MODALITY-ASSIGNMENT SENSITIVITY IN UNIVERSAL MULTIMODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

In mixed-modality retrieval, fixing candidate identities and modality counts still leaves open which entities are represented as images or text. We study sensitivity to this *modality assignment* across 16 representative universal multimodal embedding models across three image–text datasets. For each pool, the query, instruction, positive, and negative identities are fixed. First, we increase the number of hard negatives represented in the modality opposite to the positive while decreasing that number among controls, preserving the total image and text counts. Second, we keep each negative group half image and half text and enumerate all assignments of which entities appear in each modality. Retrieval shows assignment-sensitivity under both settings. We conduct analyses and find that rank sensitivity is concentrated among a few competitors. A mean-shift approximation built on the original scores in the positive modality reproduces a substantial part of both coverage-path endpoint changes and fixed-group-count sensitivity despite candidate-level score errors. An audit further reveals attainable reversals in model comparisons that could be missed by a specific assignment. These findings identify modality assignment as an evaluation variable, alongside candidate identities and modality counts, in mixed-modality retrieval. Code is available at this link.

1 INTRODUCTION

Universal multimodal embedding models (MEMs) aim to map text, images, and other content types into a shared retrieval space, enabling search across heterogeneous collections with a common similarity measure (Zhou et al., 2024; Chen et al., 2025; Meng et al., 2025; Li et al., 2026a). This shared space supports instruction-guided retrieval across modalities and tasks (Jiang et al., 2024a; Zhang et al., 2024a; Gu et al., 2025). In practice, the same entity often has multiple representations (Girdhar et al., 2023; Huang et al., 2026). For example, an object may appear as an image or as a textual description (Plummer et al., 2015; Radford et al., 2021). In mixed candidate pools, modality-dependent retrieval behavior makes the choice of representation relevant to evaluation (Lin et al., 2025; Li et al., 2025). We call the mapping from candidate entity to representation modality the *modality assignment*.

Prior work seeks to improve cross-modal comparability (Zhang et al., 2024b; An et al., 2025; Li et al., 2025). We ask a complementary evaluation question: *how strongly do retrieval outcomes and model comparisons depend on modality assignment, given fixed candidate identities and modality counts?* Consider two candidate pools that contain the same entities and the same numbers of image and text candidates, but differ in which entities are shown in each modality. Paired images and descriptions may expose different information, and instructions may impose modality requirements. We measure the sensitivity of evaluation conclusions to such *modality reassignment*.

We study this question by evaluating 16 representative universal MEMs on retrieval pools constructed from three paired image–text datasets. For each model and candidate pool, the query, instruction, positive, and negative entities remain fixed across assignments. First, we vary how many hard negatives use a modality other than the positive (the non-target modality), while moving normal negatives, which serve as a control group, in the opposite direction to preserve overall modality counts. Retrieval performance shifts sharply for many models. We then additionally fix the modality counts within both the hard-negative and control groups and enumerate all possible assignments

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.

within these groups. Retrieval remains highly dependent on which entities are presented in each modality, even under these stronger controls.

We characterize the effects through score shifts and candidate influence. For each model, dataset, and retrieval direction, the average cross-modal score shift is predictable despite substantial variation at the candidate level. Furthermore, rank sensitivity is concentrated in a small subset of hard negatives. A diagnostic approximation that replaces each negative's non-target score with its original target score plus a mean shift reproduces a substantial part of the effects in both settings.

These findings identify modality assignment as an important evaluation variable beyond candidate identities and modality counts. Reassignment can reverse model comparisons of aggregate performance even when both models face the same assignment. Some of these comparisons do not reverse along the recorded coverage path. Thus, a reported score or model comparison under one assignment does not, by itself, establish stability under other assignments that satisfy the same constraints.

2 Related Work and Our Contribution

Universal multimodal embedding models (MEMs) extend large-scale image–text alignment (Radford et al., 2021; Jia et al., 2021) toward instruction-guided retrieval across heterogeneous inputs and tasks (Wei et al., 2024; Jiang et al., 2024a;b; Zhang et al., 2024a; Li et al., 2026b; Meng et al., 2025; Li et al., 2026a). This shared space can nevertheless retain a **modality gap**, which has been linked to initialization (Liang et al., 2022), optimization (Fahim et al., 2024), and information imbalance between modalities (Schrodi et al., 2025). Geometric analyses identify off-center ellipsoidal structures in image and text representations (Levi & Gilboa, 2024). Encoder sharing (Eslami & de Melo, 2025), geometric corrections (Zhang et al., 2024b), and embedding standardization (An et al., 2025) aim to reduce this separation and improve cross-modal comparability. However, reducing the modality gap does not uniformly improve instance-wise performance (Grassucci et al., 2026). In retrieval, training-based methods seek to address wrong-modality negatives (Lin et al., 2025), uneven negative difficulty (Chen et al., 2026), and overreliance on textual cues (Juan et al., 2026). Inference-time alternatives include score standardization with statistics estimated from pseudo-positive pairs (Yamashita et al., 2025) and routing queries to modality-specific corpora (Yeo et al., 2026).

Representation-dependent behavior. Many studies examine how model behavior changes when semantically related content is represented in different modalities. CyCLIP shows that image and text representations learned through contrastive training can yield inconsistent predictions (Goel et al., 2022). Mistretta et al. (2025) further exposes intra-modal misalignment by showing that image-to-image and text-to-text retrieval can improve when query representations are inverted into the complementary modality. Decomposition-based analyses characterize when removing modality-specific components changes rankings, relating these changes to candidate-dependent modality contributions and semantic score margins (Dhimoïla et al., 2026). MixBench links document-representation replacement to same-modality ranking bias, while GR-CLIP mitigates this bias through modality-wise embedding centering (Li et al., 2025). OmniSET uses semantically matched instances to diagnose directional bias and insufficient instruction control (Huang et al., 2026). Beyond retrieval, REST tests reasoning consistency across equivalent representations (van Sprang et al., 2026), while MATE examines entity linking between images and text (Alonso et al., 2025).

Our position. Prior work establishes that modality-dependent representations and scores can affect retrieval. We focus on how strongly retrieval outcomes and pairwise model comparisons depend on *which entities compete in each modality* under matched candidate identities and modality counts. For each model and pool, we keep the query, instruction, positive, and negative entities fixed across assignments. Then we vary the allocation of modalities between hard and normal negatives while preserving overall counts. The second experiment additionally fixes each group's counts and varies only which entities receive each modality. Our analysis also separates *predicting aggregate retrieval effects* from *reconstructing individual scores*. We use the original target scores and a cross-fitted mean shift to approximate non-target scores, treating the approximation as a diagnostic rather than a calibration method. Candidate-level analysis then examines where rank variation is concentrated and how score shifts interact with the positive-score boundary.

Our contributions are threefold. First, we quantify retrieval sensitivity to *modality assignment* under fixed candidate identities and overall modality counts, showing that substantial sensitivity

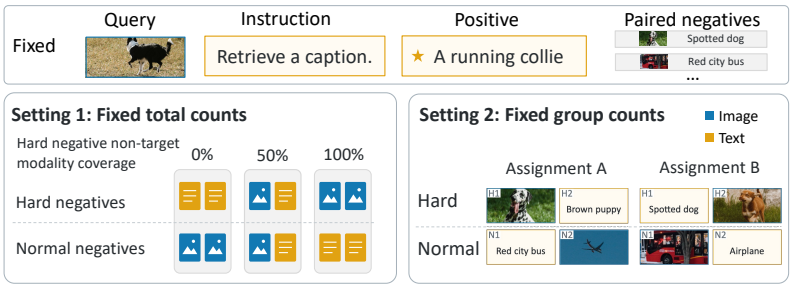


Figure 1: **Controlled modality reassignment with fixed candidate identities.** The query, instruction, positive, and paired negative entities remain fixed throughout. **Left:** we vary hard-negative non-target-modality coverage while changing controls in the opposite direction, preserving the overall image–text counts. **Right:** we additionally fix the modality counts within both negative groups and vary only which entities are represented as images or text.

persists even when counts are fixed within both negative groups. Second, we characterize the score shifts and concentration of candidate influence, and show that a mean-shift approximation reproduces a substantial part of the observations. Finally, we show that comparisons of aggregate model performance can be reversed under reassignment even when both models face the same assignment.

3 Setup

Instruction-conditioned retrieval. Given a query q, instruction I, and candidate representation c, an embedding model with parameters θ ranks candidates by cosine similarity

$$s_\theta(q, I, c) = \frac{f_\theta(q, I)^\top g_\theta(c)}{\|f_\theta(q, I)\|_2 \|g_\theta(c)\|_2}, \tag{1}$$

where f_θ and g_θ denote the model's query and candidate encoding interfaces. We evaluate image-to-text and text-to-image retrieval, with the direction defined by the query and positive modalities. The positive and all negatives are jointly ranked in a mixed-modality candidate pool.

Candidate pool construction. An entity denotes a dataset item with paired image and text representations. Each pool i contains one positive and two disjoint sets of k negative entities: hard negatives H_i and controls C_i. For each model and pool, the query, instruction, positive representation, negative identities, and evaluation labels remain fixed across assignments; only the representation selected for each negative changes.

For each retrieval direction, the positive's modality is the *target modality* M_T, and the other is the *non-target modality* M_N. For negative j in pool i, let $z_{ij} = 1$ if it is represented in M_N and $z_{ij} = 0$ if it is represented in M_T. We define the *hard-negative coverage* λ as the fraction of H_i assigned to M_N and constrain the two groups by

$$\sum_{j \in H_i} z_{ij} = k\lambda, \qquad \sum_{j \in C_i} z_{ij} = k(1-\lambda). \tag{2}$$

Modality reassignment. Along the *coverage path*, we vary $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. At each coverage level, we randomly select $k\lambda$ hard negatives to receive M_N and $k(1-\lambda)$ controls to receive M_N, with all remaining negatives represented in M_T. Thus, changing λ reallocates non-target representations between hard negatives and controls while preserving the overall modality counts. We repeat this random assignment three times, synchronized with the three independently sampled control sets described below, and average metrics over the resulting replicates.

We then hold $\lambda = 0.5$, additionally fixing the modality counts within each negative group. Any $k/2$ entities from H_i and any $k/2$ from C_i can be assigned to M_N, with the remainder assigned to M_T. The resulting assignment set $\mathcal{Z}_i$ has size $|\mathcal{Z}_i| = \binom{k}{k/2}^2$. We enumerate all legal assignments within each control replicate to measure sensitivity to *which* entities receive each modality under

unchanged group-wise counts. Fig. 1 illustrates an example. In this work, we use $k = 8$, yielding 17 candidates in a pool and 4900 legal assignments per pool at $\lambda = 0.5$.

Models. We evaluate 16 representative universal MEMs: E5-V-8B (Jiang et al., 2024a), VLM2Vec-Full, VLM2Vec-Qwen2VL-2B/7B (Jiang et al., 2024b), MM-Embed (Lin et al., 2025), BGE-VL-MLLM-S2 (Zhou et al., 2025), GME-Qwen2-VL-2B/7B-Instruct (Zhang et al., 2024a), Nomic Embed Multimodal 7B (Nomic-AI, 2025), Jina Embeddings v4 (Günther et al., 2025), Ops-MM-embedding-v1-2B/7B (OpenSearch-AI, 2025), VLM2Vec-V2 (Meng et al., 2025), RzenEmbed-v2-7B (Jian et al., 2025), and Qwen3-VL-Embedding-2B/8B (Li et al., 2026a). Each model uses its official or supported instructions, fixed across assignments without prompt tuning.

Datasets. We construct controlled retrieval pools from three paired image–text datasets: Flickr30K (Young et al., 2014), OmniSET (Huang et al., 2026), and noCaps (Agrawal et al., 2019). Each entity has an image and one or more textual descriptions. As shown in Tab. 1, we mine hard negatives for Flickr30K and noCaps using caption-based TF–IDF similarity. We aggregate each entity's available descriptions, compute TF–IDF representations, and rank all non-self entities by cosine similarity. Candidates are reviewed in descending similarity order by OpenAI GPT-5.6 Sol. Rejected candidates are skipped and replaced by the next candidate in the ranking until k accepted hard negatives are obtained. OmniSET directly uses k released hard negatives.

Table 1: Paired image–text sources and hard-negative construction. Query counts are per retrieval direction.

Dataset	Queries	Hard negatives
OmniSET	101	Released negatives
Flickr30K	1000	Caption TF–IDF
noCaps	4500	Caption TF–IDF

For each query, after excluding the positive and hard negatives, we independently sample three sets of k control entities. Each control set defines one evaluation replicate. For the coverage path, modality assignments at each λ are randomly sampled within each replicate according to Eq. 2. The control sets and their corresponding assignments are fixed before model evaluation and shared across models. Metrics at each coverage level are averaged over the replicates. At $\lambda = 0.5$, fixed-group-count sensitivity is computed by enumerating all legal assignments separately for each replicate.

Metrics. For a fixed model and specified assignment, let r denote the rank of the positive candidate. We report Hit@1 and mean reciprocal rank (MRR):

$$\text{Hit@1} = \mathbb{E}[\mathbf{1}(r = 1)], \qquad \text{MRR} = \mathbb{E}[1/r]. \tag{3}$$

Higher values indicate better retrieval performance.

4 Results and Analyses

We first examine retrieval under changing hard-negative coverage and reassignment at fixed composition (Sec. 4.1). We then analyze the average modality shifts and influential candidates behind these effects (Sec. 4.2), and test whether the same mean-shift approximation can predict both (Sec. 4.3). Finally, we examine the consequences for relative model performance (Sec. 4.4).

4.1 Retrieval Depends on Modality Assignment

Changing hard-negative coverage. Along the reassignment path, increasing λ moves more hard negatives into the non-target modality while moving controls in the opposite direction. Candidate entities and the total numbers of image and text candidates remain unchanged.

Finding 1. Retrieval performance generally declines as more hard negatives are presented in the non-target modality, despite unchanged candidate entities and overall modality counts.

Fig. 2 shows a consistent pattern across the three sources. From $\lambda = 0$ to $\lambda = 1$, both Hit@1 and MRR decrease for the same 14 of the 16 models across all three datasets. The two exceptions, MM-Embed and Nomic Embed Multimodal 7B, improve on both metrics. Averaged over all models, Hit@1 drops by 15.4, 12.8, and 19.8 percentage points on Flickr30K, OmniSET, and noCaps, respectively, while MRR drops by 0.122, 0.107, and 0.174. The magnitude of the response nevertheless varies substantially across models.

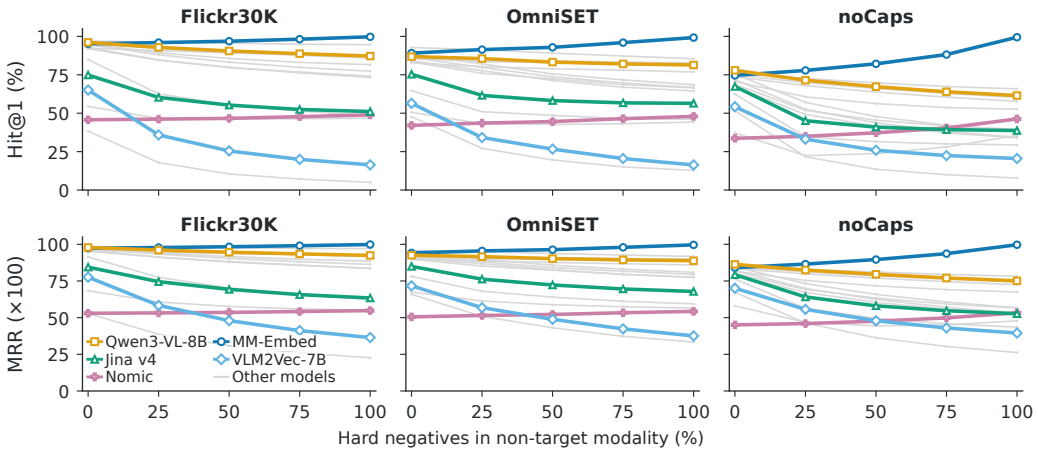


Figure 2: **Modality allocation changes retrieval despite fixed overall modality counts.** Within each replicate, the positive and candidate identities remain fixed while non-target coverage increases among hard negatives and decreases among controls. Curves average three randomized replicates.

The path changes both hard negatives and controls, so we distinguish their contributions as follows. Starting from $\lambda = 0$, we switch hard negatives alone, controls alone, or both. The single-group interventions change the overall modality counts, while the joint intervention preserves them. Tab. 2 reports the mean changes in Hit@1 across the 16 models. Switching hard negatives alone produces declines close to the joint change across all three sources, whereas switching controls alone improves performance. Although these effects are not additive, the contrast identifies hard-negative reassignment as the main contributor to the average decline.

Table 2: Mean Hit@1 changes from $\lambda = 0$ in percentage points.

	Hard	Ctrl	Joint
Flickr30K	−15.6	+10.3	−15.4
OmniSET	−13.3	+9.2	−12.8
noCaps	−19.6	+3.2	−19.8

Reassignment at fixed composition. The five-level path preserves overall modality counts but changes their allocation between hard negatives and controls. We next fix the counts within each group. At $\lambda = 0.5$, each group contains the same number of image and text candidates. We enumerate all legal assignments while retaining the same query, positive, and candidate entities.

Finding 2. Fixing modality composition does not fix retrieval behavior. Even with unchanged modality counts in both negative groups, retrieval depends on which entities receive each modality.

Averaged across all models, Hit@1 is assignment-dependent for 17.3%, 23.1%, and 27.3% of candidate pools on Flickr30K, OmniSET, and noCaps, respectively. In these pools, the same positive ranks first under some legal assignments but not others. The positive rank varies across assignments in an even larger fraction of pools, at 32.2%, 38.9%, and 57.0%, respectively, showing that reassignment can also change the positive's position without changing top-1 success. These changes occur while keeping the query, positive candidate entities, and modality counts fixed within both negative groups. Together, the two experiments show that retrieval depends on both the number of hard negatives in each modality and which entities receive which representations.

4.2 Modality Bias and Influential Hard Negatives

The two reassignment experiments show that retrieval can change even when candidate entities remain fixed. To understand these changes, we first separate the average score shift associated with modality from candidate-specific deviations. We then examine whether rank variation is spread across the pool or concentrated in a few competitors.

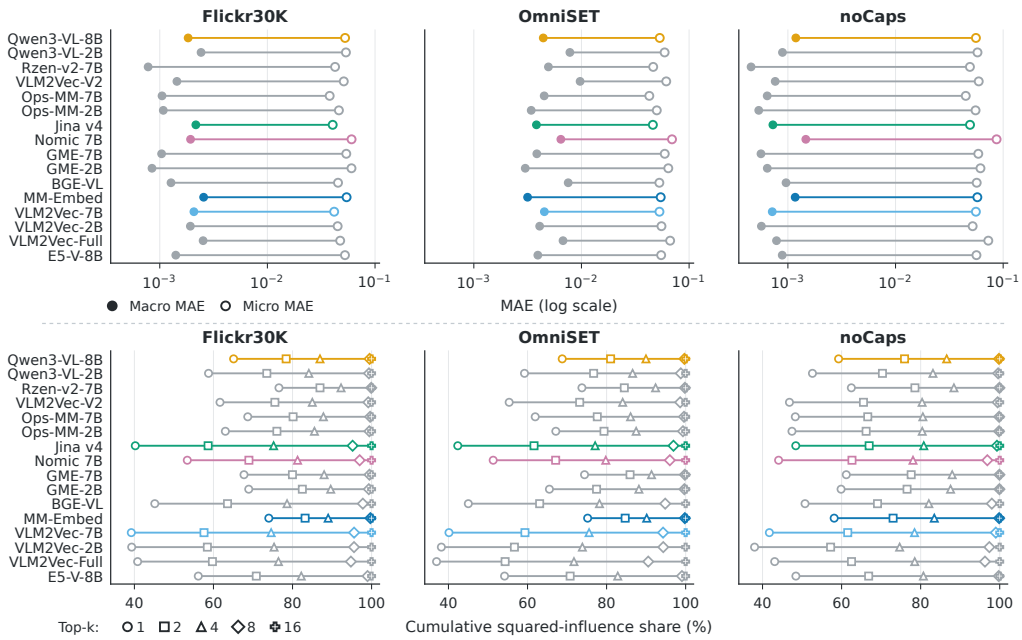


Figure 3: **Top: mean shifts are reproducible across folds.** Cross-fitted mean-shift MAE against held-out fold means (macro) and individual shifts (micro), averaged over retrieval directions. **Bottom: candidate influence is concentrated.** Cumulative squared-influence shares of the top-k negatives, averaged over pools whose positive rank varies under reassignment with $\lambda = 0.5$. Influence is the conditional difference in expected positive rank between a negative's two modalities.

Finding 3. Modality bias is stable at the population level despite substantial candidate-level error. A few competing candidates account for most assignment-induced rank variation.

Population-level modality bias. For negative candidate j in pool i, let $s_{ij}^{M_T}$ and $s_{ij}^{M_N}$ denote its target- and non-target-modality scores under a fixed model. Its cross-modal score shift is $\delta_{ij} = s_{ij}^{M_N} - s_{ij}^{M_T}$, with positive values indicating a preference for the non-target representation. We estimate the mean shift μ for each model, dataset, and retrieval direction using five-fold cross-fitting. Each estimate uses only the other four query folds, averaging negatives within each query and then over queries. All pools and both retrieval directions of a query remain in the same fold.

We assess the same estimate using mean absolute error (MAE) at two levels. *Macro MAE* measures the error against the average shift of the held-out fold, whereas *micro MAE* measures the error relative to individual candidate shifts. Both errors are computed separately for each retrieval direction and then averaged equally across directions. Fig. 3 compares them for all models. Averaged across models, macro MAE is 1.65×10^{-3}, 5.12×10^{-3}, and 0.81×10^{-3} on Flickr30K, OmniSET, and noCaps, respectively, whereas micro MAE is 49.12×10^{-3}, 55.59×10^{-3}, and 58.13×10^{-3}. The average shift is thus reproducible across query folds, even though it leaves substantial candidate-specific variation less predictable.

Concentrated candidate influence. Substantial candidate-level error does not mean that all candidates matter equally for retrieval. We therefore measure each negative's influence at fixed composition by comparing the expected positive rank under its non-target and target representations,

$$\mathcal{I}_{ij} = \mathbb{E}[r_i(z) \mid z_{ij} = 1] - \mathbb{E}[r_i(z) \mid z_{ij} = 0], \tag{4}$$

where both expectations average uniformly over legal assignments that preserve each group's modality counts. With this setting, rank variance is proportional to the sum of squared influences and can be measured through normalized shares. We order candidates by absolute influence within each pool and compute their cumulative shares.

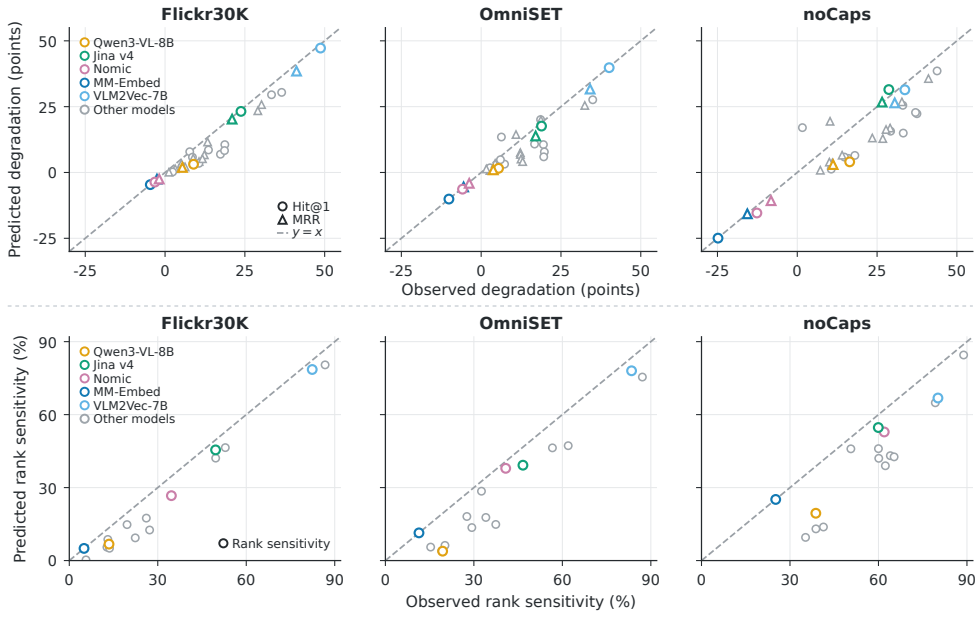


Figure 4: **A mean-shift approximation partially reproduces both retrieval effects.** Predictions use $\widehat{s}^{M_N} = s^{M_T} + \mu$ for negatives, where M_T is the positive's modality, M_N the other, and μ a cross-fitted mean shift, while other scores remain unchanged. Top: per-model endpoint drops from $\lambda = 0$ to 1 at fixed overall modality counts. Negative values indicate improvement. Bottom: percentages of pools whose positive rank varies across assignments with $\lambda = 0.5$.

Fig. 3 shows that rank variation is concentrated, although the degree of concentration differs across models. Across three datasets, the most influential candidate alone accounts for $50.7-57.4\%$ of rank variance on average. The top four account for 83.3%, 83.5%, and 82.0% on Flickr30K, OmniSET, and noCaps, respectively. Hard negatives, rather than controls, account for most of this variation, contributing 86.1%, 86.9%, and 93.7% of the total variance shares.

The analyses identify a stable average modality shift and a concentration of rank sensitivity among a small subset of competitors. We next test whether adding this mean shift to the original target-modality scores can reproduce the aggregate retrieval changes in both reassignment experiments.

4.3 A Mean Shift Predicts Both Retrieval Effects

We now test whether the population-level shift can reproduce the retrieval changes observed in Section 4.1. For each held-out query, we replace every negative's non-target score with

$$\widehat{s}_{ij}^{M_N} = s_{ij}^{M_T} + \mu, \tag{5}$$

where μ is estimated from the other four query folds. Target scores and the positive score remain unchanged. The same estimate is applied to hard negatives and controls in both reassignment experiments, without further fitting. This preserves the original target-side competition while replacing candidate-specific cross-modal shifts with a common offset.

Finding 4. The same mean-shift approximation captures a substantial part of both coverage-dependent performance changes and fixed-composition sensitivity, despite substantial errors in individual score shifts.

From score differences to relative ordering. Accurate scores are not required for accurate positive ranks. A negative contributes to the positive rank based on whether it scores above the positive, not on how much it exceeds it. For a hard negative initially below the positive, a positive shift only needs to exceed their score gap to reverse their ordering. The approximation succeeds whenever the predicted and actual non-target scores remain on the same side of the positive. Across all datasets,

predictions of hard negatives that outrank the positive achieve precision of $81.4 - 87.4\%$ and recall of $61.7 - 76.6\%$. Thus, the mean shift captures many of these ordering changes without precisely recovering their score magnitudes.

Predicting both retrieval effects. Fig. 4 compares observed and predicted outcomes. Along the reassignment path, the approximation predicts mean MRR drops of 0.091, 0.080, and 0.116 points on Flickr30K, OmniSET, and noCaps, compared with observed drops of 0.122, 0.107, and 0.174. Hit@1 follows the same aggregate pattern. At fixed composition, the predicted positive rank-changing rates are 25.3%, 28.0%, and 41.5%, compared with 32.2%, 38.9%, and 57.0% for the original scores. Predicted Hit-sensitive rates are likewise lower, at 13.0%, 14.8%, and 17.1%, versus 17.3%, 23.1%, and 27.3%. The approximation therefore reproduces a substantial part of both effects but systematically underestimates their aggregate magnitude.

Model differences. MM-Embed illustrates why score error and retrieval error need not coincide. Its mean shift is strongly negative in both retrieval directions. On Flickr30K and OmniSET, both the original and approximated non-target scores remain below zero for every negative candidate. Consequently, the approximation exactly reproduces positive ranks and Hit@1 outcomes across all fixed-composition assignments despite non-zero candidate-level score error. Moving hard negatives into the non-target modality can therefore improve retrieval by lowering their scores, rather than because the model has no modality bias.

The shared approximation connects the two observed phenomena without requiring accurate reconstruction of candidate-level scores. Its remaining errors also show that the population-level bias is still part of the explanation. Our analysis goes beyond identifying this bias to predicting its aggregate retrieval consequences, showing how a stable mean shift combined with the candidate competition can account for retrieval performance at macro level.

4.4 Consequences for Relative Model Performance

The preceding analyses show that models respond differently to modality reassignment. Such differences do not necessarily change their relative performance, since two models can gain or lose together. We now examine whether reassignment reverses relative aggregate performance comparisons across all model pairs.

Finding 5. Modality assignment can reverse relative model performance. Many comparisons that never reverse along the coverage path can still reverse when candidate assignments change at fixed composition.

Comparisons under shared assignments. For each dataset and metric, we orient the performance gap toward the model leading under the original $\lambda = 0.5$ assignment. We then compare the smallest aggregate gap for reassignment path and the smallest attainable gap at fixed $\lambda = 0.5$. For the fixed-composition setting, we minimize the within-pool difference over all possible assignments. A negative gap means the original leader has been overtaken.

Reversals under both experiments. Fig. 5 shows that fixed-composition reassignment exposes substantially more reversible comparisons than the recorded coverage path. For Hit@1, 15.8%, 15.0%, and 23.3% pairs reverse along the path on Flickr30K, OmniSET, and noCaps, respectively, compared with 43.3%, 55.0%, and 60.8% pairs under fixed-composition reassignment. MRR follows the same pattern, with 12.5%, 11.7%, and 17.5% path reversals compared with 40.8%, 50.0%, and 57.5% fixed-composition reversals. These are changes in aggregate model performance, not merely differences in individual-query outcomes.

Agreement along the path is insufficient. Among pairs that never strictly reverse along the five recorded levels, 33%, 47%, and 50% still admit a Hit@1 reversal at fixed composition. The corresponding percentages for MRR are 32%, 43%, and 49%. Thus, across all sources and both metrics, many fixed-composition reversals are absent from the five-level path. Testing several hard-negative coverage levels can therefore miss substantial dependence on which entities receive each modality.

These results do not imply that every model comparison is unstable, nor do attainable reversals establish how frequently they occur under a particular assignment policy. They show that modality assignment can affect not only a model's retrieval score but also its measured advantage over another model, even when both face the same entities and identical modality counts.

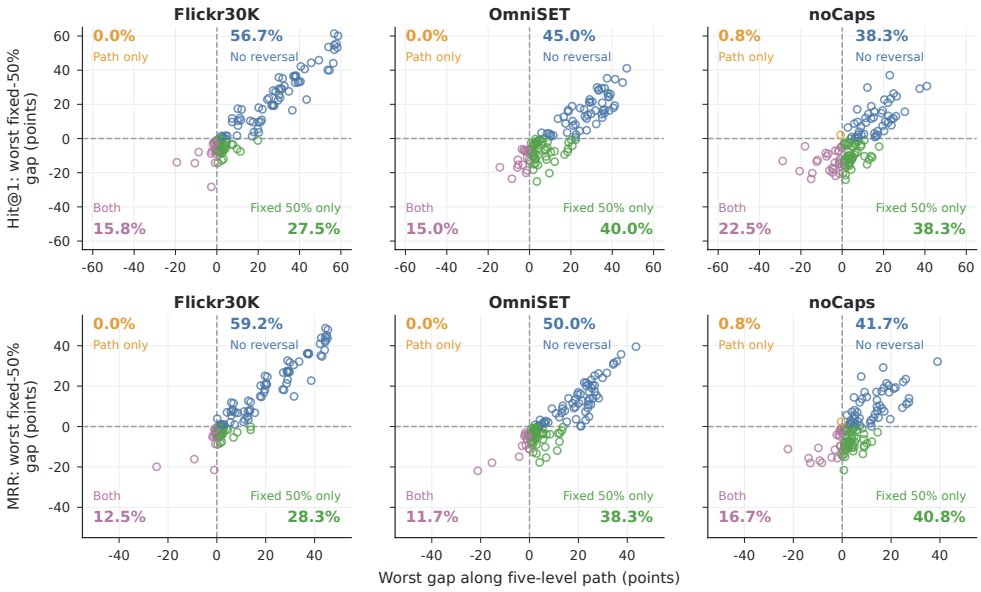


Figure 5: **Fixed-count reassignment reveals reversals missed by the sampled coverage.** Each point is one of model pairs, with gaps oriented toward the leader at the originally sampled baseline. Horizontal: minimum aggregate gap over five coverage levels. Vertical: average per-pool minimum over assignments shared by both models, with $\lambda = 0.5$. Negative values indicate reversals. Percentages are calculated over all 120 model pairs.

5 Discussion

Discussion 1: Retrieval quality and assignment stability are distinct. A high retrieval score under one assignment does not establish stability under other assignments of the same entities, even with fixed group-wise modality counts. Sensitivity is not inherently undesirable: paired representations may expose different information, and instructions may impose modality requirements. The evaluation issue is how much reported scores and model advantages depend on these choices. Mixed-modality evaluation should therefore specify both modality counts and assignment rules, and distinguish performance under the reported rule from robustness to alternative assignments.

Discussion 2: Aggregate retrieval effects can be predictable despite errors in individual scores. The mean-shift approximation retains candidate-specific target scores and the positive score. Its predictive value comes from combining this score configuration with a cross-fitted offset, not from the offset alone. Because positive rank depends on which negatives outrank the positive rather than on exact score differences, imperfect score reconstruction can still yield accurate retrieval predictions. This provides a partial score-level account of both coverage-path endpoint changes and fixed-group-count sensitivity. Its underestimation of model-averaged effects shows that ranking-relevant variation remains not fully captured. The approximation is a diagnostic of how shifts in observed scores affect retrieval at the macro level.

6 Conclusion

This study shows that fixing candidate identities and modality counts does not ensure assignment-stable evaluation in universal multimodal retrieval. Sensitivity persists under fixed group-wise counts, and a cross-fitted mean-shift approximation built on retained target scores provides a partial score-level account of the retrieval effects. Shared within-pool reassignment can also reverse aggregate model comparisons that were missed by evaluations based on specific samples. Mixed-modality evaluation should therefore report assignment rules alongside modality counts and distinguish performance under a specified rule from robustness to alternative assignments.

LLM USAGE STATEMENT

In this work, we used generative AI tools for dataset review and quality checking, partial code implementation based on human-provided examples, and analysis and interpretation of experimental results. We have not used generative AI tools for the other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing parts of the research paper to improve readability. We have reviewed all AI-assisted work. Specifically, AI-assisted dataset reviews were manually checked by the authors; AI-generated or AI-assisted code was reviewed, edited, and validated by the authors before being used in experiments; AI-assisted analyses were checked against the underlying experimental results; and AI-assisted manuscript edits were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956. IEEE, 2019.

Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez De Lacalle. Vision-language models struggle to align entities across modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18846–18862, 2025.

Na Min An, Eunki Kim, James Thorne, and Hyunjung Shim. I0t: Embedding standardization method towards zero modality gap. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27182–27199, 2025.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8254–8275, 2025.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19430–19443, 2026.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Picard. Cross-modal redundancy and the geometry of vision-language embeddings. *arXiv preprint arXiv:2602.06218*, 2026.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In *International Conference on Learning Representations*, volume 2025, pp. 82070–82088, 2025.

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. *arXiv preprint arXiv:2601.18525*, 2026.

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.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 531–550, 2025.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, et al. Mmeb-v3: Measuring the performance gaps of omni-modality embedding models. *arXiv preprint arXiv:2604.23321*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024a.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024b.

Li Juan, Chuanghao Ding, Xujie Zhang, and Cam-Tu Nguyen. Mimic: Mitigating visual modality collapse in universal multimodal retrieval while avoiding semantic misalignment. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3208–3220, 2026.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. *arXiv preprint arXiv:2411.14517*, 2024.

Binxu Li, Yuhui Zhang, Xiaohan Wang, Weixin Liang, Ludwig Schmidt, and Serena Yeung-Levy. Closing the modality gap for mixed modality search. *arXiv preprint arXiv:2507.19054*, 2025.

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.

Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. U-marvel: Unveiling key factors for universal multimodal retrieval via embedding learning with mllms. In *International Conference on Learning Representations*, volume 2026, pp. 121298–121319, 2026b.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 44215–44234, 2025.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *International Conference on Learning Representations*, volume 2025, pp. 77750–77771, 2025.

Nomic-AI. Nomic embed multimodal 7b, 2025. URL `https://huggingface.co/nomic-ai/nomic-embed-multimodal-7b`.

OpenSearch-AI. Ops-mm-embedding-v1, 2025. URL https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-2B.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 27836–27864, 2025.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. Same content, different answers: Cross-modal inconsistency in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8781–8790, 2026.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Shuhei Yamashita, Daiki Shirafuji, and Tatsuhiko Saito. Bridging the modality gap by similarity standardization with pseudo-positive samples. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pp. 126–139, 2025.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024a.

Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *International Conference on Learning Representations*, volume 2024, pp. 51711–51736, 2024b.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3185–3200, 2024.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025.

APPENDIX

A EXPERIMENTAL DETAILS

A.1 CANDIDATE POOL CONSTRUCTION

Flickr30K. We use the 1000 images in the Flickr30K Entities test set. The first accepted released caption serves as both the text query and text candidate. Mining instead aggregates entity phrases from all five captions, excluding nonvisual phrases and rejected captions. We review non-self entities in descending TF–IDF cosine similarity and retain the first eight accepted negatives, continuing down the ranking after each rejection. Codex/GPT-5 checks query and candidate images together with their captions. Pairs must be clear negatives in both retrieval directions; duplicates, image–caption mismatches, and ambiguous positives are rejected.

OmniSET. We use all 101 queries from the released MSCOCO cross-modal retrieval subset and the first eight released hard negatives per query. Text queries use the released query text; text candidates use the first catalog caption for each entity.

noCaps. We use all 4500 validation images. The caption with the smallest annotation ID is encoded. All ten captions are concatenated only for mining. We review non-self entities in descending TF–IDF cosine similarity and retain the first eight accepted negatives, continuing down the ranking after each rejection.

Mining and controls. TF–IDF uses lowercase words, English stop-word removal, unigrams and bigrams, and smoothed inverse document frequencies. Similarity ties are resolved by ascending image ID. Controls are sampled without replacement after excluding the positive and hard negatives. Flickr30K also excludes rejected hard-negative relations; OmniSET excludes the entire released hard-negative list. Each query has three control replicates, each with eight controls and the same eight hard negatives. For Flickr30K and noCaps, the control sets partition 24 sampled entities into three disjoint groups. All three datasets undergo visual review; Flickr30K controls use the same screening as its hard negatives. Candidate identities and labels are fixed before model evaluation and shared across models.

A.2 MODEL INPUTS AND ENCODING

We use each model's supported encoding interface without prompt tuning, retaining native wrappers and image placeholders. Where supported, the direction-specific instruction D is *"Find the image that best matches the given text:"* for text-to-image retrieval and *"Find the text that best matches the given image:"* for image-to-text retrieval. Tab. 3 specifies the query and candidate inputs.

All models produce one ℓ_2-normalized vector per input. MM-Embed uses its latent-attention retriever, Jina Embeddings v4 uses token-mean pooling, and the other models use their supported last-token pooling. Nomic Embed Multimodal 7B uses its dense rather than late-interaction interface. Model-specific image processing is retained: BGE-VL-MLLM-S2 inputs are first resized to 512×512, VLM2Vec-Full uses 16 crops, and MM-Embed uses the global image patch. We cache embeddings and paired candidate scores; reassignment requires neither training nor re-encoding.

A.3 ASSIGNMENTS AND AGGREGATION

Coverage path. Each control replicate uses one fixed random permutation per negative group. Starting at $\lambda = 0$, the first $k\lambda$ hard negatives switch from M_T to M_N, while the first $k\lambda$ controls switch from M_N to M_T. The five coverage levels are nested; both retrieval directions use the same path in relative target/non-target roles.

Fixed-group-count reassignment. At $\lambda = 0.5$, we enumerate all $\binom{8}{4}^2 = 4{,}900$ assignments with four non-target negatives per group within each replicate.

Ranking and aggregation. Candidates are ranked by decreasing cosine similarity, with exact ties resolved by ascending integer candidate ID. Scores are not rounded before comparison. For a dataset

Table 3: Model-specific query and candidate inputs. D denotes the direction-specific retrieval instruction.

Models	Query and candidate inputs
Qwen3-VL-Embedding-2B/8B	**Query:** Use D as the system prompt. **Candidate:** Use *"Represent the user's input."* as the system prompt.
VLM2Vec-Full, VLM2Vec-Qwen2VL-2B/7B, VLM2Vec-V2	**Query:** Include D in the input content. **Candidate:** Use captions without added instructions and native templates for images.
BGE-VL-MLLM-S2	**Query:** Use D as the query prompt. **Candidate:** Add no prompt.
GME-Qwen2-VL-2B/7B-Instruct, Ops-MM-embedding-v1-2B/7B	**Query:** Use D as the system prompt. **Candidate:** Use *"You are a helpful assistant."* as the system prompt.
RzenEmbed-v2-7B	**Query:** Include D in the input content. **Candidate:** Use captions without a prefix; include *"Represent the given image."* for images.
MM-Embed	**Query:** Use `Instruct:` D, followed by `Query:` and the input. **Candidate:** Add no instruction.
E5-V-8B	**Query:** Use the native *"Summary above sentence in one word:"* or *"Summary above image in one word:"* template by modality, without D. **Candidate:** Use the same modality-specific templates, without D.
Jina Embeddings v4	**Query:** Prefix text with `Query:`; use *"Describe the image."* for images. Add no D. **Candidate:** Prefix text with `Passage:`; use *"Describe the image."* for images. Add no D.
Nomic Embed Multimodal 7B	**Query:** Append ten `<|endoftext|>` tokens to text; use *"Describe the image."* for images. Add no D. **Candidate:** Use captions without a prefix and *"Describe the image."* for images. Add no D.

with Q queries and $R = 3$ control replicates, the reported score is

$$S_m = \frac{1}{2QR} \sum_{q=1}^{Q} \sum_{d \in \{T \to I, I \to T\}} \sum_{\rho=1}^{R} u_{m,q,d,\rho}, \qquad u \in \{\mathbf{1}[r = 1],\, 1/r\}. \tag{6}$$

Here u is the per-pool utility and r the positive rank. Model averages weight all 16 models equally. The Hit@1-sensitive indicator is $\mathbf{1}[\min_z r(z) = 1 < \max_z r(z)]$. The rank-sensitive indicator is $\mathbf{1}[\min_z r(z) < \max_z r(z)]$. We compute these indicators within each fixed pool before applying the same aggregation.

B Retrieval Sensitivity and Score-Level Results

B.1 Coverage Paths and Fixed-Count Sensitivity

Tabs. 4–6 report all five coverage levels and both fixed-count sensitivity measures for every model. Hit@1 is in percent and MRR is multiplied by 100. Endpoint drops subtract the value at $\lambda = 1$ from that at $\lambda = 0$. Negative drops indicate improvement. All three tables use the same model order and average both retrieval directions.

B.2 Retrieval Directions and Negative Groups

Direction averages can hide contrasting responses. Tab. 7 separates endpoint drops by direction. Tab. 8 reports hard-only, control-only, and joint changes from the same $\lambda = 0$ baseline. The first

Table 4: Flickr30K: complete coverage paths and fixed-count sensitivity. Coverage columns give 100λ. D_H, D_R are endpoint drops; F_H, F_r are percentages of Hit@1-sensitive and rank-sensitive pools at $\lambda = 0.5$.

	Hit@1						MRR $\times 100$						Sensitivity	
Model	0	25	50	75	100	D_H	0	25	50	75	100	D_R	F_H	F_r
Qwen3-VL-Embedding-8B	96.1	92.8	90.5	88.8	87.2	8.9	97.9	96.1	94.6	93.5	92.4	5.4	11.7	13.5
Qwen3-VL-Embedding-2B	95.2	89.3	85.6	83.1	81.6	13.6	97.3	94.0	91.6	89.8	88.5	8.8	16.7	19.6
RzenEmbed-v2-7B	97.0	96.3	95.6	94.9	94.6	2.4	98.3	97.9	97.6	97.2	97.1	1.2	4.8	5.7
VLM2Vec-V2	92.0	84.5	80.0	76.2	73.3	18.6	95.3	91.2	88.3	85.8	83.6	11.7	22.1	27.2
Ops-MM-embedding-v1-7B	96.6	92.9	89.8	87.7	86.0	10.6	98.1	96.1	94.3	92.9	91.8	6.3	11.5	13.6
Ops-MM-embedding-v1-2B	94.7	88.0	83.2	79.9	77.3	17.4	96.9	93.3	90.4	88.1	86.3	10.6	19.1	22.4
Jina Embeddings v4	74.9	60.3	55.3	52.5	51.1	23.8	84.4	74.5	69.3	65.7	63.4	21.0	28.6	49.6
Nomic Embed Multimodal 7B	45.7	46.2	46.7	47.8	48.9	−3.2	53.0	53.2	53.6	54.2	54.8	−1.8	4.2	34.6
GME-Qwen2-VL-7B-Instruct	95.1	92.2	89.8	88.2	86.5	8.6	97.2	95.5	94.1	93.0	92.0	5.2	10.5	13.1
GME-Qwen2-VL-2B-Instruct	92.4	90.8	89.2	88.0	87.0	5.3	95.6	94.7	93.7	93.0	92.4	3.2	9.6	12.8
BGE-VL-MLLM-S2	85.0	62.5	55.5	51.4	48.4	36.5	91.4	77.5	70.3	65.7	62.4	29.1	30.0	49.6
MM-Embed	95.1	96.0	96.9	98.2	99.7	−4.6	97.3	97.8	98.3	99.1	99.9	−2.5	5.0	5.0
VLM2Vec-Qwen2VL-7B	65.1	35.9	25.5	20.0	16.4	48.7	77.6	58.3	48.0	41.2	36.4	41.2	43.5	82.4
VLM2Vec-Qwen2VL-2B	38.4	17.8	10.5	7.1	5.0	33.4	52.8	38.7	30.8	26.0	22.6	30.2	28.9	86.7
VLM2Vec-Full	54.5	46.6	46.2	46.4	46.8	7.7	68.4	60.7	57.6	56.0	55.0	13.3	9.7	52.9
E5-V-8B	92.8	84.6	79.6	76.8	74.1	18.7	95.9	91.2	87.9	85.7	83.6	12.3	21.0	26.1
Model mean	81.9	73.5	70.0	67.9	66.5	15.4	87.3	81.9	78.8	76.7	75.1	12.2	17.3	32.2

Table 5: OmniSET: complete coverage paths and fixed-count sensitivity. Coverage columns give 100λ. D_H, D_R are endpoint drops; F_H, F_r are percentages of Hit@1-sensitive and rank-sensitive pools at $\lambda = 0.5$.

	Hit@1						MRR $\times 100$						Sensitivity	
Model	0	25	50	75	100	D_H	0	25	50	75	100	D_R	F_H	F_r
Qwen3-VL-Embedding-8B	86.8	85.5	83.3	82.1	81.4	5.4	92.6	91.6	90.2	89.4	88.8	3.8	15.1	19.4
Qwen3-VL-Embedding-2B	82.9	80.3	78.9	78.0	76.8	6.0	90.3	88.7	87.6	86.8	85.9	4.4	22.2	27.6
RzenEmbed-v2-7B	92.8	91.3	89.1	87.2	85.4	7.3	95.8	95.0	93.7	92.7	91.7	4.0	11.9	15.3
VLM2Vec-V2	84.1	77.7	71.3	66.9	64.5	19.6	90.5	86.8	82.7	79.6	77.5	13.0	29.7	37.4
Ops-MM-embedding-v1-7B	88.3	82.1	75.5	71.7	68.5	19.8	93.3	89.7	85.8	83.1	80.9	12.5	24.2	29.3
Ops-MM-embedding-v1-2B	86.2	79.8	73.7	69.7	66.7	19.5	91.8	88.0	84.2	81.7	79.5	12.3	25.1	34.1
Jina Embeddings v4	75.3	61.6	58.2	56.7	56.4	18.9	84.9	76.2	72.3	69.6	67.9	17.1	27.6	46.6
Nomic Embed Multimodal 7B	42.1	43.6	44.6	46.4	47.9	−5.8	50.5	51.5	52.2	53.4	54.3	−3.7	7.1	40.8
GME-Qwen2-VL-7B-Instruct	85.5	84.7	84.1	83.2	82.8	2.8	91.6	91.2	90.9	90.3	90.0	1.6	13.3	19.1
GME-Qwen2-VL-2B-Instruct	85.7	83.6	83.0	81.9	81.4	4.4	91.3	90.2	89.7	89.0	88.5	2.7	15.1	20.2
BGE-VL-MLLM-S2	64.7	51.0	48.6	46.4	46.0	18.6	78.1	68.0	64.0	61.2	59.4	18.7	30.3	56.6
MM-Embed	89.1	91.4	92.9	95.9	99.2	−10.1	94.1	95.5	96.4	98.0	99.6	−5.5	11.4	11.4
VLM2Vec-Qwen2VL-7B	56.4	34.2	26.6	20.5	16.3	40.1	71.7	56.8	48.9	42.4	37.6	34.1	43.7	83.5
VLM2Vec-Qwen2VL-2B	47.7	27.1	19.6	15.1	12.8	35.0	65.9	50.9	43.1	37.3	33.4	32.5	43.2	87.1
VLM2Vec-Full	50.6	43.2	42.7	43.2	44.3	6.3	67.4	61.3	58.8	57.4	56.5	10.9	28.2	62.0
E5-V-8B	83.3	76.0	72.4	68.6	66.5	16.7	89.8	85.2	82.4	79.5	77.6	12.3	21.2	32.6
Model mean	75.1	68.3	65.3	63.3	62.3	12.8	83.7	79.2	76.4	74.4	73.1	10.7	23.1	38.9

Table 6: noCaps: complete coverage paths and fixed-count sensitivity. Coverage columns give 100λ. D_H, D_R are endpoint drops; F_H, F_r are percentages of Hit@1-sensitive and rank-sensitive pools at $\lambda = 0.5$.

	Hit@1						MRR $\times 100$						Sensitivity	
Model	0	25	50	75	100	D_H	0	25	50	75	100	D_R	F_H	F_r
Qwen3-VL-Embedding-8B	77.9	71.5	67.3	63.9	61.5	16.4	86.3	82.4	79.5	77.0	75.1	11.2	25.1	38.8
Qwen3-VL-Embedding-2B	70.6	60.7	56.3	53.7	52.6	18.0	81.5	75.3	71.7	69.1	67.5	14.0	30.2	50.6
RzenEmbed-v2-7B	76.5	72.6	69.7	67.4	65.8	10.6	85.4	83.0	81.1	79.4	78.2	7.2	22.0	35.3
VLM2Vec-V2	67.2	49.2	41.6	37.2	34.0	33.2	79.0	67.2	60.4	55.8	52.2	26.8	33.9	65.3
Ops-MM-embedding-v1-7B	75.4	57.1	47.8	42.2	38.4	37.0	84.5	73.2	65.9	60.7	56.7	27.8	35.5	60.1
Ops-MM-embedding-v1-2B	71.9	52.6	43.7	38.1	34.4	37.5	82.1	69.9	62.5	57.1	53.0	29.1	35.5	64.0
Jina Embeddings v4	67.4	45.1	41.0	39.3	38.7	28.7	79.3	64.1	58.1	54.7	52.7	26.6	24.7	59.9
Nomic Embed Multimodal 7B	33.6	35.0	37.2	40.4	46.3	−12.7	45.1	46.0	47.6	49.8	53.3	−8.2	13.5	62.0
GME-Qwen2-VL-7B-Instruct	75.1	69.9	66.1	62.9	60.3	14.8	84.4	81.2	78.6	76.3	74.3	10.1	23.3	38.8
GME-Qwen2-VL-2B-Instruct	73.7	68.0	63.8	60.5	57.7	16.0	83.4	79.9	77.0	74.6	72.4	11.0	24.7	41.3
BGE-VL-MLLM-S2	62.4	34.7	31.5	30.1	29.4	33.1	76.3	56.3	49.5	45.8	43.5	32.8	20.2	62.3
MM-Embed	74.6	77.8	82.1	88.2	99.4	−24.8	84.1	86.5	89.6	93.6	99.7	−15.6	22.4	25.2
VLM2Vec-Qwen2VL-7B	54.2	33.1	25.9	22.4	20.5	33.7	70.0	55.6	47.9	43.0	39.5	30.5	34.7	80.2
VLM2Vec-Qwen2VL-2B	51.6	21.7	13.5	10.0	7.8	43.8	67.3	46.4	36.3	30.3	26.2	41.1	30.2	88.9
VLM2Vec-Full	37.1	22.3	23.8	28.0	35.5	1.6	58.1	46.8	44.4	45.0	47.8	10.3	30.4	79.3
E5-V-8B	68.7	51.9	45.4	41.7	39.5	29.2	80.1	69.3	63.3	59.3	56.6	23.5	30.8	60.0
Model mean	64.9	51.4	47.3	45.4	45.1	19.8	76.7	67.7	63.3	60.7	59.3	17.4	27.3	57.0

Table 7: Directional endpoint drops for all models. Hit@1 differences use percentage points; MRR differences use scaled MRR points. Negative values indicate improvement.

	Flickr30K				OmniSET				noCaps			
	D_H		D_R		D_H		D_R		D_H		D_R	
Model	$T{\to}I$	$I{\to}T$	$T{\to}I$	$I{\to}T$	$T{\to}I$	$I{\to}T$	$T{\to}I$	$I{\to}T$	$T{\to}I$	$I{\to}T$	$T{\to}I$	$I{\to}T$
Qwen3-VL-Embedding-8B	17.8	0.1	10.9	0.0	7.5	3.4	4.8	2.8	23.3	9.6	16.0	6.5
Qwen3-VL-Embedding-2B	28.8	−1.6	18.5	−0.9	17.8	−5.7	12.2	−3.4	38.0	−1.9	29.3	−1.2
RzenEmbed-v2-7B	0.9	3.9	0.5	2.0	5.0	9.7	2.5	5.6	6.4	14.9	3.6	10.8
VLM2Vec-V2	7.9	29.4	4.6	18.9	8.7	30.5	5.6	20.4	23.6	42.8	16.0	37.6
Ops-MM-embedding-v1-7B	10.0	11.2	5.8	6.8	17.8	21.8	11.3	13.7	37.9	36.2	28.4	27.2
Ops-MM-embedding-v1-2B	6.6	28.1	3.8	17.5	12.5	26.5	7.2	17.4	30.2	44.8	21.4	36.8
Jina Embeddings v4	−0.3	47.8	−0.2	42.2	−8.9	46.7	−5.6	39.7	1.2	56.1	1.0	52.1
Nomic Embed Multimodal 7B	−6.6	0.3	−4.0	0.3	−11.7	0.0	−6.9	−0.6	−25.6	0.3	−17.0	0.6
GME-Qwen2-VL-7B-Instruct	12.5	4.6	7.7	2.6	2.8	2.8	1.6	1.6	17.6	11.9	11.5	8.6
GME-Qwen2-VL-2B-Instruct	7.6	3.1	4.7	1.6	5.3	3.4	2.6	2.9	13.4	18.6	8.7	13.3
BGE-VL-MLLM-S2	64.9	8.1	53.4	4.7	39.2	−2.0	38.2	−0.8	54.8	11.3	56.9	8.7
MM-Embed	−5.1	−4.1	−2.9	−2.2	−11.1	−9.1	−6.1	−4.9	−24.6	−25.1	−15.5	−15.7
VLM2Vec-Qwen2VL-7B	59.0	38.4	50.1	32.2	44.0	36.2	34.3	34.0	54.8	12.6	46.5	14.6
VLM2Vec-Qwen2VL-2B	59.5	7.2	49.3	11.1	41.4	28.5	32.5	32.4	52.6	35.0	44.9	37.2
VLM2Vec-Full	19.7	−4.3	28.9	−2.3	24.0	−11.3	28.1	−6.2	39.3	−36.1	44.7	−24.2
E5-V-8B	0.8	36.6	0.3	24.3	4.8	28.7	2.2	22.3	9.7	48.7	5.5	41.5

two interventions change overall modality counts. Only the joint intervention preserves them. Their changes are not additive contributions to the joint effect.

B.3 Score Shifts and Candidate Influence

Tab. 9 gives the two score-shift errors and the concentration statistics for every model. Small fold-mean error coexists with larger candidate-level error. The influence columns describe where rank variation is concentrated; they do not measure the fraction of pools that are sensitive.

Table 8: Single-group changes from $\lambda = 0$ for all models. Each cell gives Hit@1 / scaled MRR; negative values indicate degradation. Joint changes are $-D_H$ and $-D_R$ in Tabs. 4–6.

	Flickr30K		OmniSET		noCaps	
Model	Hard	Control	Hard	Control	Hard	Control
Qwen3-VL-Embedding-8B	−9.1 / −5.6	0.5 / 0.3	−6.3 / −4.3	1.0 / 0.6	−16.5 / −11.3	0.2 / 0.1
Qwen3-VL-Embedding-2B	−13.9 / −9.0	1.2 / 0.7	−6.7 / −5.1	3.2 / 1.9	−18.0 / −14.1	0.6 / 0.5
RzenEmbed-v2-7B	−2.5 / −1.3	0.1 / 0.0	−7.4 / −4.1	0.1 / 0.1	−10.7 / −7.2	0.0 / 0.0
VLM2Vec-V2	−19.2 / −12.2	1.7 / 1.0	−20.6 / −14.0	4.0 / 2.3	−33.3 / −27.0	1.0 / 0.7
Ops-MM-embedding-v1-7B	−10.8 / −6.4	0.4 / 0.2	−20.4 / −12.9	1.4 / 0.8	−37.1 / −27.9	0.4 / 0.2
Ops-MM-embedding-v1-2B	−17.7 / −10.9	0.9 / 0.5	−20.4 / −12.9	2.2 / 1.2	−37.6 / −29.2	0.4 / 0.3
Jina Embeddings v4	−24.1 / −22.1	14.9 / 9.7	−18.9 / −17.6	7.4 / 4.5	−28.7 / −26.7	1.9 / 1.2
Nomic Embed Multimodal 7B	3.5 / −0.3	30.3 / 30.9	6.9 / 2.0	21.4 / 25.1	12.8 / 6.1	15.0 / 18.8
GME-Qwen2-VL-7B-Instruct	−8.7 / −5.3	0.3 / 0.2	−3.2 / −1.8	0.2 / 0.2	−14.8 / −10.1	0.1 / 0.1
GME-Qwen2-VL-2B-Instruct	−5.6 / −3.3	0.3 / 0.2	−4.9 / −3.0	0.6 / 0.3	−16.0 / −11.0	0.1 / 0.1
BGE-VL-MLLM-S2	−36.9 / −29.7	11.4 / 6.6	−19.8 / −20.7	23.9 / 15.5	−33.1 / −33.2	10.8 / 7.0
MM-Embed	4.9 / 2.7	−0.1 / −0.1	10.9 / 5.9	−0.5 / −0.3	25.3 / 15.9	0.0 / 0.0
VLM2Vec-Qwen2VL-7B	−49.3 / −42.8	23.3 / 15.5	−41.4 / −36.7	25.1 / 17.0	−33.8 / −30.7	2.5 / 1.6
VLM2Vec-Qwen2VL-2B	−33.6 / −33.1	43.6 / 36.1	−36.5 / −35.4	30.5 / 21.1	−43.9 / −41.7	5.2 / 4.4
VLM2Vec-Full	−7.9 / −14.8	34.0 / 25.0	−6.4 / −12.8	24.9 / 17.4	0.4 / −9.5	12.3 / 8.2
E5-V-8B	−18.9 / −12.5	1.3 / 0.8	−17.1 / −12.6	2.0 / 1.2	−29.3 / −23.6	0.5 / 0.3
Model mean	−15.6 / −12.9	10.3 / 8.0	−13.3 / −11.6	9.2 / 6.8	−19.6 / −17.6	3.2 / 2.7

Table 9: Score-shift prediction errors and squared-influence shares. Macro and micro MAE are in units of 10^{-3}. Top-4 and Hard are percentages of normalized squared influence, averaged over each model's rank-sensitive pools.

	Flickr30K				OmniSET				noCaps			
Model	Macro	Micro	Top-4	Hard	Macro	Micro	Top-4	Hard	Macro	Micro	Top-4	Hard
Qwen3-VL-Embedding-8B	1.8	52.7	86.9	96.7	4.4	53.4	90.0	92.6	1.2	55.8	86.6	98.9
Qwen3-VL-Embedding-2B	2.4	53.6	84.1	95.9	7.8	59.2	86.6	90.6	0.9	57.7	83.1	98.1
RzenEmbed-v2-7B	0.8	42.5	92.3	96.3	4.9	46.4	92.4	95.7	0.5	49.1	88.4	99.3
VLM2Vec-V2	1.4	51.2	85.0	93.6	9.7	61.3	84.1	90.0	0.8	59.3	80.4	97.0
Ops-MM-embedding-v1-7B	1.1	37.9	87.8	96.9	4.5	42.7	86.1	96.1	0.6	44.9	80.6	99.0
Ops-MM-embedding-v1-2B	1.1	46.2	85.5	96.3	3.4	50.1	87.5	93.9	0.5	55.4	80.4	98.8
Jina Embeddings v4	2.2	40.4	75.2	80.3	3.8	46.1	77.1	88.7	0.7	49.2	80.8	95.9
Nomic Embed Multimodal 7B	1.9	60.5	81.3	59.7	6.4	69.4	79.8	69.9	1.5	86.9	78.1	67.8
GME-Qwen2-VL-7B-Instruct	1.0	54.1	88.0	96.2	3.9	59.4	91.4	94.2	0.6	58.7	88.0	99.3
GME-Qwen2-VL-2B-Instruct	0.9	60.4	89.7	94.2	3.0	64.1	88.2	96.5	0.6	61.5	87.6	99.4
BGE-VL-MLLM-S2	1.3	45.3	78.6	86.9	7.5	52.9	78.2	72.4	1.0	56.8	82.1	82.1
MM-Embed	2.6	54.4	89.0	96.5	3.2	54.5	90.1	94.5	1.2	57.6	83.5	99.1
VLM2Vec-Qwen2VL-7B	2.1	41.7	74.6	80.3	4.5	53.0	75.6	77.9	0.7	55.8	78.5	95.1
VLM2Vec-Qwen2VL-2B	1.9	44.9	75.3	64.6	4.1	55.4	73.9	71.6	0.6	52.0	74.7	89.2
VLM2Vec-Full	2.5	47.4	76.4	48.1	6.7	66.7	71.7	71.0	0.8	72.8	78.5	80.9
E5-V-8B	1.4	52.6	82.2	95.7	3.9	55.0	82.8	94.7	0.9	56.5	80.7	98.6
Model mean	1.6	49.1	83.3	86.1	5.1	55.6	83.5	86.9	0.8	58.1	82.0	93.7

Table 10: Mean-shift predictions for all models. Endpoint drops and sensitivity fractions use the definitions and scales in Tabs. 4–6. Predictions preserve the original target-score configuration.

	Flickr30K				OmniSET				noCaps			
Model	$\widehat{D}_H$	$\widehat{D}_R$	$\widehat{F}_H$	$\widehat{F}_r$	$\widehat{D}_H$	$\widehat{D}_R$	$\widehat{F}_H$	$\widehat{F}_r$	$\widehat{D}_H$	$\widehat{D}_R$	$\widehat{F}_H$	$\widehat{F}_r$
Qwen3-VL-Embedding-8B	3.1	1.9	5.3	6.8	1.6	1.0	2.2	3.9	4.0	3.0	10.0	19.5
Qwen3-VL-Embedding-2B	8.6	5.4	13.2	14.8	4.9	3.5	13.3	18.1	6.5	6.6	28.4	45.9
RzenEmbed-v2-7B	0.2	0.1	0.2	0.3	3.2	1.9	3.6	5.5	1.3	0.9	3.9	9.5
VLM2Vec-V2	8.3	5.2	9.1	12.6	5.9	4.1	9.8	14.9	15.0	12.9	18.9	42.6
Ops-MM-embedding-v1-7B	3.5	2.1	3.7	5.1	7.8	5.1	8.6	13.6	22.8	16.3	22.3	42.0
Ops-MM-embedding-v1-2B	6.9	4.1	7.4	9.3	10.6	6.9	12.1	17.7	22.3	16.9	21.2	43.1
Jina Embeddings v4	23.2	20.3	25.4	45.5	17.6	13.9	28.5	39.2	31.5	26.7	22.4	54.7
Nomic Embed Multimodal 7B	−3.7	−2.5	4.1	26.7	−6.3	−4.1	7.6	37.9	−15.4	−10.7	14.0	52.9
GME-Qwen2-VL-7B-Instruct	5.8	3.3	6.9	8.7	1.8	1.3	2.0	4.0	5.8	3.9	5.9	13.0
GME-Qwen2-VL-2B-Instruct	2.3	1.4	3.3	5.5	0.8	0.7	2.1	6.2	5.3	3.7	5.4	13.8
BGE-VL-MLLM-S2	30.4	23.4	28.3	42.1	20.1	19.9	20.4	46.3	25.5	26.8	6.2	39.0
MM-Embed	−4.6	−2.5	5.0	5.0	−10.1	−5.5	11.4	11.4	−24.9	−15.7	22.5	25.1
VLM2Vec-Qwen2VL-7B	47.3	38.5	46.1	78.6	39.8	31.7	41.0	78.0	31.4	26.4	25.6	66.8
VLM2Vec-Qwen2VL-2B	29.5	25.8	29.0	80.5	27.6	25.4	34.3	75.4	38.6	35.7	29.2	84.5
VLM2Vec-Full	7.9	11.6	7.5	46.4	13.5	14.4	19.9	47.2	17.0	19.4	14.0	64.9
E5-V-8B	10.6	6.7	13.8	17.5	10.8	7.6	19.5	28.5	15.7	13.1	23.4	46.0
Model mean	11.2	9.1	13.0	25.3	9.3	8.0	14.8	28.0	12.6	11.6	17.1	41.5

Table 11: Prediction correspondence (%). P_b, R_b: precision and recall for hard negatives that outrank the positive in M_N. P_r, R_r: precision and recall for rank-sensitive pools. The two pairs use different prediction units and denominators.

	Flickr30K				OmniSET				noCaps			
Model	P_b	R_b	P_r	R_r	P_b	R_b	P_r	R_r	P_b	R_b	P_r	R_r
Qwen3-VL-Embedding-8B	64.8	30.9	80.8	40.6	58.3	38.4	87.2	17.3	67.2	43.8	85.6	43.0
Qwen3-VL-Embedding-2B	67.6	43.2	81.4	61.4	60.0	46.2	90.7	59.5	70.3	56.0	87.5	79.4
RzenEmbed-v2-7B	49.5	31.5	64.7	3.2	38.9	29.8	55.4	20.0	65.7	43.2	85.3	23.0
VLM2Vec-V2	70.9	40.1	85.0	39.2	63.9	32.7	90.0	35.7	79.8	53.9	92.3	60.2
Ops-MM-embedding-v1-7B	74.7	35.0	78.0	29.0	72.2	41.9	89.1	41.2	77.0	51.5	90.0	62.9
Ops-MM-embedding-v1-2B	77.6	37.5	89.8	37.5	69.8	49.0	82.1	42.7	79.1	53.0	91.1	61.4
Jina Embeddings v4	80.5	74.5	94.7	86.8	80.9	61.0	92.4	77.7	82.4	72.5	85.6	78.1
Nomic Embed Multimodal 7B	99.5	99.7	96.7	74.5	99.9	99.8	98.4	91.5	99.5	98.7	99.3	84.8
GME-Qwen2-VL-7B-Instruct	59.2	43.2	64.5	42.9	52.6	54.5	77.5	16.1	67.1	48.9	82.2	27.6
GME-Qwen2-VL-2B-Instruct	62.0	44.9	69.8	30.2	53.1	44.2	82.5	25.5	67.3	47.1	83.9	28.0
BGE-VL-MLLM-S2	82.8	64.4	91.8	77.9	77.5	77.5	95.1	77.8	89.4	85.0	89.2	55.8
MM-Embed	—	—	100.0	100.0	—	—	100.0	100.0	—	0.0	100.0	99.8
VLM2Vec-Qwen2VL-7B	80.9	71.8	92.3	88.0	80.1	71.7	92.6	86.6	75.2	64.4	92.4	77.0
VLM2Vec-Qwen2VL-2B	90.7	85.8	95.1	88.2	84.3	73.4	95.7	82.8	84.8	77.1	94.5	89.9
VLM2Vec-Full	90.6	91.5	95.9	84.2	73.8	73.3	98.5	75.1	78.3	85.2	97.0	79.4
E5-V-8B	81.9	46.1	89.5	59.8	83.9	59.0	83.3	72.9	85.0	59.5	92.9	71.3

B.4 Mean-Shift Prediction

Tab. 10 reports predictions using $s^{M_T} + \mu$ for all non-target negatives. The shift comes from the other four query folds; target and positive scores remain unchanged. Both experiments use the same estimate without further fitting.

Tab. 11 checks both negative–positive comparisons and the identity of sensitive pools. Boundary events include negatives already above the positive in M_T. Ratios pool counts within each model and dataset.

Table 12: Reversals among 120 pairs per row and metric. Fixed only means a negative fixed-count minimum and a nonnegative path minimum. Overall averages directions before orienting pairs.

		Hit@1			MRR		
Dataset	Scope	Path	Fixed	Fixed only	Path	Fixed	Fixed only
Flickr30K	Overall	19	52	33	15	49	34
Flickr30K	T→I	33	63	30	32	61	29
Flickr30K	I→T	20	40	20	17	39	22
OmniSET	Overall	18	66	48	14	60	46
OmniSET	T→I	29	70	41	29	67	38
OmniSET	I→T	32	58	26	26	57	31
noCaps	Overall	28	73	46	21	69	49
noCaps	T→I	35	68	34	39	68	31
noCaps	I→T	34	55	25	33	50	21

Table 13: Aggregate witnesses. Leader and Other are scores under the shared assignment; the remaining columns are gaps. Both metrics and their gaps are multiplied by 100.

Dataset	Metric	Baseline	Path min.	Leader	Other	Witness gap
Flickr30K	Hit@1	5.7	3.2	80.6	92.5	−11.9
Flickr30K	MRR	3.3	2.0	88.2	95.7	−7.5
OmniSET	Hit@1	4.1	2.9	78.7	90.5	−11.8
OmniSET	MRR	2.1	1.0	86.9	94.8	−7.9
noCaps	Hit@1	7.5	3.1	55.9	71.0	−15.2
noCaps	MRR	5.3	1.9	71.7	81.9	−10.2

C Pairwise Model Comparisons

C.1 Shared-Assignment Audit

We orient each comparison toward the strict leader at the recorded $\lambda = 0.5$ baseline. The path audit minimizes five aggregate gaps. The fixed-count audit averages per-pool minima over $4{,}900$ assignments shared by both models, retaining minimizing assignments as witnesses. Negative gaps indicate strict reversals. Signs use exact Hit@1 counts and reciprocal-rank numerators.

Assignments may differ across pools and model pairs; repeated entities need not keep the same modality across pools. These are attainable pairwise reversals, not one reversed leaderboard or reversal probabilities.

C.2 Reversal Counts and Examples

Tab. 12 includes both directions and their aggregate; all baselines are strict. OmniSET's 48 overall Hit@1 fixed-only reversals include 47 path-positive pairs and one path-tied pair.

Tab. 13 gives path-positive reversal examples: Qwen3-VL-Embedding-2B versus VLM2Vec-V2 on Flickr30K, and GME-Qwen2-VL-2B-Instruct versus Qwen3-VL-Embedding-2B on OmniSET and noCaps. The first model is the baseline leader; both models use the same assignment within each pool.