BANDRANK: INTERVAL-ANCHORED LISTMLE FOR THRESHOLD-TRANSFERABLE POINTWISE RERANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pointwise neural rerankers are widely deployed for their $\mathcal{O}(N)$ inference efficiency, yet learning them effectively involves balancing within-query discrimination against cross-query score consistency. Hard pointwise targets discard distinctions between relevance grades, whereas listwise ranking objectives are shift-invariant and lack an absolute anchor, leading to cross-query scale drift that degrades global thresholding. We propose **BandRank**, an interval-anchored ListMLE training framework. BandRank couples ListMLE with a smooth two-sided softplus penalty defined over query-shared score bands, encouraging predictions to remain within grade-specific intervals while ListMLE governs the relative ordering. To evaluate cross-query score reliability, we examine threshold transfer using transferred balanced accuracy (tr. BA) under frozen validation cuts. On the MultiVENT 2.0 video benchmark, BandRank improves on the previous state of the art by 7.6 points in nDCG@5 and 5.1 points in Recall@5, while improving tr. BA over a pointwise BCE baseline by up to 6.1 points. Furthermore, BandRank transfers thresholds stably even under calibration-data scarcity, and the same objective and band layout carry over to visual document (MMDocIR) and text passage (TREC DL) reranking. Code is available at `https://anonymous.4open.science/r/ICLR2027-13167-7F87`.

1 INTRODUCTION

Neural rerankers refine candidates returned by first-stage retrievers in two-stage retrieval systems (Nogueira & Cho, 2019; Karpukhin et al., 2020). Pointwise rerankers score query–candidate pairs independently, enabling parallel $\mathcal{O}(N)$ inference without cross-candidate decoding (Li et al., 2026; Weller et al., 2025). Beyond sorting candidates within a query, their scores also support global thresholding for filtering irrelevant items, selecting retrieval cutoffs, and triggering downstream responses. These applications require a shared threshold to remain reliable across heterogeneous queries, making both *within-query discrimination* and *cross-query score consistency* essential.

Existing training objectives each satisfy only one of these requirements. Hard pointwise targets can compress graded relevance into rigid values and encourage score concentration (Cai et al., 2026), obscuring distinctions such as partial visual overlap versus exact semantic matches. Listwise objectives provide richer relative supervision, but objectives such as ListMLE (Xia et al., 2008) are invariant to additive score shifts and leave the score origin unidentified (Yan et al., 2022). Strong rankings can therefore coexist with query-dependent offsets that undermine global thresholding. Simply combining these objectives is not necessarily sufficient: point targets constrain score gaps as well as the score origin, potentially conflicting with the ranking loss. Prior work identifies incompatible optima for sigmoid and softmax cross-entropy under binary relevance (Bai et al., 2023).

Our key insight is that absolute anchoring need not prescribe an exact score for every candidate. Instead, *query-shared intervals* can establish grade-level score semantics while preserving flexibility for relative ordering. We instantiate this idea in **BandRank**, an interval-anchored ListMLE training framework. ListMLE learns fine-grained candidate orderings, while a smooth two-sided softplus penalty anchors predictions to predefined grade-specific score bands. The anchor has bounded logit-

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.

space gradients and approximately linear growth outside the bands; for the outermost grades, it imposes only a one-sided constraint.

Training supervision combines coarse relevance tiers with fine-grained intra-tier orderings obtained through multi-run vision-language model (VLM) consensus. Both objectives train the same pointwise scorer, preserving standard $\mathcal{O}(N)$ inference. We evaluate score reliability through *transferred balanced accuracy* (tr. BA): a threshold selected on validation queries is frozen and applied to held-out queries. We also test the case where only a few labeled dev queries are available for choosing the threshold.

Our main contributions are as follows:

1. **BandRank Framework:** We propose BandRank, a training framework that couples ListMLE with query-shared interval anchoring via smooth softplus penalties. The anchor identifies the score origin that ListMLE leaves free, placing each relevance tier at a designed location on the absolute scale while preserving single-pass $\mathcal{O}(N)$ inference. In a controlled comparison with target locations matched, the interval anchor combines with ListMLE more favorably than a soft-target point anchor (65.3 vs. 63.2 nDCG@5; 70.4 vs. 68.9 judged tr. BA)
2. **Controllable Operating Regions:** We show that interval anchors allow flexible intra-tier ranking while placing the operating threshold where the band layout predicts. By adjusting the predefined band layout, practitioners can directly steer the operating thresholds without sacrificing ranking fidelity, even under scarce-data calibration regimes.
3. **Cross-Modal Empirical Validation:** We conduct comprehensive experiments across video (MultiVENT 2.0 (Kriz et al., 2025)), visual document (MMDocIR (Dong et al., 2025)), and text retrieval (TREC DL (Craswell et al., 2020; 2021)), and show that BandRank outperforms strong open-weight baselines in ranking metrics (e.g., nDCG) and threshold generalization (tr. BA), supporting its broad cross-modal applicability.

2 Related Work

2.1 Neural Reranking and Supervision Signals

Neural rerankers are commonly formulated as pointwise or listwise (Nogueira & Cho, 2019). Listwise rerankers (Liu et al., 2026; Xu et al., 2025) model inter-candidate interactions but output permutations over a candidate window, carrying no magnitude comparable across queries. Pointwise rerankers (Nogueira & Cho, 2019; Li et al., 2026; Skow et al., 2026; Weller et al., 2025) score candidates independently with parallel $\mathcal{O}(N)$ inference, at the cost of list-level context.

Pointwise rerankers are traditionally trained with binary cross-entropy. Reasoning- and reinforcement-based rerankers (Weller et al., 2025; Liu et al., 2026; Zhuang et al., 2026) improve discrimination on complex queries, but target orderings, selections, or binary judgments rather than a cross-query score scale. For graded relevance, prior work adopts generative integer targets (Cai et al., 2026), teacher-distilled soft ranking (Sun et al., 2026), or softened pointwise targets with distillation and a pairwise term (Skow et al., 2026). Most signals map each grade to a fixed scalar, turning ambiguity between adjacent grades into conflicting supervision and suppressing the intra-tier score variance essential for ranking.

2.2 Calibration and Cross-Query Score Comparability

Classical IR models per-query score distributions to derive probabilities and cut-offs (Manmatha et al., 2001; Arampatzis & Robertson, 2011), while generic monotonic calibrators (Platt, 1999; Guo et al., 2017) apply a single global mapping and cannot correct *query-dependent* drift. In learning-to-rank, Yan et al. (2022) show that listwise objectives are shift-invariant and restore an origin via multi-objective losses or a virtual reference candidate with a fixed score; Bai et al. (2023) show that sigmoid and softmax cross-entropy have incompatible optima and propose ListCE for binary relevance. Industrial systems decouple the two objectives via dual logits (Sheng et al., 2023) or gradient detachment (Zhang et al., 2024). GroupRank (Long et al., 2026) regularizes within-group score distributions, and Ren et al. (2025) calibrate list-view scores with point-view estimates; both remain tied to jointly scored candidates and thus fall outside our frozen-threshold pointwise protocol.

Most closely related to our work, Yu et al. (2025) study scale calibration for neural rankers with graded labels, and Meng et al. (2026) concurrently treat reranker scores as absolute judgments.

Our approach differs in two respects. First, prior anchors regress each score toward a point target, whereas we impose a soft interval per grade that is centered on the band with bounded, saturating gradients outside it, and one-sided for the outermost grades. Second, unlike the single reference candidate of Yan et al. (2022), our anchor applies a query-shared, grade-conditional penalty to every candidate, fixing where each tier lies on the absolute scale while ListMLE governs the ordering. We evaluate frozen-threshold transfer rather than distributional calibration.

3 METHOD

3.1 LISTWISE RANKING DATASET CONSTRUCTION

The training data is constructed upon the extended query set of MultiVENT 2.0 (Kriz et al., 2025; Skow et al., 2026), which expands the original human annotations to a total of 9,267 training queries over 108,500 news videos across six languages.

To mitigate the computational and cognitive burden of ranking the entire candidate pool simultaneously, we construct listwise training data via a two-stage pipeline (Figure 1): coarse-grained tiering followed by fine-grained intra-tier ranking, both annotated by Qwen3.6-27B (Qwen Team, 2026).

Coarse-grained Tiering. We first employ an embedding model (Ma et al., 2025) to retrieve the top 10 candidate videos for each query. Subsequently, Qwen3.6 evaluates the relevance of each query–candidate pair based exclusively on the sampled video frames. We restrict the annotator to frames so that training labels use only the evidence available to the reranker at inference. Following a predefined rubric, each candidate is assigned to one of four grades: **Grade 3** (*Exact Match*), **Grade 2** (*High Visual Deception*), **Grade 1** (*Weakly Related*), and **Grade 0** (*Irrelevant*).

Fine-grained Intra-tier Ranking. To break ties within grades 1–3, we conduct a second, listwise comparative ranking stage over candidate windows of at most five videos. To mitigate the inherent position bias of the VLM annotator, each window is evaluated under five input permutations (the identity order, its reversal, and three random shuffles). We then aggregate the resulting rankings by extracting their longest common subsequence (LCS) to form the final intra-tier rank.

Dataset Statistics. The pipeline yields 8,938 query candidate sets with a total of 98,339 query–video pairs. Each query is paired with 11 candidates on average, typically consisting of one ground-truth video and ten graded negative candidates. Across all evaluated pairs, the label distribution for Grades 0–3 is 45,610, 30,492, 13,277, and 8,960, respectively. More details on the dataset construction are provided in Appendix B.1.

3.2 INTERVAL-ANCHORED LISTMLE TRAINING

The overall objective balances listwise relative ordering with cross-query absolute scale calibration:

$$\mathcal{L} = \lambda_{\text{ListMLE}}\,\mathcal{L}_{\text{ListMLE}} + \lambda_{\text{Anchor}}\,\mathcal{L}_{\text{Anchor}}, \tag{1}$$

where λ_{ListMLE} and λ_{Anchor} are balancing hyperparameters.

Listwise Ordering. For a given query paired with n candidate videos, let $\pi = (\pi_1, \ldots, \pi_n)$ denote the target permutation. This golden order is hierarchically determined: candidates are sorted by descending *grade*, with ties broken by ascending *intra-grade rank*. Let z_i be the predicted raw logit for candidate i; the reported relevance score is $s_i = \sigma(z_i) \in [0, 1]$, obtained through the sigmoid function $\sigma(\cdot)$. Both loss terms act directly on the logits z_i. To learn the target order of candidates, ListMLE minimizes the negative log-likelihood of the ideal permutation π under the Plackett–Luce model:

$$\mathcal{L}_{\text{ListMLE}} = -\frac{1}{n}\sum_{i=1}^{n}\left(z_{\pi_i} - \log\sum_{k=i}^{n}\exp(z_{\pi_k})\right). \tag{2}$$

While Eq. 2 effectively optimizes intra-list ranking, it is shift-invariant. This unconstrained degree of freedom leaves scores uncalibrated and incomparable across queries.

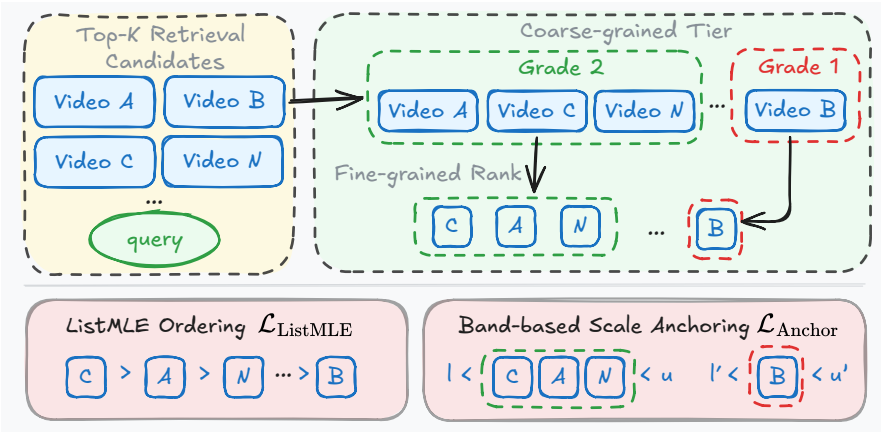


Figure 1: Overview of the training pipeline. Top-K candidates retrieved for each query are hierarchically organized via **(1) Coarse-grained Tiering** and **(2) Fine-grained Intra-tier Ranking**. The resulting golden permutation is supervised by $\mathcal{L}_{\text{ListMLE}}$ for relative intra-list ordering and regularized by $\mathcal{L}_{\text{Anchor}}$ into cross-query, grade-specific score bands.

Band-based Scale Anchoring. To align scores globally, we assign each relevance grade $g \in \{0, 1, 2, 3\}$ a score band $[b_g, b_{g+1}]$ with boundaries $(b_0, \ldots, b_4) = (0,\ 0.4,\ 0.6,\ 0.8,\ 1)$ and penalize out-of-band predictions. The penalty is applied in logit space, $[\ell_g, u_g] = [\sigma^{-1}(b_g),\ \sigma^{-1}(b_{g+1})]$, which avoids the vanishing factor $\sigma'(z)$ in saturated regions and keeps the anchor's scale comparable to $\mathcal{L}_{\text{ListMLE}}$. For a batch of N candidates with logits z_j:

$$\mathcal{L}_{\text{Anchor}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\beta}\Big[\text{softplus}\big(\beta\,(\ell_{g_j} - z_j)\big) + \text{softplus}\big(\beta\,(z_j - u_{g_j})\big)\Big], \tag{3}$$

where $\beta > 0$ controls the sharpness of the band edges ($\beta = 1$ by default). Since $\sigma^{-1}(0) = -\infty$ and $\sigma^{-1}(1) = +\infty$, the outer bounds vanish for Grades 0 and 3, and the anchor reduces to a one-sided barrier that only pushes Grade 0 down and Grade 3 up, leaving their separation to $\mathcal{L}_{\text{ListMLE}}$. For Grades 1 and 2, the anchor gradient vanishes at the logit-space band midpoint and grows smoothly outside the band (see Appendix A). In all cases its magnitude is bounded by 1, so each candidate, including a mislabeled one, contributes at most $\lambda_{\text{Anchor}}/N$ to the gradient. Because the boundaries are shared across queries, each grade is encouraged to occupy the same region of the score scale for every query.

4 Experiments

4.1 Experimental Setup

Benchmarks. Primary evaluations are conducted on MultiVENT 2.0 (Kriz et al., 2025) (2,546 test queries, 109,800 videos), reranking the top-100 candidate pools retrieved by OmniEmbed (Ma et al., 2025). To test generalization to other modalities, we further extend evaluations to visual document retrieval on MMDocIR (Dong et al., 2025) and text passage ranking on TREC DL19/20 (Craswell et al., 2020; 2021); each additional modality is trained separately under the same objective and band layout, with training-data construction and test-set isolation described in Appendix B.3 and benchmark-specific candidate pools and label conventions in Appendix C.4.

Evaluation Metrics and Threshold Transfer Protocol. For conventional intra-query ranking, we report standard nDCG@K and Recall@K. To quantify whether a single cut-off transfers across queries, we establish *transferred balanced accuracy* (tr. BA) alongside threshold-free AUC. Specifically, to prevent calibration leakage, queries are randomly partitioned into dev/test splits ($30/70\%$) over 20 random seeds. On each dev split, an optimal decision boundary is calibrated via $\tau^* = \arg\max_\tau \frac{1}{2}(\text{TPR}(\tau) + \text{TNR}(\tau))$; this τ^* is subsequently frozen and evaluated on the

corresponding test split, with the mean test BA reported. Appendix C.3 details the threshold search, query eligibility, and calibration protocols.

Our calibration benchmark focuses on the lower **Grade 2 vs. Grade 0** boundary (partially relevant vs. irrelevant),[1] with negatives drawn from the OmniEmbed top-100 pools: *Judged Negatives* (n=974 human-annotated Grade 0) and *Unjudged Negatives* (Grade 0 plus all unlabeled top-100 candidates, $n\approx$245K), the latter reflecting the realistic deployment setting where a threshold filters the full, largely unlabeled candidate list.

Implementation Details. For multimodal benchmarks (MultiVENT 2.0 and MMDocIR), we use Qwen3-VL-8B-Instruct (Qwen Team, 2025b), and Qwen3-8B (Qwen Team, 2025a) for TREC DL. Following the tiered data in Sec. 3.1, each training ranking list comprises one golden candidate (Grade 3) and seven lower-tier candidates (Grades 0–2) sampled from the candidate pool. All models are fine-tuned using LoRA (Hu et al., 2022) on attention and MLP projections. We fix the objective trade-off weights to $\lambda_{\text{ListMLE}} = \lambda_{\text{Anchor}} = 1.0$ and barrier steepness to $\beta = 1$ (sensitivity analysis in Appendix D.1), with band boundaries set to $(b_1, b_2, b_3) = (0.4, 0.6, 0.8)$. Full training configurations and scoring templates are detailed in Appendix C.1.

4.2 Baselines

We evaluate two representative models on the same OmniEmbed top-100 candidates: Qwen3-VL-Reranker-8B (Li et al., 2026), serving as a strong zero-shot reference, and RankVideo-8B (Skow et al., 2026), the previous supervised state-of-the-art on MultiVENT 2.0.

To eliminate discrepancies in candidate pool capacity and isolate the gain of our training objective, we construct a pointwise counterpart trained under identical settings to BandRank. It uses the exact same candidate pools, differing solely in optimizing the standard binary cross-entropy loss by binarizing all seven lower-tier candidates as negative (0). This model serves as the direct reference for our improvements in Table 2 and the ablation in Sec. 4.3.3. The score extraction protocols and metric applicability for published baselines are detailed in Appendix C.2.

4.3 Main Results on MultiVENT 2.0

4.3.1 Ranking Quality

As shown in Table 1, BandRank achieves new state-of-the-art ranking performance on MultiVENT 2.0, consistently outperforming all baselines across every evaluation cutoff.

Table 1: Within-query ranking quality on MultiVENT 2.0. Best results are in **bold**.

		nDCG@K (↑)					**Recall@K (↑)**				
	Model	1	5	10	20	50	1	5	10	20	50
Dense	OmniEmbed	50.5	50.1	49.5	52.4	55.4	23.9	43.5	52.3	59.8	69.0
Reranker	Qwen3-VL-Reranker-8B	42.2	45.0	47.2	52.3	54.7	20.8	41.0	53.7	66.2	73.2
	RankVideo-8B	58.6	57.7	57.0	59.7	61.6	26.7	48.7	59.4	66.8	72.6
	Pointwise BCE	61.9	59.9	58.5	60.9	62.6	27.8	49.4	59.4	65.9	71.5
	BandRank	**66.4**	**65.3**	**63.6**	**65.7**	**66.8**	**30.7**	**53.8**	**64.3**	**70.3**	**73.9**

Compared to the prior supervised SOTA RankVideo-8B, BandRank demonstrates decisive advantages, boosting top-tier precision by $+7.8$ nDCG@1 and Recall@10 by $+4.9$. More importantly, against the controlled Pointwise BCE baseline under identical candidate pools, our method delivers substantial margins across the board (e.g., $+4.5$ on nDCG@1 and $+5.4$ on nDCG@5). These gains substantiate that our joint listwise-anchor formulation fosters sharper discriminative capability at early ranks while maintaining robust coverage across deeper positions.

[1] Grade 3 is well separated from irrelevant candidates (Appendix D.3), while Grade 1 is omitted from boundary tuning due to severe annotation scarcity (only 112 instances in top-100 pools) and high label ambiguity.

4.3.2 CROSS-QUERY THRESHOLD TRANSFER

Table 2: Cross-query threshold transfer and tier separability on MultiVENT 2.0. The lower boundary (Grade 2 vs. irrelevant) is evaluated using judged Grade 0 negatives and unjudged candidates treated as negatives. The upper boundary (Grade 3 vs. Grade 2) evaluates separation between two already-relevant tiers. n counts negatives pooled over the top-100 candidates of all 2,546 queries. Metrics are in %; best in **bold**.

	Lower: G2 vs. G0				**Upper: G3 vs. G2**	
	(judged, $n = 974$)		(unjudged, $n \approx 245$K)			
Model	tr. BA	AUC	tr. BA	AUC	tr. BA	AUC
Pointwise BCE	68.3	74.0	79.5	85.2	**67.5**	**73.2**
Qwen3-VL-Reranker-8B	**76.1**	**81.7**	82.9	89.7	55.6	60.2
RankVideo-8B	63.3	69.4	77.0	84.5	66.0	72.0
BandRank	70.4	77.5	**85.6**	**92.3**	66.8	72.0

Beyond within-query ordering, we evaluate the cross-query tier separability to examine score consistency across different queries (Table 2). We highlight three major findings:

Consistent gains across negative distributions. BandRank achieves the highest lower-boundary performance on the unjudged pool (85.6% tr. BA, 92.3% AUC), outperforming pointwise BCE, Qwen3-VL-Reranker-8B, and RankVideo-8B by 6.1, 2.7, and 8.6 tr. BA points. Its gain over BCE holds across all 20 query splits under both conventions (+2.1 judged, +6.1 unjudged). These settings offer complementary views: judged negatives form a sparse human-verified subset (10.1% of human labels), whereas the unjudged setting spans the full pool, treating unlabeled items as negatives. Model ranking reflects this difference: Qwen3-VL-Reranker-8B leads on judged data (76.1% tr. BA), while BandRank leads on the unjudged pool.

Ranking quality does not guarantee threshold transfer. Within-query ranking metrics are invariant to query-specific monotone transformations of the scores and therefore cannot reveal whether a tier lands at the same absolute score across queries. Tables 1 and 2 exhibit this dissociation directly: although RankVideo-8B outperforms Qwen3-VL-Reranker-8B by 12.7 nDCG@5 points (57.7 vs. 45.0), it substantially trails it on lower-boundary transfer under both negative conventions (63.3 vs. 76.1 judged, 77.0 vs. 82.9 unjudged tr. BA). Notably, despite its more modest nDCG@5 score (45.0), Qwen3-VL-Reranker-8B attains the highest judged-negative transfer across all compared models (76.1 tr. BA). Ranking-only evaluation therefore cannot certify that a reranker's scores support a shared cut-off across queries.

Boundary-specific performance. Lower-boundary gains do not fully transfer to the upper boundary: BandRank achieves 85.6% tr. BA on unjudged negatives, but its upper-boundary score is slightly below BCE (66.8% vs. 67.5%). This matches the band layout: Grades 0 and 2 are buffered by the Grade-1 interval $[0.4, 0.6]$, whereas Grades 2 and 3 share the boundary at 0.8. Correspondingly, score-distribution diagnostics reveal greater lower-tier and smaller upper-tier separation than BCE (Appendix D.2). On the coarser Grade-3-vs-unjudged boundary, BandRank still leads with 90.1% tr. BA, compared to 88.3% for BCE and $\leq$ 87.8% for external rerankers (Appendix D.3). Thus, the shortfall against BCE is confined solely to separating Grades 3 and 2.

4.3.3 ABLATION AND MECHANISM ANALYSIS

We conduct ablation studies on MultiVENT 2.0 to dissect the individual contributions of our loss formulations and investigate the underlying calibration mechanisms.

Complementary Effects of the Objectives. As shown in Table 3, isolated ListMLE substantially boosts permutation quality over BCE (64.5 vs. 59.9 nDCG@5), but its shift-invariance leaves the global score scale unconstrained. The band anchor alone exhibits the opposite profile: with no ranking objective at all, its transfer nearly matches ListMLE (69.3/84.0 vs. 69.9/84.4 judged/unjudged tr. BA), i.e., it provides a usable global scale, although its ranking performance remains below the

Table 3: Ablation of loss components on MultiVENT 2.0. Best results are in **bold**.

BCE	ListMLE	Anchor	Lower (tr. BA)		Upper (tr. BA)	Ranking	
			Judged	Unjudg.		nDCG@5	R@5
✓	–	–	68.3	79.5	**67.5**	59.9	49.4
–	✓	–	69.9	84.4	66.1	64.5	53.0
–	–	Band	69.3	84.0	67.0	64.0	53.1
–	–	Point	68.9	83.1	66.0	63.9	52.8
–	✓	Point	68.9	83.3	66.1	63.2	52.0
–	✓	**Band**	**70.4**	**85.6**	66.8	**65.3**	**53.8**

joint model (64.0 vs. 65.3 nDCG@5). Combining the objectives yields the best ranking (65.3 nDCG@5) and transfer (70.4/85.6 judged/unjudged tr. BA), conceding only 0.2 points to the band anchor alone on the upper boundary. Over ListMLE alone, the judged gain is modest (+0.5); the anchor mainly fixes *where* the threshold lies ($\tau^* = 0.516$, Table 12).

Band vs. Point Anchoring. To examine how the choice of anchoring loss affects joint optimization with ListMLE, rows 4–5 replace the band anchor with a point anchor based on soft-target cross-entropy:

$$\mathcal{L}_{\text{point}} = -\frac{1}{N}\sum_{j=1}^{N}\left[t_{g_j}\log\sigma(z_j) + (1-t_{g_j})\log(1-\sigma(z_j))\right]. \tag{4}$$

We set $(t_0, t_1, t_2, t_3) \approx (0,\ 0.5,\ 0.7101,\ 1)$ to match the band anchor's minima for Grades 1–2 and limiting optima for Grades 0 and 3, while keeping candidate-level averaging and λ_{Anchor} unchanged.

Adding the point anchor to ListMLE reduces nDCG@5 from 64.5 to 63.2 and judged tr. BA from 69.9 to 68.9, whereas the band anchor improves both to 65.3 and 70.4. The two anchors rank comparably alone (63.9 vs. 64.0 nDCG@5); the gap arises only when combined with ListMLE, which lowers the point anchor by 0.7 but raises the band anchor by 1.3. At the tested weight, band anchoring is therefore more compatible with ListMLE, even when the point targets are aligned with the band anchor's optima. This comparison controls for target placement, but not gradient shape or effective regularization strength (Appendix A).

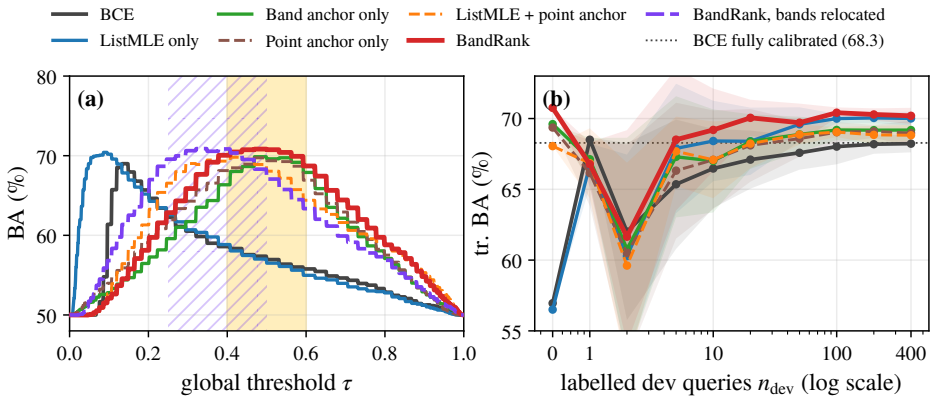


Figure 2: **(a)** BA vs. global τ on the lower boundary (G2 vs. judged G0; all queries, no tuning). Shaded: designed G1 band $[0.4, 0.6]$; purple dash-dotted: BandRank with bands at $(0.25, 0.5, 0.75)$ (hatched). **(b)** tr. BA with τ tuned on n_{dev} queries (20 splits; shaded: ± 1 std); n_{dev}=0: τ=0.5; dotted: BCE with 30% dev (reference).

Threshold Robustness and Limited-Data Calibration. Figure 2(a) shows that anchoring moves the BA peak into the designed band: BandRank maintains $\geq$ 68.7% BA over $[0.4, 0.6]$, whereas

BCE and ListMLE alone peak near $\tau = 0.1$ and fall to 57% at $\tau = 0.5$. With only 10 queries (Figure 2(b)), BandRank achieves 69.2% tr. BA, compared with 66.5% for BCE at the same budget and 68.3% for BCE using the full development split (30% of queries). Without threshold tuning, the default $\tau = 0.5$ yields 70.8% BA. These results support the utility of the learned score scale under limited calibration data.

Controllable Operating Boundaries. Beyond robustness to threshold choice, we examine whether the location of the operating region can be adjusted through the band layout. Changing the predefined boundaries from $(0.4, 0.6, 0.8)$ to $(0.25, 0.5, 0.75)$ shifts the high-BA region toward lower thresholds (Figure 2(a)), while preserving ranking quality (65.3 nDCG@5 for both layouts). Although only two layouts were tested, this comparison suggests that the band layout provides a way to adjust the operating region without sacrificing ranking quality. The dev-selected threshold confirms that the operating point is designed rather than discovered: BandRank selects $\tau^* = 0.516$, within 0.016 of the band midpoint, whereas BCE, ListMLE only, and Qwen3-VL-Reranker-8B land at 0.137, 0.086, and 0.125. What the anchor fixes is this *location*—the across-split dispersion of τ^* is already small for every fine-tuned model and, once normalized by tier separation, does not favor our objective. The same frozen threshold also transfers to held-out languages, held-out query types, and a five-fold reduction in pool depth. Appendix D.4 reports these analyses and the scope they do not cover; additional limited-data calibration results are in Appendix D.3.

Test-time query-wise normalization is complementary, not a replacement. A natural alternative to training-time anchoring is to remove each query's offset at inference using pool statistics, as in metasearch score normalization and ranked-list truncation (Montague & Aslam, 2001; Arampatzis et al., 2009). On MultiVENT, mean centering helps both models: it raises BandRank from 70.4 to 73.5 and ListMLE only from 69.9 to 71.4, so a normalized ListMLE baseline does exceed raw BandRank on this boundary, while the gap between the two widens from 0.5 to 2.1 once both are normalized. The mechanisms nonetheless differ. Applying a global τ after mean centering is equivalent to an *adaptive* threshold $\tau + \mu_{\text{pool}}$ on raw scores, so each decision depends on which other candidates were retrieved, and the approach does not apply when items must be judged individually. Its effect is also benchmark-dependent: $+4.7$ tr. BA on MMDocIR but -12.4 on TREC DL, consistent with the far more variable per-query positive rate on TREC DL (coefficient of variation 1.58 vs. 0.61). Query-wise normalization is therefore best treated as an optional, per-benchmark add-on to an anchored scale rather than a substitute for it (Appendix E).

Table 4: Cross-modal ranking performance. **(a)** MMDocIR top-20 reranking (macro-averaged over 10 domains). **(b)** TREC DL passage reranking on BM25 top-100 (nDCG@10). Qwen3-VL-Rer.-8B and Qwen3-Rer.-8B denote Qwen3-VL-Reranker-8B and Qwen3-Reranker-8B, respectively. Best open results are in **bold**.

(a) Visual Document Retrieval (MMDocIR)

	$\text{DSE}_{\textit{wiki-ss}}$			ColQwen2		
Model	R@1	R@3	R@5	R@1	R@3	R@5
First stage						
Dense retriever	46.0	69.5	77.1	60.4	79.0	84.1
Listwise						
UniME	51.2	70.9	78.0	59.3	79.8	84.7
LamRA	63.4	77.6	82.7	65.4	82.0	86.7
MM-R5	66.1	79.1	83.8	68.5	83.0	86.6
ZipRerank	64.2	84.8	89.0	66.1	84.9	89.3
Pointwise						
LamRA	62.2	82.4	87.1	62.4	82.9	87.2
Qwen3-VL-Rer.-8B	**69.4**	**86.1**	**89.5**	**70.2**	**86.9**	**90.4**
BandRank	66.1	85.5	89.2	66.9	86.0	89.6

(b) Text Passage Ranking (TREC DL)

Model	DL19	DL20	Avg.
First stage			
BM25	50.6	48.0	49.3
Listwise			
Rank-R1	72.2	67.7	70.0
Rearank	74.8	70.1	72.5
Pointwise			
JudgeRank	65.4	59.9	62.6
Rank1	69.0	65.1	67.1
Qwen3-Rer.-8B	73.9	67.4	70.7
ERank-32B	70.8	65.5	68.1
BandRank	**74.9**	**70.4**	**72.6**

4.4 Generalization across Modalities

We evaluate BandRank on MMDocIR for visual document retrieval and TREC DL19/20 for passage ranking, keeping the training objective and band boundaries unchanged. On MMDocIR, following ZipRerank (Sun et al., 2026), we rerank the top-20 pages from two retrievers (Ma et al., 2024; Faysse et al., 2025) and compare against visual rerankers including UniME (Gu et al., 2025), LamRA (Liu et al., 2025), MM-R5 (Xu et al., 2025), ZipRerank, and Qwen3-VL-Reranker-8B. On TREC DL, we rerank BM25's top-100 passages and compare against open-source passage rerankers including Rank-R1-7B (Zhuang et al., 2026), Rearank-7B (Zhang et al., 2025a), JudgeRank-8B (Niu et al., 2024), Rank1-7B (Weller et al., 2025), ERank (Cai et al., 2026), and Qwen3-Reranker-8B (Zhang et al., 2025b).

Ranking Performance. On MMDocIR (Table 4(a)), BandRank outperforms all open listwise rerankers at R@3 and R@5 under both retrievers. With $\mathrm{DSE}_{\textit{wiki-ss}}$, it surpasses ZipRerank by 1.9 R@1 and MM-R5 by 6.4 R@3 / 5.4 R@5 points. While the same-size Qwen3-VL-Reranker-8B has higher Recall (up to $+3.3$ R@1), BandRank yields higher tr. BA (Table 5). On TREC DL (Table 4(b)), BandRank averages 72.6 nDCG@10, beating ERank-32B by 4.5 and Qwen3-Reranker-8B by 1.9 points, and slightly tops Rearank-7B without test-time reasoning.

Cross-Query Threshold Transfer. Table 5 evaluates threshold transfer across modalities using the protocol in Sec. 4.1 (comparing score-producing pointwise rerankers; score extraction details in Appendix C.2). On MMDocIR, golden pages are positives and all remaining in-document candidates are negatives. On TREC DL, we evaluate Grade 2 vs. Grade 0, treating unjudged passages as negatives. BandRank achieves the highest tr. BA on all evaluated settings, even where AUC is tied or slightly lower (90.2 vs. 90.5 on MMDocIR-DSE; 93.4 vs. 94.0 on TREC DL). The two metrics therefore come apart: tr. BA asks whether one frozen cut transfers, whereas AUC averages over all thresholds. On MMDocIR, it improves tr. BA over Qwen3-VL-Reranker-8B by 1.6–1.8 points, winning all 20 query splits under both retrievers. On TREC DL, it exceeds Qwen3-Reranker-8B by 0.4 points ($13/20$ splits) and ERank-14B by 2.9 points; the advantage over ERank also persists when both models are scored only on candidates for which ERank emits a parseable generation. Together, these results demonstrate that BandRank maintains robust cross-query threshold transfer across video, visual document, and text retrieval using identical band boundaries, while retaining competitive ranking quality.

Table 5: tr. BA (%) under the protocol of Sec. 4.1 (20 seeded $30\%/70\%$ splits; dev-frozen threshold). Best open results in **bold**.

Dataset	Model	tr. BA	AUC
MMDocIR ($\mathrm{DSE}_{\textit{wiki-ss}}$)	LamRA	78.8	86.9
	Qwen3-VL-Rer.-8B	81.1	**90.5**
	BandRank	**82.7**	90.2
MMDocIR (ColQwen2)	LamRA	78.6	86.8
	Qwen3-VL-Rer.-8B	81.4	**90.7**
	BandRank	**83.2**	**90.7**
TREC DL 19+20	BM25	58.3	64.4
	ERank-14B	84.5	91.2
	Rank1-7B	84.6	93.0
	Qwen3-Rer.-8B	87.0	**94.0**
	BandRank	**87.4**	93.4

5 Conclusion

In this paper, we studied the tension between within-query discrimination and cross-query score consistency in pointwise rerankers, and proposed **BandRank**, which couples ListMLE with a query-shared, grade-conditional softplus interval anchor at unchanged $\mathcal{O}(N)$ inference cost. Because the bands are the same for every query, the anchor removes the per-query score offset that ListMLE alone cannot determine, while ListMLE still decides the order within each list. On MultiVENT 2.0, BandRank sets a new state of the art in ranking quality, improves frozen-threshold transfer over a controlled BCE baseline on every query split, and selects its operating threshold where the band layout predicts, which makes an untuned default threshold usable and ten calibration queries sufficient. With the same objective and band layout, it achieves the highest tr. BA among evaluated pointwise rerankers on MMDocIR and TREC DL, with ranking quality competitive on MMDocIR and state of the art among compared open rerankers on TREC DL.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic training supervision: a frozen Qwen3.6-27B annotator assigns coarse relevance tiers and intra-tier rankings (Sec. 3.1, Appendix B); these labels are used only for training, never for evaluation. We also used AI assistants to provide feedback on the manuscript, including checking the consistency of reported numbers and the scope of claims against our results. We have not used generative AI tools to develop the conceptual framework, formulate mathematical claims, or propose hypotheses. Additionally, we used generative AI tools to edit the paper for readability and to assist with figure rendering and code editing. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work fine-tunes rerankers on public retrieval benchmarks (MultiVENT 2.0, MMDocIR, TREC DL, and MS MARCO training queries) and does not collect new human-subject data. Evaluation labels are the official human judgments released with each benchmark; the VLM annotator is used only to manufacture training supervision. The MultiVENT annotator sees sampled video frames and no titles, transcripts, or other metadata, which limits exposure of personal text while also inducing a visual-only rubric shift relative to the human judgments. We release training and evaluation code at the anonymous URL given in the abstract; we do not redistribute the underlying video, document, or passage corpora, which remain available from their original providers under their existing licenses.

REPRODUCIBILITY STATEMENT

An anonymous implementation, including training and evaluation scripts, is linked in the abstract. Sec. 3.2 states the objective and the scoring parameterization; Appendix B documents candidate construction, the two-stage annotation pipeline, and the per-modality training corpora; Appendix C records optimizer settings (Table 6), the scoring template, the threshold-transfer protocol, and the per-benchmark evaluation instantiation; Appendix A derives the joint gradients; and Appendix D reports the sensitivity, geometry, and robustness checks referenced in the main text.

REFERENCES

Avi Arampatzis and Stephen Robertson. Modeling Score Distributions in Information Retrieval. *Information Retrieval*, 14(1):26–46, 2011.

Avi Arampatzis, Jaap Kamps, and Stephen Robertson. Where to Stop Reading a Ranked List?: Threshold Optimization using Truncated Score Distributions. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 524–531. ACM, 2009.

Aijun Bai, Rolf Jagerman, Zhen Qin, Le Yan, Pratyush Kar, Bing-Rong Lin, Xuanhui Wang, Michael Bendersky, and Marc Najork. Regression Compatible Listwise Objectives for Calibrated Ranking with Binary Relevance. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 4502–4508, 2023.

Yuzheng Cai, Yanzhao Zhang, Dingkun Long, Mingxin Li, Pengjun Xie, and Weiguo Zheng. ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30121–30129, 2026.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019)*, volume 500-331 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2020 deep learning track. In *Proceedings of the Twenty-Ninth Text REtrieval Confer-*

ence (TREC 2020), volume 500-332 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2021.

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 (EMNLP)*, pp. 30971–31005. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1576.

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.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2860–2869, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Eugene Yang, and Benjamin Van Durme. MultiVENT 2.0: A Massive Multilingual Benchmark for Event-Centric Video Retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24149–24158. IEEE, 2025.

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

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. ReasonRank: Empowering Passage Ranking with Strong Reasoning Ability. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22007–22031, 2026.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025. IEEE, 2025.

Meixiu Long, Duolin Sun, Dan Yang, Yihan Jiao, Lei Liu, Jiahai Wang, BinBin Hu, Yue Shen, Jie Feng, Zhehao Tan, et al. GroupRank: A Groupwise Paradigm for Effective and Efficient Passage Reranking with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25165–25186, 2026.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying Multimodal Retrieval via Document Screenshot Embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505, 2024.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. Tevatron 2.0: Unified Document Retrieval Toolkit across Scale, Language, and Modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4061–4065, 2025.

Raghavan Manmatha, Toni Rath, and Fangfang Feng. Modeling Score Distributions for Combining the Outputs of Search Engines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–275, 2001.

Chuan Meng, Jiqun Liu, Mohammad Aliannejadi, Fengran Mo, Jeff Dalton, and Maarten de Rijke. Re-Rankers as Relevance Judges. *arXiv preprint arXiv:2601.04455*, 2026.

Mark H. Montague and Javed A. Aslam. Relevance Score Normalization for Metasearch. In *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management*, pp. 427–433. ACM, 2001.

Tong Niu, Shafiq Joty, Ye Liu, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. JudgeRank: Leveraging Large Language Models for Reasoning-Intensive Reranking. *arXiv preprint arXiv:2411.00142*, 2024.

Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

John Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Qwen Team. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a.

Qwen Team. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025b.

Qwen Team. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. Self-Calibrated Listwise Reranking with Large Language Models. In *Proceedings of the ACM on Web Conference 2025*, pp. 3692–3701, 2025.

Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. Joint Optimization of Ranking and Calibration with Contextualized Hybrid Model. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4813–4822, 2023.

Tyler Skow, Alexander Martin, Benjamin Van Durme, Rama Chellappa, and Reno Kriz. RANKVIDEO: Reasoning Reranking for Text-to-Video Retrieval. *arXiv preprint arXiv:2602.02444*, 2026.

Yiqun Sun, Pengfei Wei, and Lawrence B. Hsieh. Very Efficient Listwise Multimodal Reranking for Long Documents. In *Forty-third International Conference on Machine Learning*, 2026.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-Time Compute for Reranking in Information Retrieval. In *Second Conference on Language Modeling*, 2025.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise Approach to Learning to Rank: Theory and Algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. MM-R5: MultiModal Reasoning-Enhanced ReRanker via Reinforcement Learning for Document Retrieval. *arXiv preprint arXiv:2506.12364*, 2025.

Le Yan, Zhen Qin, Xuanhui Wang, Michael Bendersky, and Marc Najork. Scale Calibration of Deep Ranking Models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4300–4309, 2022.

Puxuan Yu, Daniel Cohen, Hemank Lamba, Joel Tetreault, and Alejandro Jaimes. Explain then Rank: Scale Calibration of Neural Rankers Using Natural Language Explanations from Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22716–22730, 2025.

Le Zhang, Bo Wang, Xipeng Qiu, Siva Reddy, and Aishwarya Agrawal. REARANK: Reasoning Re-ranking Agent via Reinforcement Learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2458–2471, 2025a.

Shunyu Zhang, Hu Liu, Wentian Bao, Enyun Yu, and Yang Song. A Self-boosted Framework for Calibrated Ranking. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6226–6235, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*, 2025b.

Shengyao Zhuang, Xueguang Ma, Zheng Yao, Shuai Wang, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4419–4425, 2026.

A Analysis of the Training Objective

ListMLE encourages relative ordering, while the interval anchor controls grade-dependent score location. We analyze both objectives with respect to raw logits z_j; reported scores are $s_j = \sigma(z_j)$. For a query with n candidates indexed by target order, define the Plackett–Luce selection probabilities

$$p_{ij} = \frac{\exp(z_j)}{\sum_{k=i}^{n} \exp(z_k)}, \qquad i \le j. \tag{5}$$

ListMLE: Relative Ordering. Differentiating Eq. 2 gives

$$\frac{\partial \mathcal{L}_{\text{ListMLE}}}{\partial z_j} = \frac{1}{n}\left(\sum_{i=1}^{j} p_{ij} - 1\right). \tag{6}$$

For $n > 1$, gradient descent promotes the first-ranked candidate and suppresses the last-ranked candidate. Intermediate candidates balance suppression at earlier selection stages against promotion at their own stage. The gradient direction therefore depends on target position and competing logits, rather than grade alone. Since $\sum_j \partial \mathcal{L}_{\text{ListMLE}}/\partial z_j = 0$, ListMLE is invariant to a common logit shift and does not fix an absolute score origin.

Anchor: Grade-Dependent Score Location. Let $\ell_g = \text{logit}(b_g)$ and $u_g = \text{logit}(b_{g+1})$. Writing the per-candidate anchor loss as $a_g(z)$, its derivative is

$$a_g'(z) = \sigma\big(\beta(z - u_g)\big) - \sigma\big(\beta(\ell_g - z)\big). \tag{7}$$

For finite endpoints, this derivative is negative below $m_g = (\ell_g + u_g)/2$ and positive above it. The unique minimum is therefore at the logit-space midpoint m_g, corresponding to score $\sigma(m_g)$. The derivative is bounded in magnitude by one and saturates far outside the band.

At finite β, the anchor retains a midpoint preference even inside the interval; it is not a zero-penalty region. As $\beta \to \infty$, the loss approaches

$$a_g(z) \longrightarrow \max(0, \ell_g - z) + \max(0, z - u_g), \tag{8}$$

which is flat inside the band and linear outside.

For the outer grades, the infinite-boundary term vanishes:

$$a_0'(z) = \sigma\big(\beta(z - u_0)\big) > 0, \qquad a_3'(z) = -\sigma\big(\beta(\ell_3 - z)\big) < 0 \tag{9}$$

for finite z. Thus, Grade 0 is pushed downward and Grade 3 upward, with loss infima approached as scores tend to 0 and 1, respectively. The inner bands provide two-sided location control, whereas the outer bands impose one-sided penalties.

Joint Optimization. For a query-level objective with both terms averaged over its n candidates, the combined gradient is

$$\frac{\partial \mathcal{L}_q}{\partial z_j} = \frac{1}{n}\left[\lambda_{\text{ListMLE}}\left(\sum_{i=1}^{j} p_{ij} - 1\right) + \lambda_{\text{Anchor}}\, a_{g_j}'(z_j)\right]. \tag{10}$$

The ListMLE term depends on relative position and competing logits; the anchor term depends on grade and absolute logit location. The terms reinforce each other when they favor the same adjustment and compete otherwise. Within-grade score variation therefore reflects a balance between ranking pressure and the anchor's location preference.

Band-Aligned Point Anchor: Gradient Comparison. For the point loss in Eq. 4, the intermediate-grade targets are constructed as $t_g = \sigma(m_g)$ for $g \in \{1, 2\}$, where $m_g = (\ell_g + u_g)/2$. The outer-grade targets are $t_0 = 0$ and $t_3 = 1$. These endpoints are left unclipped, and cross-entropy is evaluated in logit space for numerical stability.

Writing the per-candidate point loss as $c_g(z)$, its derivative is

$$c_g'(z) = \sigma(z) - t_g. \tag{11}$$

For Grades 1–2, this derivative vanishes at $z = \text{logit}(t_g) = m_g$, matching the band anchor's minimum. For Grade 0, $c_0'(z) = \sigma(z) > 0$; for Grade 3, $c_3'(z) = \sigma(z) - 1 < 0$. Both anchors therefore have the same one-sided directions and limiting optima for the outer grades. In particular, neither anchor pulls low-scoring Grade-0 candidates upward.

B Training Data and Annotation

B.1 MultiVENT candidate construction and label processing

Candidate pool construction. The starting point is the pointwise training data of RankVideo (Skow et al., 2026) over the extended MultiVENT 2.0 training queries, which provides for each query one or more golden positive videos and a set of mined hard negatives. We keep the 8,938 queries that have at least one golden positive, and supplement each query with additional negatives from the OmniEmbed (Ma et al., 2025) top-100 retrieval list in descending retrieval score, skipping documents already present, until every query has ten negatives. This yields 89,379 query–negative pairs sent to the annotator; the negatives mix the pre-mined hard negatives with freshly retrieved candidates.

Coarse-grained tiering. Each of the 89,379 query–candidate pairs is graded by Qwen3.6-27B (Qwen Team, 2026) in an independent pointwise call. The annotator sees only sampled video frames and assigns one of nine tiers (T0–T8) under the rubric in Figure 3, which map onto the four grades as follows: grade 3 (*Exact Match*) = T0; grade 2 (*High Visual Deception*, hard negatives) = T1–T3; grade 1 (*Weakly Related*, soft negatives) = T4–T6; grade 0 (*Irrelevant*) = T7–T8. The nine-tier design forces the annotator to first localize *why* a video is deceptive (same subjects but a different occurrence; same action but different subjects; same event but the target action missing) before committing to a grade, which we found to produce substantially more hard negatives (grade 2) than directly asking for a four-way judgement. Golden positives are not re-annotated and are directly assigned grade 3.

Fine-grained intra-tier ranking. Ties within each tier are broken by a second, listwise annotation stage (Figure 4). Three design choices are worth noting. *(i) Which tiers are ranked.* Only grades 1–3 enter this stage; grade 0 videos keep their retrieval order, since ordering among irrelevant items carries no training signal. *(ii) Reclassification of annotator grade 3.* Manual inspection showed that negatives the annotator placed in grade 3 are overwhelmingly strong matches rather than exact matches, so all such negatives are reclassified as grade 2. *(iii) Windowing.* Within each (query, grade) group, candidates are sorted by retrieval rank, truncated to the top five, and ranked in windows of at most five videos labeled A–E.

To mitigate the position bias of the annotator, every window is ranked under five input permutations—the identity order, its reversal, and three random shuffles from a fixed seed. The five resulting rankings are then aggregated by extracting their longest common subsequence to form the definitive intra-tier rank, preserving the relative order consistently agreed upon across permutations. Failed runs are discarded before aggregation; 99.2% of windows have all five runs valid, and the five permutations show moderate-to-high consistency with a mean pairwise Spearman correlation of 0.64 (median 0.75).

Rank-list assembly. The two stages are merged into one rank list per query: candidates are ordered by descending grade, and within a grade by the fine-grained order where available, falling back to retrieval rank otherwise (all of grade 0 and groups with a single candidate). Because a few event queries carry many golden videos (e.g., up to 19 for *2022 Russian invasion of Ukraine*), the number of grade-3 positives per query is capped at three, keeping the fine-grained order. The final dataset contains 8,938 rank lists and 98,339 query–video pairs (11.0 candidates per query on average, typically one golden video plus ten graded negatives), with grade distribution 45,610 / 30,492 / 13,277 / 8,960 for grades 0–3.

B.2 Agreement with Human Annotations

To check whether the two-stage annotation pipeline agrees with human judgments, we apply it to the official MultiVENT 2.0 test judgments (2,546 queries, 16,116 query–video pairs graded on the same 0–3 scale), with the annotator blind to the human labels. These annotations are used only for this analysis and never for training or evaluating rerankers. As in Appendix B.1, Qwen3.6-27B first grades each pair independently, then breaks within-grade ties in windows of at most five videos, each ranked under five input permutations and aggregated by their longest common subsequence.

Ordering by grade and then by intra-grade rank gives a machine ranking of each query's judged videos, which we compare with the ranking implied by the human grades.

We evaluate alignment using pairwise concordance, defined as the proportion of video pairs (within the same query and having different human grades) where the machine ranking agrees with human preference; ties in machine predictions receive a score of $\frac{1}{2}$ (chance level: 50%). On this metric, our full pipeline achieves an 80.1% concordance. Across the 2,059 queries with at least two distinct human grades, it yields an average tie-corrected Kendall's τ_b of 0.55 (median 0.63).

System

You are an expert video retrieval reranker. You will evaluate a `video` against an `original_query` based STRICTLY on VISUAL EVIDENCE (no metadata or titles provided).

WORKFLOW
1. Assign a TIER (T0-T8) to the video based on the rubric below.
2. Derive the `graded_relevance` (3 to 0) directly from the assigned Tier.
3. Rate the video strictly according to the Tier hierarchy (T0 is the best match, T8 is the worst).

TIER & GRADE RUBRIC

GRADE 3 (Exact Match / Positive)
• T0: Unambiguous, exact visual match. The requested subject, action, and visual context are clearly visible (includes both raw/primary footage and official news/highlight clips depicting the exact event).
GRADE 2 (Hard Negative / High Visual Deception)
• T1: SAME subject(s) AND SAME action type, but a DIFFERENT specific occurrence (e.g., the target athlete scoring/dunking, but in a different match or year).
• T2: SAME specific action type and setting, but DIFFERENT subject(s) (e.g., a visually identical sports dunk or car crash, but involving different individuals or vehicles).
• T3: SAME visual setting and event, but the core target action is missing or obstructed (e.g., showing the correct match, but filming only the crowd or bench instead of the target play).
GRADE 1 (Soft Negative / Weakly Related)
• T4: SAME subject(s), but a COMPLETELY DIFFERENT action or setting (e.g., the target athlete in a press conference or restaurant rather than playing).
• T5: Content ABOUT the query without showing the actual event footage (e.g., studio commentary, tactical whiteboards, or vloggers talking to the camera).
• T6: Same broad topic or category, but visually distinct (e.g., query requests WW2 historical footage, but video shows modern military exercises).
GRADE 0 (Irrelevant / Easy Negative)
• T7: Surface-level visual overlap only, but fundamentally unrelated to the query intent (e.g., accidental background color or shape similarity).
• T8: Categorically unrelated content (e.g., off-topic videos, music videos, or random memes).

OUTPUT FORMAT
Output a single, valid JSON object evaluating this video. Do not include markdown code fences (```json).

```
{
  "tier": "T0",
  "graded_relevance": 3,
  "justification": "Brief visual evidence from the video supporting the assigned
tier."
}
```

User

```
original_query: {query}
video: <video>
```
Grade the video strictly on its actual visual content, in the JSON format specified above.

Figure 3: Annotator prompt for coarse-grained tiering.

B.3 Training Data for MMDocIR and TREC DL

We instantiate the proposed two-stage annotation pipeline for TREC DL and MMDocIR using benchmark-specific rubrics: the official four-level scale (perfectly / highly / related / irrelevant) for TREC DL, and page-level answer relevance for MMDocIR. To prevent data contamination, all training queries are kept strictly disjoint from evaluation splits:

- **TREC DL:** Queries are gathered from the MS MARCO v1 passage training split alongside the ReasonIR-HQ and Promptriever pools (Cai et al., 2026). Candidate passages retrieved via BM25 top-k are labeled under the four-point rubric, where ground-truth (qrels) positives are assigned grade 3 and unverified grade-3 predictions are demoted to grade 2. This yields 517,939 rank lists (898,082 pairs), disjoint from the DL19/20 topics.
- **MMDocIR:** Using the official training split and the first-stage retriever, we construct 64,994 rank lists (580,020 pairs) disjoint from the test queries. Golden pages from the official binary annotations are pegged at grade 3, and annotator grade-3 predictions on non-golden pages are demoted to grade 2, exactly as for TREC DL.

In both benchmarks the annotator thus emits multi-level graded labels, and the training supervision is four-level on *both* benchmarks; MMDocIR's binary official labels are used directly only at evaluation time (Appendix C.4).

System

You are an expert video retrieval reranker specializing in listwise reranking.

CONTEXT
In the coarse retrieval stage, a candidate pool of videos was retrieved and assigned the same relevance grade. Your goal is to break the tie among these equally-ranked videos by evaluating their ACTUAL VISUAL CONTENT (represented as sampled frames) against the user's `original_query`.

INPUT STRUCTURE
1. `original_query`: The user's search query.
2. A list of candidate videos labeled sequentially as "Video A", "Video B", "Video C", etc. Each video is provided as a sequence of uniformly sampled visual frames.

EVALUATION CRITERIA
1. **Visual Relevance First**: Judge the true relevance based strictly on visual evidence in the frames—such as key subjects, actions, specific events, visual text (OCR/subtitles), and background context.
2. **Strict Total Order**: Provide a strict, non-tied ranking from best (1st place) to worst.
3. **Completeness Constraint**: The `ranked` array MUST contain every candidate video label (e.g., A, B, C...) exactly once. No missing or duplicate labels are allowed.

OUTPUT FORMAT
Output ONLY a single, valid JSON object without markdown code fences (```json), commentary, or extra text. Use the following exact JSON structure:

```
{
  "ranked": [
    {
      "label": "A",
      "reason": "Single concise sentence explaining visual alignment or discrepancy with the query."
    }
  ],
  "summary": "1-2 sentences highlighting the key visual differences that led to this ranking."
}
```

User

```
original_query: {query}
Video A: <video>
Video B: <video>
...
```
Rank the videos on their actual visual content, in the JSON format specified above.

Figure 4: Annotator prompt for fine-grained intra-tier ranking.

B.4 Annotation Prompts

In this section, we present the prompt templates employed in our two-stage data generation pipeline (Sec. 3.1). Figure 3 specifies the coarse-grained tiering prompt, where an annotator VLM assigns

candidates to structured relevance tiers (T0–T8) mapped onto a four-point scale. To resolve intra-tier ties, Figure 4 illustrates the listwise prompt that determines a strict total ordering based on comparative visual evidence.

C Experimental and Evaluation Details

This section details the model scoring mechanism and the shared training configuration for the main experiments.

C.1 Scoring Parameterization and Training Recipes

Following the single-token formulation in Qwen3-VL-Reranker, our model adopts a *pointwise scoring* scheme. Specifically, the relevance score for a query-candidate pair (q, d) is computed from the logit difference between the `yes` and `no` tokens at the final generation step, squashed into $[0, 1]$ via a sigmoid function:

$$s(q, d) = \sigma\big(\text{logit}(\texttt{yes}) - \text{logit}(\texttt{no})\big). \tag{12}$$

The exact scoring prompt template is illustrated in Figure 5. Under this formulation, listwise supervision is imposed solely through the training loss without altering the underlying generative architecture.

System

Judge whether the Document meets the requirements based on the Query and the Instruct provided. Note that the answer can only be "yes" or "no".

User

```
<Instruct>: Given a search query, retrieve relevant candidates that answer the query.
<Query>: {query}
<Document>: {video}
```

Figure 5: Pointwise scoring template of the reranker, shared by training and evaluation.

For model training, all compared variants follow a unified recipe unless specified otherwise. We optimize the network using AdamW under DeepSpeed ZeRO-2, applying LoRA adaptation to the language projection layers while keeping the vision encoder frozen. Detailed training hyperparameters and configurations are summarized in Table 6.

Table 6: Training configurations and hyperparameters for all variants.

Configuration / Parameter	**Value / Specification**
Model Architecture & Scoring	
Backbone	Qwen3-VL-8B-Instruct
LoRA Settings	$r = 16$, $\alpha = 32$, dropout 0.05, bias: none
LoRA Target Modules	Language tower (`q, k, v, o, gate, up, down_proj`)
Frozen Components	Vision tower, all non-LoRA parameters
Scoring Formulation	$z = \text{logit}(\texttt{yes}) - \text{logit}(\texttt{no})$, $s = \sigma(z)$
Loss Objective	
Total Objective	$\mathcal{L} = \lambda_{\text{ListMLE}}\mathcal{L}_{\text{ListMLE}} + \lambda_{\text{Anchor}}\mathcal{L}_{\text{Anchor}}$
Hyperparameters	$\lambda_{\text{ListMLE}} = 1.0$, $\lambda_{\text{Anchor}} = 1.0$
Anchor Bands (Grades 0–3)	$[0, 0.4]$, $[0.4, 0.6]$, $[0.6, 0.8]$, $[0.8, 1.0]$
Training & Optimization	
Optimizer	AdamW (weight decay $= 0$, grad clip $= 1.0$)
Learning Rate	10^{-5} (cosine decay, warmup ratio 0.03)
Batch Layout	8 docs/query, 1 query/device, 4 grad accum steps
Total Effective Batch	256 rows ($8 \times 4 \times 8$ GPUs; 32 query groups)
Training Epochs	2
Precision & Framework	`bf16`, DeepSpeed ZeRO-2, gradient checkpointing
Video Input	32 frames per video, sampled at 2 fps

C.2 Baseline scores and metric applicability

This section describes score extraction and metric reporting for the baselines evaluated in Sec. 4.4. For ZipRerank (Sun et al., 2026), MM-R5 (Xu et al., 2025), and UniME (Gu et al., 2025), we use their window-relative ranking outputs. These outputs encode relative position within a candidate list rather than a standalone relevance score. We therefore report ranking metrics for these models and omit them from the tr. BA comparison in Table 5.

For the pointwise baselines, Qwen3-VL-Reranker-8B (Li et al., 2026) and Qwen3-Reranker-8B (Zhang et al., 2025b) use $s = \sigma(\ell_{\text{yes}} - \ell_{\text{no}})$, where ℓ denotes a token logit; Rank1-7B (Weller et al., 2025) similarly uses $s = \sigma(\ell_{\text{true}} - \ell_{\text{false}})$, and LamRA in pointwise mode uses the binary softmax $s = \sigma(\ell_{\text{Yes}} - \ell_{\text{No}})$ at the first generated token. For these probability-scale outputs, we report tr. BA and AUC. ERank-14B (Cai et al., 2026) instead returns a confidence-weighted predicted grade, $s_i = \text{grade}_i \cdot \Pr[\text{token} = \text{grade}_i]$. We evaluate this score with threshold-based metrics.

C.3 Threshold-transfer evaluation protocol

We select a separate threshold τ^* for each model by maximizing BA over all pooled dev query–candidate pairs, then apply it to the pooled test pairs using $\hat{y} = \mathbb{1}[s > \tau^*]$. Thus, evaluation is at the pooled-pair level rather than macro-averaged across queries. Threshold search exhaustively evaluates the $m + 1$ partitions induced by the m distinct dev scores, including the all-positive and all-negative predictions. Ties are resolved in favor of the largest searched threshold.

Dev and test queries are disjoint. Each eligible query contains candidates from both evaluated classes; we retain the intersection across models. We use 30% dev and 70% test splits over 20 random seeds (0–19) and report the mean. Where a paired comparison is discussed in the text, we also report the win count on identical splits.

Figure 2(b) varies the calibration budget over $n_{\text{dev}} \in \{1, 2, 5, 10, 20, 50, 100, 200, 400\}$ queries, using the rest for testing. At $n_{\text{dev}} = 0$, we use $\tau = 0.5$ without calibration. We also report two test-optimized references: the pooled oracle optimizes one threshold over pooled evaluation pairs, while the per-query oracle optimizes a threshold per query and macro-averages BA (Table 16). Neither is used for threshold-transfer evaluation.

C.4 Benchmark-Specific Evaluation Settings

MultiVENT 2.0 (Main Benchmark). MultiVENT 2.0 (Kriz et al., 2025) is a large-scale, multilingual benchmark designed for event-centric video retrieval across diverse real-world news and user-generated footage. Its test split comprises 2,546 queries evaluated against a 109,800-video corpus. We evaluate on the OmniEmbed top-100 candidate pool (254,600 pairs; first-stage Recall@100 = 74.9), containing 9,634 four-level human-annotated pairs (grades 3/2/1/0: 5,313 / 3,235 / 112 / 974) and 244,966 unjudged candidates. We report nDCG@K, Recall@K, and tr. BA across the lower (G_2 vs. G_0) and upper (G_3 vs. G_2) boundaries against both judged and unjudged negatives to reflect deployed search distributions.

TREC DL 19/20. The TREC Deep Learning (DL) tracks (Craswell et al., 2020; 2021) serve as standard benchmarks for text passage reranking, built upon the MS MARCO dataset with fine-grained human relevance judgments. Following the ERank protocol, candidate pools consist of the top-100 passages per topic retrieved via BM25 from Pyserini's prebuilt `msmarco-v1-passage` index. Evaluations use official four-level graded qrels across judged topics (43 for DL19, 54 for DL20). We report nDCG@10 (annual and averaged) and tr. BA across the decision boundary (G_2 vs. G_0), where the lower boundary is assessed against both judged negatives and the pooled judged/unjudged set.

MMDocIR. MMDocIR (Dong et al., 2025) benchmarks page-level retrieval from visually rich PDFs across 10 domains. Following ZipRerank, we rerank the top-20 pages retrieved within each query's source document by $\text{DSE}_{\textit{wiki-ss}}$ or ColQwen2. Although training uses four-level annotator grades (Appendix B.3), evaluation follows the official binary labels: pages in the ground-truth set are positive, and all remaining candidate pages are treated as implicit negatives, with no separate unjudged category. We report official Recall@K for $K \in \{1, 3, 5\}$ and tr. BA on this single relevant-

Table 7: Benchmark evaluation profiles and protocols. Scoring rule, two-stage training pipeline, and threshold-transfer protocol are identical across benchmarks. G_k denotes relevance grade k.

Dimension	MultiVENT 2.0	TREC DL 19/20	MMDocIR
Benchmark Profile			
Modality	Video frames	Text passages	Page images
Domain / Task	Video retrieval	Passage rerank	Doc/page retrieval
Queries / Topics	2,546	43 / 54	1,658
Relevance labels	4-level (G_0–G_3)	4-level (G_0–G_3)	Binary (rel./irrel.)
First-stage retriever	OmniEmbed	BM25	DSE / ColQwen2
Candidate pool	Top-100 / query	Top-100 / topic	Top-20 (in-doc)
Evaluation Protocol			
Ranking metrics	nDCG@K, Recall@K	nDCG@10	Recall@$\{1,3,5\}$
BA boundaries	G_2 vs G_0 ; G_3 vs G_2	G_2 vs G_0 ; G_3 vs G_2	Rel. vs Irrel.
BA negative pool	Judged / Unjudged	Judged / +Unjudged	In-doc non-target

versus-irrelevant boundary. Because positive labels require evidence supporting the answer, this boundary differs from the Grade-2-versus-Grade-0 comparison on MultiVENT; their BA values therefore measure different operational distinctions.

All three benchmarks share an identical scoring rule and threshold-transfer protocol, while spanning diverse modalities, candidate retrieval strategies, and label granularities (Table 7). Rather than enforcing artificial metric uniformity across mismatched label spaces, we evaluate each benchmark using its community-standard ranking metric—nDCG and Recall via `pytrec_eval` (exponential gain $2^{\text{rel}}-1$) for MultiVENT and TREC DL, and the official page-level evaluator for MMDocIR—supplemented by the shared tr. BA task.

D Additional Experiments and Diagnostics

D.1 Hyperparameter Sensitivity

Throughout the main text, we fix $\lambda_{\text{Anchor}} = 1$ and $\beta = 1$ without tuning. Here, we evaluate their sensitivity (Table 8) across varying configurations while keeping the rest of the setup identical to Sec. 4.1 (including $\lambda_{\text{ListMLE}} = 1$, band boundaries $(0.4, 0.6, 0.8)$, and the evaluation protocol over 20 splits).

(a) Anchor weight λ_{Anchor}.

	Lower (tr. BA)		Upper	Ranking
λ_{Anchor}	Judged	Unjudg.	G3/G2	nDCG@5
0.25	**70.6**	85.3	66.2	65.0
0.5	70.5	85.4	66.6	**65.3**
0.75	70.3	85.3	66.7	**65.3**
1.0	70.4	**85.6**	66.8	**65.3**
2.0	70.1	85.2	**67.3**	65.1
4.0	68.3	82.6	66.7	63.0

(b) Barrier steepness β.

	Lower (tr. BA)		Upper	Ranking
β	Judged	Unjudg.	G3/G2	nDCG@5
0.5	**70.9**	85.1	66.5	**65.5**
1.0	70.4	**85.6**	66.8	65.3
2.0	70.2	**85.6**	**67.1**	65.1

Table 8: Hyperparameter sensitivity. Default settings are shaded. Best values within each subtable are bold, including ties at the reported precision.

Sensitivity to anchor weight λ_{Anchor}. Table 8a shows limited sensitivity across the tested settings in the eightfold range $\lambda_{\text{Anchor}} \in [0.25, 2.0]$. Judged tr. BA, unjudged tr. BA, and nDCG@5 vary by at most 0.5, 0.4, and 0.3 points, respectively, while upper G3/G2 performance varies by 1.1 points. Every tested setting in this range outperforms the pointwise BCE baseline on both negative-class comparisons in all 20 splits (Δjudged $\in [+1.9, +2.3]$; Δunjudged $\in [+5.7, +6.1]$).

At $\lambda_{\text{Anchor}} = 4.0$, performance declines: nDCG@5 decreases from 65.3 under the default configuration to 63.0, and judged tr. BA is close to the BCE baseline (Δjudged $= +0.1$ points, with wins in $12/20$ splits). This degradation is consistent with excessive emphasis on band anchoring at the expense of within-band ranking. Although $\lambda_{\text{Anchor}} = 0.25$ yields slightly higher judged tr. BA than the default (70.6 vs. 70.4), the default achieves the highest unjudged tr. BA and ties for the highest nDCG@5 at the reported precision. These results support the use of the fixed default $\lambda_{\text{Anchor}} = 1$ without task-specific tuning.

Sensitivity to barrier steepness β. The anchor barrier $\text{softplus}(\beta x)/\beta$ has an asymptotic out-of-band slope of one, while β controls the sharpness of the transition. Across $\beta \in \{0.5, 1, 2\}$, judged tr. BA, unjudged tr. BA, upper G3/G2 performance, and nDCG@5 vary by at most 0.7, 0.5, 0.6, and 0.4 points, respectively (Table 8b).

No tested setting consistently dominates across all four metrics: $\beta = 0.5$ gives the highest judged tr. BA and nDCG@5, whereas $\beta = 2$ gives the highest upper G3/G2 performance. The default $\beta = 1$ ties for the highest unjudged tr. BA at the reported precision and remains competitive on the other metrics. These results indicate limited sensitivity within the tested range and support retaining the fixed default $\beta = 1$.

D.2 Score Distributions and Tier Geometry

Tier separation. We characterize score separation at the two evaluated boundaries using Cohen's d on judged MultiVENT 2.0 test candidates at reranking depth 100: $d_{2\text{v}0} = \frac{\mu_2 - \mu_0}{\sqrt{(\sigma_2^2 + \sigma_0^2)/2}}$, $d_{3\text{v}2} = \frac{\mu_3 - \mu_2}{\sqrt{(\sigma_3^2 + \sigma_2^2)/2}}$, where μ_g and σ_g are the mean and standard deviation of sigmoid scores for grade g.

Table 9: Tier separation for the ablation arms in Table 3, using judged test candidates. Ratio denotes $d_{2\text{v}0}/d_{3\text{v}2}$.

Method	$d_{2\text{v}0}$	$d_{3\text{v}2}$	Ratio
BCE	0.66	0.88	0.75
ListMLE only	0.74	0.82	0.91
Band anchor only	1.01	0.82	1.24
Point anchor only	1.00	0.77	1.29
ListMLE + point	0.96	0.77	1.25
BandRank	1.08	0.76	1.42

Compared with BCE, BandRank exhibits greater lower-tier separation ($d_{2\text{v}0}$: $0.66 \rightarrow 1.08$) but smaller upper-tier separation ($d_{3\text{v}2}$: $0.88 \rightarrow 0.76$), increasing their ratio from 0.75 to 1.42 (Table 9). This pattern is consistent with the target-band layout: the Grade-1 interval $[0.4, 0.6]$ separates the bands for Grades 0 and 2, whereas the Grade-2 and Grade-3 bands meet at 0.8.

However, this geometry alone does not guarantee threshold transfer. Point anchor only has a higher separation ratio than band anchor only (1.29 vs. 1.24), yet lower unjudged tr. BA (83.1% vs. 84.0%). We therefore treat these statistics as descriptive diagnostics, not as sufficient conditions for successful threshold transfer (Sec. 4.3.3).

Score-density analysis. Figure 6 complements the separation statistics with score-density plots. For BCE, 36% of judged scores lie within 0.13 ± 0.05, and the Grade-0 and Grade-2 medians are only 0.08 apart (0.12 vs. 0.20). For BandRank, the medians are 0.27 apart (0.38 vs. 0.65), straddling the Grade-1 interval. Within-grade variation remains substantial: the Grade-0 standard deviation is 0.22, compared with BCE's 0.15. These distributional differences are consistent with Figure 2(a) (Sec. 4.3.3): BA remains near its maximum across the Grade-1 band. On the fixed sigmoid-score scale, this indicates reduced sensitivity to threshold choice on the evaluated candidates.

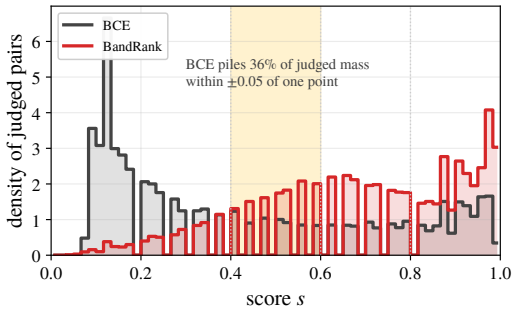


Figure 6: Score densities on MultiVENT 2.0 test candidates (depth 100). Shading: Grade-1 band $[0.4, 0.6]$; dotted lines: boundaries $(0.4, 0.6, 0.8)$.

D.3 Extended Threshold-Transfer Results

We provide additional results supporting the boundary-specific comparison in Sec. 4.3.2 and the operating-point analyses in Sec. 4.3.3, following the protocol in Appendix C.3.

Grade 3 versus the unjudged pool. Table 10 evaluates human Grade-3 candidates against unjudged candidates treated as negatives on MultiVENT 2.0, using the same query-splitting protocol and 20 seeds as Table 2. Here, pooled-oracle BA serves as a test-optimized, non-deployable upper bound reference. BandRank achieves the highest tr. BA (90.1%), exceeding Pointwise BCE (88.3%) and both external rerankers (86.6%–87.8%). Thus, BandRank's shortfall relative to Pointwise BCE on Grade 3 versus Grade 2 does not extend to this coarser comparison. Moreover, both the ListMLE-only and band-anchor-only variants also outperform the baselines, demonstrating that the advantage on this boundary does not arise exclusively from combining the two objectives.

Table 10: Performance on Grade 3 versus unjudged candidates on MultiVENT 2.0. Best results per column are in bold.

Model	tr. BA (%)	Pooled-oracle BA (%)
RankVideo-8B	86.6	86.8
Qwen3-VL-Reranker-8B	87.8	87.9
Pointwise BCE	88.3	88.5
ListMLE only	89.2	89.4
Band anchor only	89.7	90.0
BandRank	**90.1**	**90.4**

Limited-data calibration. Table 11 complements Figure 2(b) by evaluating human Grade-2 versus judged Grade-0 candidates under a fixed budget of ten dev queries against the standard 30% dev split. With ten queries, BandRank achieves 69.2% tr. BA, exceeding Pointwise BCE both at the same budget (66.5%) and with the full dev split (68.3%). BandRank also shows the smallest observed decrease across the two settings among the evaluated variants; the band-anchor-only variant does not share this pattern. These results support the limited-data performance of the joint objective, rather than a calibration-efficiency benefit attributable to anchoring alone. Under the zero-calibration setting, the predefined threshold $\tau = 0.5$ yields 70.8% BA. As specified in Appendix C.3, the queries remaining after dev selection are used for testing, so the test sets vary with the calibration budget.

Table 11: tr. BA (%) on Grade 2 vs. Grade 0 with limited dev queries (20 seeds). Best in bold.

	Lower (tr. BA)	
Model	30% dev	$n_{\text{dev}} = 10$
Pointwise BCE	68.3	66.5
ListMLE only	69.9	68.4
Band anchor only	69.3	67.0
BandRank	**70.4**	**69.2**

D.4 Scope of the Threshold-Transfer Claim

We test how far the frozen-threshold result of Sec. 4.1 extends. All analyses reuse the existing depth-100 predictions without retraining.

Where the threshold lands. Table 12 lists the threshold τ^* selected on dev queries. BandRank's thresholds fall where its band layout predicts: 0.516 on the lower boundary, next to the Grade-1 midpoint 0.5, and 0.785 on the upper boundary, next to the Grade-2/Grade-3 boundary 0.8. Models without an anchor select 0.09–0.16 on the lower boundary, a value that can only be found with labeled data; combined with ListMLE, the point anchor drifts to 0.429. This is why the untuned default $\tau = 0.5$ already reaches 70.8% BA for BandRank (Sec. 4.3.3). The anchor fixes *where* the threshold lies, not how much it varies. Across splits, all fine-tuned models already have small variation. Measured relative to the Grade-2/Grade-0 score gap, BandRank's variation is slightly larger than that of BCE and ListMLE only. RankVideo-8B is unstable (τ^* ranges over $[0.00, 0.85]$) because its scores saturate.

Table 12: Threshold τ^* selected on dev queries (mean $\pm$ std over 20 random splits) on MultiVENT 2.0, on the $[0, 1]$ score scale. The last row gives the location implied by BandRank's band layout $(0.4, 0.6, 0.8)$.

Model	G2 vs. judged G0	G2 vs. unjudged	G3 vs. G2
Qwen3-VL-Reranker-8B	0.125 ± 0.018	0.069 ± 0.005	0.256 ± 0.113
RankVideo-8B	0.157 ± 0.289	0.000 ± 0.000	0.999 ± 0.001
Pointwise BCE	0.137 ± 0.010	0.127 ± 0.007	0.351 ± 0.039
ListMLE only	0.086 ± 0.011	0.048 ± 0.004	0.375 ± 0.056
Band anchor only	0.531 ± 0.031	0.426 ± 0.018	0.741 ± 0.031
ListMLE + point	0.429 ± 0.043	0.309 ± 0.023	0.750 ± 0.059
BandRank	0.516 ± 0.025	0.379 ± 0.017	0.785 ± 0.033
Designed location	*0.5*	–	*0.8*

Shift in language and event type. Random splits mix all languages and topics in both dev and test. Table 13 instead tunes the threshold on the queries of a *single* group and applies it unchanged to all other groups, using two partitions provided by MultiVENT 2.0: video language (6 groups) and event type (8 groups). Among the fine-tuned models, BandRank has the best average in all four settings and the best worst case in two. Qwen3-VL-Reranker-8B remains stronger on judged negatives, as in Table 2. BandRank's averages under language shift (70.7 judged, 84.8 unjudged) stay within 0.8 points of its random-split values, so its threshold does not depend on dev and test queries sharing a language or topic.

Table 13: tr. BA (%) when the threshold is tuned on one group of queries only and applied unchanged to all other groups (MultiVENT 2.0). Avg./Worst: average and lowest result over the choice of tuning group. Best per column in **bold**.

	Tuned on one language (6 groups)				**Tuned on one event type** (8 groups)			
	Judged neg.		Unjudged neg.		Judged neg.		Unjudged neg.	
Model	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
Qwen3-VL-Reranker-8B	**75.3**	**71.1**	81.5	79.4	**75.3**	**71.1**	82.4	81.7
RankVideo-8B	63.4	61.2	77.0	75.9	63.1	60.3	75.9	72.8
Pointwise BCE	68.2	66.6	79.3	77.7	67.7	65.0	78.8	76.0
ListMLE only	69.4	67.5	84.3	**83.8**	68.9	65.3	84.2	82.6
Band anchor only	68.8	67.5	83.5	81.7	68.8	66.0	83.5	81.4
ListMLE + point	69.2	68.4	83.1	82.1	68.9	66.9	83.0	81.7
BandRank	70.7	69.8	**84.8**	83.0	69.7	66.7	**85.2**	**83.6**

Candidate-pool depth. Reranking fewer candidates changes the ratio of positives to negatives and removes the easiest negatives. Table 14 repeats the protocol using only the top 20, 50, or 100 candidates per query. At every depth, BandRank is the best fine-tuned model on judged negatives and the best model overall on unjudged negatives. Its tr. BA decreases as the pool shrinks ($70.4 \rightarrow 69.5$ judged, $85.6 \rightarrow 80.6$ unjudged), while its judged threshold stays inside the Grade-1 band (0.502–0.525). BCE and ListMLE only keep their thresholds near 0.13 and 0.05–0.10.

D.5 Per-Query Significance of Ranking Gains

Table 15 tests whether the ranking gains in Tables 1 and 3 hold at the query level rather than being driven by a few queries. For each of the 2,546 test queries, we compute BandRank's nDCG@5 minus that of each model. BandRank's advantage is significant against every model, and all confidence intervals exclude zero. The smallest margin is against ListMLE only (+0.7): adding the band anchor yields a real but modest ranking gain, and its main benefit lies in threshold transfer. These intervals reflect query sampling only, not variation across fine-tuning seeds or pseudo-label generation.

Table 14: tr. BA (%) on MultiVENT 2.0 when only the top-d candidates of each query are reranked and evaluated. Best per column in **bold**.

	G2 vs. judged G0			G2 vs. unjudged		
Model	d=20	d=50	d=100	d=20	d=50	d=100
Qwen3-VL-Reranker-8B	**74.1**	**75.1**	**76.1**	78.8	81.4	82.9
RankVideo-8B	63.3	63.3	63.3	72.4	74.9	77.0
Pointwise BCE	67.8	68.2	68.3	76.0	77.9	79.5
ListMLE only	69.0	69.5	69.9	79.6	82.4	84.4
Band anchor only	68.8	69.0	69.3	79.3	81.8	84.0
ListMLE + point	68.7	68.8	68.9	78.5	81.4	83.3
BandRank	69.5	69.9	70.4	**80.6**	**83.6**	**85.6**

Table 15: Per-query nDCG@5 comparison on MultiVENT 2.0 (2,546 queries). Δ: mean gain of BandRank over the row model (points). CI: 95% bootstrap confidence interval over queries. Win/Loss: queries where BandRank scores higher/lower; the rest are ties. All differences are significant under a Wilcoxon signed-rank test ($p < 0.01$).

Model	Δ	**95% CI**	**Win / Loss**
Qwen3-VL-Reranker-8B	$+20.3$	$[+19.2, +21.5]$	1456 / 271
RankVideo-8B	$+7.6$	$[+6.7, +8.4]$	1018 / 380
Pointwise BCE	$+5.4$	$[+4.7, +6.0]$	843 / 312
ListMLE + point	$+2.1$	$[+1.6, +2.5]$	601 / 350
Band anchor only	$+1.3$	$[+1.0, +1.6]$	506 / 262
ListMLE only	$+0.8$	$[+0.4, +1.1]$	427 / 361

E Supplementary Analysis of Query-Wise Score Normalization

We examine whether training-free, query-wise score transformations can substitute for training-time anchoring. A strictly monotone global map (e.g., Platt scaling) only relocates the dev-selected threshold and leaves tr. BA unchanged, whereas query-conditioned shifts preserve within-query rankings but alter cross-query comparisons. Using label-free statistics of each query's retrieved pool, we evaluate mean centering ($s - \mu_{\text{pool}}$), standardization ($(s - \mu_{\text{pool}})/\sigma_{\text{pool}}$), pool-max shifting ($s - s_{\max}^{(\text{pool})}$), and a first-stage shift ($s - 0.5\, s_{\max}^{(1)}$, where $s_{\max}^{(1)}$ is the query's top first-stage retrieval score). Thresholds are selected on dev and frozen on test as in Sec. 4.1. The per-query threshold oracle is a test-optimized, macro-averaged diagnostic, not an attainable bound on pooled tr. BA.

Table 16: tr. BA (%) on the MultiVENT 2.0 lower boundary (Grade 2 vs. judged Grade 0) under query-wise score normalization. Bold indicates the best non-oracle result for each model. The per-query threshold oracle is an empirical diagnostic reference.

Transformation	**BCE**	**ListMLE-only**	**BandRank**	**Qwen3-VL-Reranker-8B**
Raw scores (s)	68.3	69.9	70.4	76.1
Mean-centered ($s - \mu_{\text{pool}}$)	68.9	**71.4**	**73.5**	75.6
Standardized ($\frac{s-\mu_{\text{pool}}}{\sigma_{\text{pool}}}$)	68.3	69.9	72.2	73.4
First-stage shifted ($s - 0.5 s_{\max}^{(1)}$)	**69.2**	71.2	71.4	**76.5**
Pool-max shifted ($s - s_{\max}^{(\text{pool})}$)	54.8	55.9	67.3	71.1
Per-query threshold oracle	*87.5*	*88.3*	*89.0*	*90.4*

On MultiVENT 2.0, mean centering is the most effective transformation for the fine-tuned models. It raises BandRank from 70.4 to 73.5 and ListMLE only from 69.9 to 71.4, widening the gap between them from 0.5 to 2.1 points; normalization thus complements rather than replaces interval anchoring. The benefit is model-dependent: Qwen3-VL-Reranker-8B changes by less than one point under mean centering or the first-stage shift, and pool-max shifting degrades every model.

Query-wise normalization nonetheless has two limitations. First, applying a global τ after mean centering equals an adaptive threshold $\tau + \mu_{\text{pool}}$, making decisions dependent on the retriever and pool depth. Second, its effect varies across benchmarks: on each primary boundary, mean centering alters BandRank by $+4.7$ ($82.7 \rightarrow 87.4$) on MMDocIR (DSE) but -12.4 ($83.8 \rightarrow 71.4$) on TREC DL, while pool-max shifting yields $+6.4$ and -1.6, respectively.[2] Because the per-query positive rate spans 0–100% on TREC DL (vs. 0–40% on MMDocIR), μ_{pool} partly reflects candidate relevance; subtracting it discards signal alongside the offset. Thus, query-wise normalization remains an optional per-benchmark adjustment, whereas our anchored scale needs no pool statistics at inference.

[2]TREC DL uses Grade 2 vs. judged Grade 0 here; Table 5 additionally treats unjudged passages as negatives.