distilling bert with adaptive inference time. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6035–6044, 2020.

Zhuchenyang Liu, Yao Zhang, and Yu Xiao. Nanovdr: Distilling a 2b vision-language retriever into a 70m text-only encoder for visual document retrieval. *arXiv preprint arXiv:2603.12824*, 2026b.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel de Souza P Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16570–16600, 2026.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Efficient document re-ranking for transformers by precomputing term representations. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 49–58, 2020.

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

NomicAI. Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval, 2025. URL `https://nomic.ai/blog/posts/nomic-embed-multimodal`.

Revanth Gangi Reddy, Pradeep Dasigi, Md Arafat Sultan, Arman Cohan, Avirup Sil, Heng Ji, and Hannaneh Hajishirzi. Refit: Relevance feedback from a reranker during inference. *arXiv preprint arXiv:2305.11744*, 2023.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472, 2022.

Yiqun Sun, Pengfei Wei, and Lawrence B Hsieh. Very efficient listwise multimodal reranking for long documents. *arXiv preprint arXiv:2605.11864*, 2026.

Mujeen Sung, Jungsoo Park, Jaewoo Kang, Danqi Chen, and Jinhyuk Lee. Optimizing test-time query representations for dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5731–5746, 2023.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24827–24837. IEEE, 2025.

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*, volume 2026, pp. 495–522, 2026.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval*, pp. 317–324, 2021.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. Deebert: Dynamic early exiting for accelerating bert inference. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2246–2251, 2020.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pp. 21074–21098, 2025.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3712–3721. IEEE, 2019.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A PROOFS

Throughout, a candidate set has K members and $\langle x, y\rangle_K = \frac{1}{K}\sum_i x_i y_i$. A vector is standardized when $\langle x, \mathbf{1}\rangle_K = 0$ and $\langle x, x\rangle_K = 1$, which is what Equation 1 produces, and $\mathbf{y}$ is standardized the same way. All three of $z(\mathbf{b})$, $z(\mathbf{s})$ and $\mathbf{y}$ are therefore unit vectors in this inner product, so an inner product between any two of them is their correlation.

A.1 PROPOSITION 1

The objective. Write $\alpha = 1 - w$ and $\beta = w$, so that $\mathbf{f} = \alpha\, z(\mathbf{b}) + \beta\, z(\mathbf{s})$. Its correlation with $\mathbf{y}$ is the inner product of $\mathbf{y}$ with $\mathbf{f}$ normalized to unit length, and $\mathbf{f}$ is centered because both of its terms are, so no mean has to be removed. Bilinearity gives the numerator

$$\langle \mathbf{f}, \mathbf{y}\rangle_K \;=\; \alpha\, \langle z(\mathbf{b}), \mathbf{y}\rangle_K + \beta\, \langle z(\mathbf{s}), \mathbf{y}\rangle_K \;=\; \alpha\, c_b + \beta\, c_s, \tag{12}$$

and, using $\langle z(\mathbf{b}), z(\mathbf{b})\rangle_K = \langle z(\mathbf{s}), z(\mathbf{s})\rangle_K = 1$ and $\langle z(\mathbf{b}), z(\mathbf{s})\rangle_K = \rho$, the squared norm

$$\langle \mathbf{f}, \mathbf{f}\rangle_K \;=\; \alpha^2 + \beta^2 + 2\alpha\beta\rho. \tag{13}$$

Collecting $v = (\alpha, \beta)^\top$, $c = (c_b, c_s)^\top$ and $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, the correlation to be maximized is the ratio

$$\mathcal{R}(v) \;=\; \frac{v^\top c}{\sqrt{v^\top \Sigma\, v}}. \tag{14}$$

The maximizer. The eigenvalues of Σ are $1 \pm \rho$, so $|\rho| < 1$ makes Σ positive definite and $\Sigma^{1/2}$ invertible. Substituting $u = \Sigma^{1/2} v$ turns Equation 14 into $u^\top \Sigma^{-1/2} c / \|u\|$, which by the Cauchy Schwarz inequality is at most $\|\Sigma^{-1/2} c\| = \sqrt{c^\top \Sigma^{-1} c}$, with equality exactly when u is a positive multiple of $\Sigma^{-1/2} c$. Undoing the substitution, the maximizers are the positive multiples of

$$\Sigma^{-1} c \;=\; \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} c_b \\ c_s \end{pmatrix} \;\propto\; \begin{pmatrix} c_b - \rho\, c_s \\ c_s - \rho\, c_b \end{pmatrix}. \tag{15}$$

The normalization. Equation 14 is invariant to the scale of v, so the whole ray of Equation 15 attains the maximum and the constraint $\alpha + \beta = 1$ of Equation 2 selects one point of it. The two components of Equation 15 sum to

$$(c_b - \rho\, c_s) + (c_s - \rho\, c_b) \;=\; (c_b + c_s)(1 - \rho), \tag{16}$$

which is positive since both components are positive by assumption. Normalizing by this sum gives positive weights with $\alpha + \beta = 1$, so the maximizing direction is feasible and its second component is

$$w^\star \;=\; \frac{c_s - \rho\, c_b}{(c_b + c_s)(1 - \rho)}, \tag{17}$$

which is Equation 4.

The endpoint. Since the denominator is positive, $w^\star < 1$ is equivalent to $c_s - \rho\, c_b < (c_b + c_s)(1 - \rho)$. Expanding the right hand side gives $c_b + c_s - \rho\, c_b - \rho\, c_s$, and cancelling $c_s - \rho\, c_b$ from both sides leaves $0 < c_b - \rho\, c_s$, which is the stated condition. The same computation on the first component gives $w^\star > 0$ if and only if $c_s > \rho\, c_b$, so the optimum is interior exactly when each score agrees with relevance by more than a ρ scaled copy of the other would. □

Weighted positions. Fix $\omega \in \mathbb{R}^K$ with every $\omega_i > 0$ and replace $\langle \cdot, \cdot \rangle_K$ by $\langle x, y \rangle_\omega = \left(\sum_i \omega_i\right)^{-1} \sum_i \omega_i\, x_i y_i$, standardizing every vector and defining c_b, c_s and ρ in that inner product. Positive weights make $\langle \cdot, \cdot \rangle_\omega$ symmetric, bilinear and positive definite, which is everything the argument above uses. The numerator expands by bilinearity, the squared norm by bilinearity and symmetry, and the Cauchy Schwarz step holds in any inner product space. Every line therefore goes through unchanged and the statement holds verbatim. The deployed choice $\omega_i = 1/\log_2(1 + \mathrm{rank}_i)$ is the discount of the ranking metric itself, taken under the order the retriever supplies, so it is fixed before any weight is chosen and carries no free parameter.

A.2 Proposition 2

A positive semidefinite matrix. Fix a candidate set and collect the three standardized vectors $z(\mathbf{s})$, $z(\mathbf{t})$ and $\mathbf{y}$. Their matrix of pairwise inner products is

$$R \;=\; \begin{pmatrix} 1 & r & c_s \\ r & 1 & c_t \\ c_s & c_t & 1 \end{pmatrix}, \tag{18}$$

Table 4: Accuracy against measured cost. Recovery is the fraction of the headroom between the retriever and the full cross encoder that a method closes. Speedup is against the cross encoder on the same pool.

Retriever	Method	NDCG@5	Recovery	Latency (ms)	Speedup
ColNomic-7B	Retriever only	60.9	–	60	113×
	GQR	63.8	24.9%	130	52×
	RRF	62.3	11.9%	93	72×
	Score Aggregation	62.2	11.4%	93	72×
	Qwen3-Reranker-0.6B	61.6	6.4%	344	20×
	MonoQwen2-VL	65.4	39.0%	2,860	2×
	Ours (Qwen2.5-VL-7B)	69.0	70.4%	156	44×
	Ours (GLM-4.1V-9B)	71.2	89.4%	141	48×
	Full cross encoder	72.4	100.0%	6,765	1×
ColPali-v1.3	Retriever only	50.2	–	40	166×
	RRF	55.8	31.8%	74	90×
	Score Aggregation	56.8	37.9%	74	90×
	Qwen3-Reranker-0.6B	54.7	25.7%	330	20×
	MonoQwen2-VL	62.6	71.4%	2,808	2×
	Ours (Qwen2.5-VL-7B)	64.9	84.7%	136	49×
	Ours (GLM-4.1V-9B)	66.1	91.3%	122	55×
	Full cross encoder	67.6	100.0%	6,694	1×

whose diagonal is one because each vector is standardized. For any $a \in \mathbb{R}^3$, bilinearity gives $a^\top R\, a = \left\| a_1 z(\mathbf{s}) + a_2 z(\mathbf{t}) + a_3 \mathbf{y} \right\|_K^2 \geq 0$, so R is positive semidefinite and in particular $\det R \geq 0$.

A quadratic in c_s. Expanding along the first row,

$$\det R \;=\; \big(1 - c_t^2\big) - r\big(r - c_t c_s\big) + c_s\big(r c_t - c_s\big) \;=\; 1 - r^2 - c_t^2 - c_s^2 + 2\, r\, c_t\, c_s. \tag{19}$$

Requiring this to be non-negative and changing sign gives

$$c_s^2 \;-\; 2\, r\, c_t\, c_s \;+\; \big(r^2 + c_t^2 - 1\big) \;\leq\; 0. \tag{20}$$

The left hand side is a quadratic in c_s with positive leading coefficient, so the inequality confines c_s to the closed interval between its two roots. Those roots are

$$r\, c_t \;\pm\; \sqrt{r^2 c_t^2 - r^2 - c_t^2 + 1} \;=\; r\, c_t \;\pm\; \sqrt{\big(1 - r^2\big)\big(1 - c_t^2\big)}, \tag{21}$$

the discriminant factoring because $r^2 c_t^2 - r^2 - c_t^2 + 1 = (1 - r^2)(1 - c_t^2)$. The lower root is Equation 11, and the upper root is the matching upper bound.

Monotonicity. Write $g(r) = r\, c_t - \sqrt{(1 - r^2)(1 - c_t^2)}$ for the lower root at fixed c_t. Differentiating the square root through the chain rule,

$$g'(r) \;=\; c_t \;+\; \frac{r\big(1 - c_t^2\big)}{\sqrt{\big(1 - r^2\big)\big(1 - c_t^2\big)}} \;=\; c_t \;+\; r\sqrt{\frac{1 - c_t^2}{1 - r^2}}, \tag{22}$$

whose two terms are both positive for $r, c_t \in (0, 1)$. So g is strictly increasing there, and raising the agreement of the model score with the target raises the floor under its agreement with relevance. □

B Additional Results

Ranking quality under four further metrics and the measured cost of each arm are reported here.

Table 5: The same arms under four further metrics, macro over the corpora of each benchmark. The fusion weight of every arm is set as in Tables 1 and 2 and then held fixed, so none of these four metrics had a part in choosing it.

Benchmark	Retriever	Method	NDCG@5	NDCG@10	R@5	R@10	MRR@10
ViDoRe 2	ColNomic-7B	Retriever only	60.9	63.8	58.0	69.8	71.3
		GQR	63.8	65.2	58.6	68.5	73.8
		RRF	62.3	64.8	57.6	69.9	72.5
		Score Aggregation	62.2	65.1	57.5	69.4	72.5
		Qwen3-Reranker-0.6B	61.6	64.2	58.0	69.0	71.5
		MonoQwen2-VL	65.4	67.9	60.3	72.3	75.7
		Ours (Qwen2.5-VL-7B)	69.0	70.5	63.9	73.8	78.0
		Ours (GLM-4.1V-9B)	71.2	72.2	64.4	73.6	80.4
		Full cross encoder	72.4	73.3	65.3	73.8	82.2
	ColPali-v1.3	Retriever only	50.2	53.3	47.8	60.2	61.5
		RRF	55.8	57.6	52.9	63.5	66.2
		Score Aggregation	56.8	58.7	52.8	63.4	67.3
		Qwen3-Reranker-0.6B	54.7	56.9	52.8	64.0	64.6
		MonoQwen2-VL	62.6	63.7	57.1	66.4	73.3
		Ours (Qwen2.5-VL-7B)	64.9	65.0	59.3	67.0	75.0
		Ours (GLM-4.1V-9B)	66.1	66.3	58.9	66.7	76.5
		Full cross encoder	67.6	67.6	60.9	67.9	77.8
ViDoRe 3	ColNomic-7B	Retriever only	52.8	55.3	48.2	59.8	67.0
		GQR	54.3	56.8	49.4	61.3	68.6
		RRF	51.4	54.6	47.3	60.4	66.2
		Score Aggregation	53.5	56.0	48.8	60.6	67.6
		Qwen3-Reranker-0.6B	54.0	56.7	49.4	61.3	68.3
		MonoQwen2-VL	54.5	56.8	49.7	60.8	68.5
		Ours (Qwen2.5-VL-7B)	56.2	58.2	51.6	62.3	69.4
		Ours (GLM-4.1V-9B)	55.8	57.8	50.5	61.5	69.8
		Full cross encoder	58.0	59.8	52.6	63.2	71.5
	ColPali-v1.3	Retriever only	38.8	41.3	36.4	46.5	51.8
		RRF	42.9	45.6	39.5	50.6	57.3
		Score Aggregation	44.6	46.4	40.8	50.2	59.4
		Qwen3-Reranker-0.6B	47.8	49.0	43.3	51.7	62.0
		MonoQwen2-VL	46.7	48.3	41.9	51.0	61.4
		Ours (Qwen2.5-VL-7B)	48.8	50.0	44.2	52.4	63.1
		Ours (GLM-4.1V-9B)	48.8	49.8	43.8	51.9	63.2
		Full cross encoder	51.1	51.7	45.5	53.0	65.6

B.1 Fusion at Full Depth

Proposition 1 concerns any pair of scores, so it applies to the uncompressed cross encoder as much as to the readout. Table 6 sweeps the fusion weight over the full cross encoder on all 24 subsets. No subset attains its optimum at $w = 1$, the median optimum is 0.80, and ranking by the cross encoder alone loses 1.5 points on average and up to 4.3. Keeping the retriever score in the order is therefore no correction for compression, since it improves the very model the readout reproduces.

Table 6: The full cross encoder fused with the retriever, the optimal weight located per subset on a grid of step 0.05.

Benchmark	Retriever	Subsets	Optimal at $w = 1$	Median optimal w	NDCG@5 at $w = 1$	NDCG@5 at optimum
ViDoRe 2	ColNomic-7B	4	0	0.850	72.0	73.6
ViDoRe 2	ColPali-v1.3	4	0	0.875	67.0	67.9
ViDoRe 3	ColNomic-7B	8	0	0.700	55.9	58.4
ViDoRe 3	ColPali-v1.3	8	0	0.825	50.7	51.5

B.2 Number of Calibration Queries

Figure 8 draws the calibration queries at random from the corpora a fold is fitted on, five draws per size, and scores the held out corpus of the 154 query calibration sample. None of these queries carries a relevance judgement. Sixteen queries already reach 74.3 on GLM and 71.1 on Qwen, 99.4 and 98.3 percent of the full sample, and both backbones stay above the 67.2 that GQR scores on the same queries at every size.

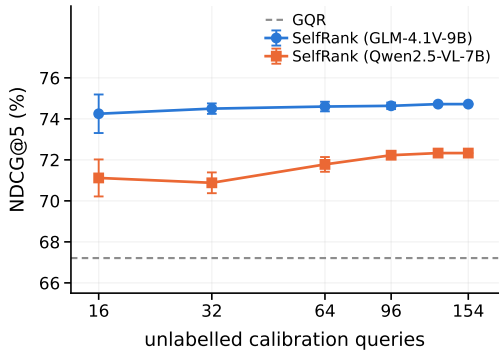

Figure 8: NDCG@5 against the number of unlabelled calibration queries, mean and standard deviation over five draws.

B.3 Label Free Regularization

Table 7 compares relevance judgements with centered teacher error for selecting λ. Within each outer fold, two corpora fit the readout and the third validates it; the modal choice across four folds is deployed. On GLM both rules deploy 10^3, yielding bit identical refitted vectors despite different fold choices. On Qwen the teacher rule selects 10^3 in every fold, scoring 72.3 against 71.3 for the labelled rule. Reported results use the teacher rule.

Table 7: The ridge strength each rule selects on the four folds, the modal value deployed, and NDCG@5 on the calibration sample at that value.

Backbone	Rule	Biomed	Econ	ESG-H	ESG-F	Deployed	NDCG@5
GLM-4.1V-9B	Relevance judgements	10^3	10^4	10^3	10^1	10^3	74.7
	Teacher alone	10^3	10^4	10^3	10^3	10^3	74.7
Qwen2.5-VL-7B	Relevance judgements	10^3	10^4	10^4	10^1	10^4	71.3
	Teacher alone	10^3	10^3	10^3	10^3	10^3	72.3

B.4 Pool Depth for Every Method

Table 8 extends Figure 7 to every reranker of Table 1, SelfRank setting its weight by Equation 4 and each baseline by NDCG@5 on the same inner corpus, at every depth. At every depth from five to a hundred candidates both SelfRank backbones rank above all three external baselines, by at least 5.0 points from ten candidates onward.

B.5 Operating Point Without Relevance Judgements

Proposition 1 requires the correlation of each score with relevance, and Table 9 shows that both can be recovered from the scores alone. Under a one factor model in which the retriever score, the readout and the full depth margin of the uncompressed model each carry relevance plus mutually uncorrelated noise, $c_b = \sqrt{\rho_{bs}\rho_{bt}/\rho_{st}}$ and $c_s = \sqrt{\rho_{bs}\rho_{st}/\rho_{bt}}$, where each ρ is a correlation between two scores and involves no relevance judgement. Read on the unlabelled queries of the corpus

Table 8: NDCG@5 on ViDoRe 2 with ColNomic candidates against the number of candidates reranked. The retriever alone scores 60.9 and GQR 63.8 at every depth.

K	RRF	Score Aggregation	Qwen3-Reranker-0.6B	SelfRank (Qwen2.5-VL-7B)	SelfRank (GLM-4.1V-9B)
5	61.9	61.4	60.8	63.7	65.2
10	62.8	62.2	60.3	67.8	69.5
15	62.6	62.5	60.5	68.9	70.8
20	62.3	62.2	61.6	69.0	71.1
25	62.0	62.5	60.8	69.2	71.1
30	61.9	62.3	61.1	69.4	71.0
40	61.5	62.6	61.1	69.5	71.0
50	61.5	62.6	61.0	69.6	70.8
60	61.6	62.5	61.7	69.6	70.4
70	61.7	62.8	61.7	69.6	70.8
80	61.5	62.9	61.9	69.8	70.6
90	61.5	64.1	62.6	70.1	70.5
100	61.5	64.0	62.1	70.2	70.5

being ranked, this estimate reaches 59.9 on average against 60.1 for the deployed rule, and sixteen unlabelled queries per corpus already give 59.8.

Table 9: NDCG@5 (%) of SelfRank under three ways of setting the fusion weight. Sweep picks w by NDCG@5 on the next corpus of the same benchmark. Proposition 1 reads the three correlations on that corpus with its relevance judgements, as in Tables 1 and 2. Label free reads only correlations between scores, on the unlabelled queries of the corpus being ranked, using all of them or sixteen drawn at random with mean and standard deviation over five draws. The third score is the full depth margin of Qwen2.5-VL-7B.

					Label free	
Benchmark	Retriever	Backbone	Sweep	Proposition 1	All queries	16 queries
ViDoRe 2	ColNomic-7B	Qwen2.5-VL-7B	69.6	69.0	69.4	69.4 $\pm$0.2
		GLM-4.1V-9B	71.3	71.2	70.7	70.2 $\pm$0.4
	ColPali-v1.3	Qwen2.5-VL-7B	65.1	64.9	65.2	65.2 $\pm$0.1
		GLM-4.1V-9B	65.8	66.1	66.3	65.8 $\pm$0.3
ViDoRe 3	ColNomic-7B	Qwen2.5-VL-7B	56.2	56.2	54.7	54.6 $\pm$0.1
		GLM-4.1V-9B	55.5	55.8	55.9	55.8 $\pm$0.1
	ColPali-v1.3	Qwen2.5-VL-7B	48.6	48.8	48.5	48.5 $\pm$0.1
		GLM-4.1V-9B	48.6	48.8	48.8	48.7 $\pm$0.0
Average			60.1	60.1	59.9	59.8