MIND THE MARGIN: WHEN CROSS-MODAL GEOMETRY MISLEADS RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc transformations of pretrained cross-modal embeddings, such as compressions and alignment maps, change retrieval in ways the standard measures of cross-modal structure do not register. Across 98 encoder pairs, three datasets, four modalities and 16 transformations, alignment maps close the modality gap by 63–91% on MSCOCO while removing 18–73% of Recall@1, and CKA and canonical correlation stay exactly unchanged under an orthogonal map that removes 40% of it. We explain the loss through the comparisons that decide retrieval. A transformation perturbs the similarity difference between two candidates, and we read off the ambient margin below which orderings become unreliable, the order-preservation threshold. Under a location–scale model, the expected loss depends on the transformation only through this scale, its severity, and on the target only through its top-two similarity gaps, its exposure. We show supporting evidence for the account. The threshold is a property of the transformation, its median stable to within 2% from 500 to 5,000 gallery items, and estimated without labels on reference encoder pairs it orders the losses of the compressions at $\rho = 0.72$ to 0.90 on three datasets. Read against it, a target's gaps order targets under a fixed compression at a mean ρ of 0.59 on MSCOCO and 0.53 on MSR-VTT, and order 29 encoder pairs by their mean loss at $\rho = 0.76$, where ambient accuracy and separability do not. Together they forecast a transformation's cost before it is applied, from an unlabelled retrieval log, for settings in which the target's embeddings are out of reach.

1 INTRODUCTION

Post-hoc transformations of pretrained embeddings are routine. Compressions (Jegou et al., 2010; Johnson et al., 2019) reduce the memory and latency of serving an embedding corpus, and alignment maps (Choi et al., 2023; Norelli et al., 2023; Moschella et al., 2022; Li et al., 2025; Maniparambil et al., 2025) re-shape the cross-modal structure of pretrained encoders. Their cost to retrieval is measured only *after* they are applied and scored against ground truth, and their effect on the embedding space is judged by measures of cross-modal structure such as the modality gap (Liang et al., 2022), the true-pair cosine, separability, centred kernel alignment (Kornblith et al., 2019) and canonical correlation.

Those measures do not register the cost. Across 3 datasets, 4 modalities, 98 encoder pairs and retrieval directions and 16 transformations—12 compressions and 4 alignment maps from prior work, applied singly or in composition—Recall@1 falls in 1,760 of 1,932 configurations while the five measures move in either direction. The alignment maps make the gap explicit: across the 29 MSCOCO encoder pairs they close the modality gap by 63–91% and raise the mean true-pair cosine by 57–169% while removing 18–73% of Recall@1, and an orthogonal map leaves CKA and canonical correlation unchanged to machine precision while removing two fifths of it. Yi et al. (2026) reported the dissociation for the modality gap.

The reason is where retrieval is decided. Recall@1 turns on whether the correct candidate outranks its strongest rival, a single comparison per query, while the measures aggregate over the whole space. We model a transformation as a perturbation of the similarity difference between two candidates and read off the ambient margin below which orderings become unreliable, the *order-preservation threshold* δ^*. Under a location–scale perturbation model, preservation depends on the margin only

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.

through its ratio to the perturbation scale, so the expected rate of order reversal depends on the transformation only through that scale, its *severity*, and on the target only through its retrieval-critical margins, its *exposure*. On a correctly answered query, the decisive margin is the gap between the two highest untransformed similarities, readable without labels.

We test the account on MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) for image–text retrieval and on MSR-VTT (Xu et al., 2016) for retrieval among text, video and audio, showing empirically that both factors are supported. The threshold behaves as a property of the transformation: its median stays within 2% as the gallery grows from 500 to $5{,}000$ items, the pairs that decide Recall@1 follow its symmetric law, and estimated on reference encoder pairs without labels it orders the losses of the compressions at $\rho = 0.72$ to 0.90 on the three datasets. The target's gaps carry its exposure, which no reference quantity supplies. Read against the threshold, they order the targets of a fixed compression at a mean ρ of 0.59 on MSCOCO, 0.53 on MSR-VTT and 0.18 on Flickr30k, and they order the 29 MSCOCO pairs by mean loss at $\rho = 0.76$, where their ambient Recall@1, separability and width reach at most 0.33. The gaps need the transformation's own threshold: read against one threshold for a whole family of compressions, they lose most of their correlation with the loss.

The account yields a forecast of a transformation's cost that reads two similarities per query from an unlabelled retrieval log and applies nothing to the target. Its *risk score* is the fraction of queries whose gap falls below the calibrated threshold, and its *predicted loss* adds the target's position among the reference risks to the average loss the transformation causes on the references (Xia et al., 2020). The predicted loss improves on that average at Recall@5 in all five compression settings, a variant using no label at any stage ranks the losses as well as the labelled one, and the gap alone flags the queries a compression breaks at a mean AUROC of 0.72 to 0.75. Pooled over transformations, the loss is dominated by which transformation is applied, which the reference average reads directly, and there the two terms raise its rank correlation in 19 of 26 settings. Scoping our forecast, it serves settings that expose only similarities, such as retrieval logs and scoring interfaces, or that precede the target's pair, such as the choice of an alignment map.

Our **contributions** are:

1. **A diagnostic for structure measures.** Post-hoc alignment maps close the modality gap and raise the true-pair cosine while removing Recall@1, and CKA and canonical correlation are provably invariant to orthogonal maps that remove about two fifths of it, over $1{,}932$ configurations, two encoder provenances and both loss normalizations.
2. **A margin account and evidence for both of its factors.** Under a location–scale perturbation model, the loss depends on the transformation only through its perturbation scale and on the target only through its decisive margins. The measured threshold is a property of the transformation that orders losses without labels, and the target's gaps order targets and encoder pairs where accuracy and separability do not.
3. **A label-free forecast from retrieval logs.** The account's two factors, read before the transformation is applied, rank and size its cost from an unlabelled log, flag the queries it breaks, and choose a compression under a tolerance, with the boundary set by what a transformed subsample of the target reveals when one is available.

2 Preliminaries and Experimental Setup

Cross-modal retrieval. Two encoders map inputs from modalities A and B into a common embedding space. A query x from modality A ranks candidate embeddings $\{y_1, \dots, y_N\}$ from modality B by the cosine similarity $s(x, y)$, one of which is its paired candidate y^+, and we score the ranking by Recall@K, the fraction of queries whose paired candidate appears among the top K.

Encoder provenance. *Jointly pretrained* pairs are trained contrastively on paired multimodal data (Radford et al., 2021), so their embeddings already share cross-modal structure useful for retrieval. *Independently pretrained* pairs are trained separately, and a post-hoc alignment map has to establish that structure in the first place before any transformation can perturb it.

Measures of cross-modal structure. Five common measures judge a pair of spaces: separability, the probability that a query's true partner outranks a random mismatch, with null value 0.5 at any gallery size, the modality gap between the two modality centroids (Liang et al., 2022), the mean cosine of true pairs, linear centred kernel alignment (CKA) (Kornblith et al., 2019), and the mean squared canonical correlation. Equation 7 in Appendix B defines each.

Local similarity and candidate ordering. Candidate ordering is determined by local similarity differences. For two candidates y_i and y_j of the same query x, their *ambient margin* is

$$m(x; y_i, y_j) = s(x, y_i) - s(x, y_j), \tag{1}$$

computed before any post-hoc transformation, written m_{ij} whenever the query is fixed, and positive exactly when y_i outranks y_j. For Recall@1, the comparison that matters is between the correct candidate and its strongest rival, the *decisive margin*

$$m^\star(x) = s(x, y^+) - \max_{y \neq y^+} s(x, y), \tag{2}$$

which is positive exactly when the correct candidate is ranked first. Margins at other depths determine the corresponding local ordering.

Embedding transformations. A post-hoc transformation T acts on an encoder pair's embeddings, giving the transformed similarity $s_T(x, y) = s(T(x), T(y))$ used for re-ranking candidates. We study two transformation classes. Post-hoc alignment maps include canonical correlation analysis (CCA) and its multi-way generalization (GCCA) (Hotelling, 1992; Carroll, 1968; Kettenring, 1971), orthogonal Procrustes and least-squares regression (Choi et al., 2023), and ASIF (Norelli et al., 2023), all of which optimize cross-modal agreement. Compressions comprise PCA, Gaussian random projection (Johnson et al., 1984) and quantization, which reduce dimensionality or numerical precision. Together, they give 16 transformations applied singly or in composition.

Datasets and encoder pairs. We work with MSCOCO (Lin et al., 2014), Flickr30k (Young et al., 2014) and MSR-VTT (Xu et al., 2016). The jointly pretrained group contributes 85 pairs and retrieval directions: author-released contrastive checkpoints on convolutional and transformer backbones over twelve pretraining corpora, tri-modal video–text–audio models, video and audio specialists, and two image–text pairs we train from scratch, with ambient Recall@1 from 2.0% to 87.7%. The independently pretrained group contributes 16 pairs, of which 13 units are used once an alignment map has created their ordering (Appendix B for details).

The phenomenon we show. A *configuration* is one target with one transformation applied, evaluated before and after on fixed candidate sets through Recall@K and the five measures of Equation 7. In Section 3 we report loss relative to the target's ambient Recall@K, keeping targets comparable even when their ambient accuracy differs by an order of magnitude. Appendix C repeats the phenomenon and the predictions in absolute Recall points, where the predicted relative loss multiplied by the ambient Recall@1 orders the absolute losses at a median ρ of 0.911. Applying all 16 transformations to each jointly pretrained pair yields $1{,}360$ configurations. For independently pretrained pairs, an alignment map first creates cross-modal structure and ordering, and the twelve compressions then act on the aligned space, giving 572 configurations over 50 aligned spaces, two further spaces with aligned Recall@1 near zero being set aside and reported separately in Appendix D.

Predictor calibrations and evaluation settings. The order-preservation threshold (Section 4) is calibrated in one of three target-free ways: on other encoder pairs from the target's dataset, on a different dataset, or on jointly pretrained MSCOCO pairs when the target is an independently pretrained pair first aligned and then compressed. A *setting* pairs one group of targets with one such reference set. Ranking and loss prediction use 26 settings, three calibrations on each of three datasets for each transformation class and one per first-stage map after alignment under each of two calibrations, and flagging broken queries uses 10.

Baselines. The reference average (Xia et al., 2020), the mean loss a transformation causes on the reference pairs, is the principal comparator on every axis and the first term of our predictor. NQC (Shtok et al., 2012) and WIG (Zhou & Croft, 2007), averaged over the target's queries and mapped to

a loss by a line fitted on the references, and the constant mean loss complete the configuration-level baselines, and NQC, WIG and two perturbation-based predictors for dense retrieval (Arabzadeh et al., 2023; Datta et al., 2025) score single queries. Applying the transformation to the whole target, or to a subsample of it, and counting the top-1 decisions it changes gives the reference for what the transformed target itself reveals.

3 The Retrieval Cost of Post-hoc Transformations

In Figure 1, we report the Recall@1 degradation phenomenon and the trend the cross-modal structure measures of Equation 7 follow across the configurations of Section 2. Recall@1 falls in 1,760 of them (median loss 48.9%), is unchanged in 44, and increases in 128 by a median of 0.30 Recall points. The five measures move in either direction: over the 1,804 configurations where Recall@1 does not improve, each has a median change within 8%, the modality gap shrinks in 48% of them and the mean true-pair cosine rises in 35%. The spaces aligned post-hoc lose Recall@1 in 501 of their 572 configurations (median 41.1%). Appendix D reports every measure on each dataset and on the aligned spaces.

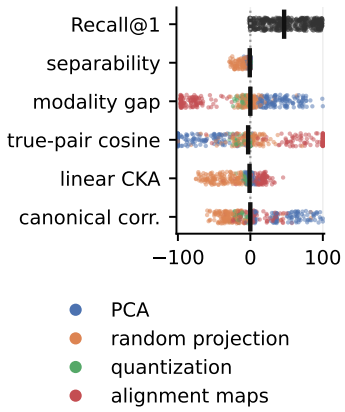


Figure 1: Relative change (%) of Recall@1 and the five measures over the 464 MSCOCO configurations. Bars mark medians and values are clipped to ±100%. Recall@1 is plotted as a loss, so its points lie to the right.

Only separability, which needs the true pairs and the transformed embeddings, follows retrieval loss closely: rank correlations with the relative Recall@1 change are −0.01 for the modality gap, 0.06 for true-pair cosine, 0.32 for canonical correlation, 0.51 for CKA, and 0.80 for separability. However, it follows only the direction of the loss and not its size: among the configurations that lose Recall@1, the relative loss exceeds the relative change in separability in 99%, by a median factor of 18 (Table 5). The four alignment maps make the imbalance explicit. Across the 29 MSCOCO pairs they close the modality gap by 63–91% and raise the mean true-pair cosine by 57–169%, while Recall@1 falls by 18.4–72.7% and mean separability stays within 1.2% of ambient (Appendix E gives more depths and the pairwise spread). Of the 340 alignment-map configurations only 11 gain Recall@1, at a median ambient 9.3% versus 46.6% for those that lose. On independently pretrained pairs, the same maps create the ordering instead (Appendix H.1), showing that retrieval-critical margins must already exist before alignment can destroy them.

We next ask what determines how much Recall@1 a transformation removes from a given encoder pair, and why the standard measures register so little of it.

4 Severity and Exposure: A Margin Account

We characterize the phenomenon of Section 3 through the candidate orderings that decide retrieval. The distinction that matters is between the margin separating two candidates and the perturbation a transformation induces on it. This yields two factors the rest of the paper works with, the *severity* of the transformation and the *exposure* of the target.

Similarity orderings and order-preservation threshold. For two candidates y_i and y_j of a query x, the ambient margin is that of Equation 1. After applying T, the transformed margin is

$$m_T(x; y_i, y_j) \;=\; s\big(T(x), T(y_i)\big) - s\big(T(x), T(y_j)\big), \tag{3}$$

and the transformation reverses the ordering exactly when the two differ in sign. An ordering is vulnerable when its ambient margin m is small relative to the scale of the induced perturbation $\Delta m = m_T - m$, which calls for a transformation-specific measure of order instability.

Definition 1 (Order-preservation threshold)**.** *Given a transformation T, the* order-preservation threshold δ^* *is the characteristic scale of the ambient margin m_{ij} at which the probability of preserving the ordering of a candidate pair rises above a fixed criterion.*

Proposition 1 (Order preservation). *Assume the perturbation that T induces on the similarity difference of a candidate pair is symmetric about zero, with a continuous and strictly increasing distribution function and scale σ_T. Then the probability that the ambient order of (y_i, y_j) survives T is a monotone function of m_{ij}/σ_T, approaching 0.5 as $m_{ij} \to 0$ and 1 as $m_{ij} \to \infty$. Moreover, under this model, the corresponding threshold scale is proportional to σ_T, with a proportionality constant determined by the perturbation distribution.*

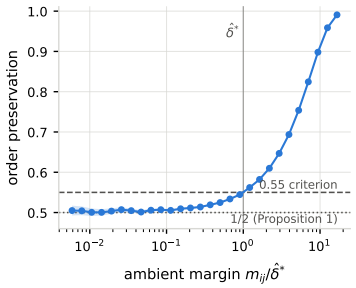


Figure 2: Measured probability that ambient ordering survives the canonical correlation solve, against its ambient margin in units of the pair's $\hat{\delta}^*$; pool of 29 MSCOCO pairs with a 95% band. Dotted line marks the 0.5 that Proposition 1 predicts at zero margin; dashed line the 0.55 criterion defining $\hat{\delta}^*$.

The proof is given in Appendix F. Each transformation thus has a characteristic scale of order instability, and each encoder pair a distribution of ambient margins, among which the few that decide the retrieval outcome determine the loss.

Corollary 1 (Severity monotonicity). *For a fixed ambient margin $m > 0$, the probability that T reverses that ordering is non-decreasing in σ_T. Consequently, for a target with a fixed distribution M of decisive margins (Equation 2) over the queries it answers correctly, the expected fraction of those queries that T answers incorrectly is non-decreasing in σ_T.*

Corollary 2 (Margin-dominance ordering). *Fix σ_T. If one target's distribution of decisive margins first-order stochastically dominates another's, its expected rate of order reversal over correctly answered queries is no greater.*

Both are proved in Appendix F, and together justify the two factors the forecast of Section 5 reads: the severity of the transformation, which enters only through σ_T and which $\hat{\delta}^*$ estimates, and the exposure of the target, which enters only through the ordering of its retrieval-critical margins.

Proposition 1 fixes δ^* only up to the perturbation scale σ_T and the shape of its distribution, neither of which a transformation gives in closed form, so we read the threshold off the preservation curve the transformation produces on reference pairs. We estimate $\hat{\delta}^*$ as the margin at which measured order preservation first exceeds 0.55, interpolating between margin-bin centres. The criterion sits just above the 0.5 that the proposition predicts at zero margin, so $\hat{\delta}^*$ marks the onset of reliable ordering. We fix it before examining retrieval loss, and raising it to 0.60 or 0.65 changes the correlation between risk and Recall@1 loss by at most 0.016 under compression and 0.037 under alignment (Appendix G).

Measured preservation approaches 0.5 as the ambient margin vanishes and rises towards 1 as it grows (Figure 2, Appendix F). On the pairs that decide Recall@1, each query's two highest candidates, it follows the same law at a median 0.47 of the random-pair scale, and the median threshold stays within 2% as the gallery grows from 500 to $5{,}000$ items, as a property of the transformation should. Read against that fixed threshold, the risk score rises with the gallery as the loss does, in the same direction in 91% of the 348 compression configurations (Appendix H).

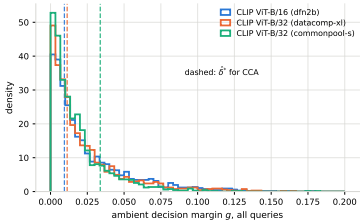


Figure 3: Ambient top-1 decision margins of the smallest-, median- and largest-risk MSCOCO pair under CCA, with each pair's $\hat{\delta}^*$ dashed.

Decisive margins. A correctly answered query becomes incorrect when the induced perturbation drives its decisive margin (Equation 2) non-positive, and Recall@K depends in the same way on the margins around the K-th position. On a correctly answered query, the strongest rival is the second-ranked candidate, so the decisive margin coincides with the top-1 decision margin (Scheffer et al., 2001)

$$g(x) \;=\; s\big(x, y_{(1)}\big) - s\big(x, y_{(2)}\big), \tag{4}$$

where $y_{(1)}$ and $y_{(2)}$ are the two highest-ranked candidates. Unlike $m^\star$, g needs no knowledge of which candidate is correct. The 29 pairs differ in how much of their decision-margin distribution lies below the transformation-specific $\hat{\delta}^*$ (Figure 3, all 29 in Figure 10).

Why standard measures understate the effect. The five measures (Eq. 7) aggregate over comparisons that do not decide the top of a ranking. Separability averages over every mismatch of every query, while Recall@1 turns on one comparison per query, the one with the smallest margin and therefore the first to fall below $\hat{\delta^*}$. The reordering is local to the top: under CCA on the 29 MSCOCO pairs, the ambient and transformed top-5 sets overlap by 0.381 on average, and 83.5% of the candidates entering the transformed top 5 come from within ambient rank 50 of a $2{,}000$-item gallery. The modality gap is a function of two centroids alone, so it cannot see competitors that decide a query. Mean true-pair cosine reads matched items only, and on MSCOCO every alignment map raises it while removing Recall@1. CKA and canonical correlations depend on cross-modal covariance only through its singular values, so they record how much structure the two spaces share and not how that structure is oriented, while every cross-modal similarity is an inner product between a vector of one space and a vector of the other. An orthogonal Procrustes map re-orients one space, which leaves both statistics unchanged to machine precision, while Recall@1 falls by 32.9 to 48.5% (Appendix F proves both invariances). The fraction of top-1 decision margins below $\hat{\delta^*}$ combines the transformation's scale with the target's exposure in one number, which Section 5 tests against measured Recall@1 loss.

Scale against margins. The account predicts that the loss depends on a transformation's scale relative to a pair's own margins, and not on either alone. The two ResNet-50 pairs we train from scratch have wide decisive margins, with medians of 0.0431 and 0.0359 over their correctly answered queries, against a range of 0.0120 to 0.0409 for the 27 public pairs. Orthogonal Procrustes and least-squares regression, at $\hat{\delta^*}$ of 0.0046 to 0.0088 on these pairs, change their Recall@1 by at most 4.3%, while ASIF, at $\hat{\delta^*}$ of 0.0504 and 0.0576, removes 62.7% and 69.0% of it. The same arithmetic explains why a distortion guarantee does not protect retrieval. Gaussian random projection to rank $k = O(\epsilon^{-2} \log n)$ preserves squared distances among n points within a factor $1 \pm \epsilon$ (Johnson et al., 1984), and on unit-normalised embeddings, where $\|x - y\|^2 = 2 - 2\, s(x, y)$, this bounds the distortion of every similarity, and hence of every margin, at $O(\epsilon)$. An ordering with ambient margin m is then protected only once ϵ falls to the order of m, which at thresholds of order 10^{-2} (Table 15) asks for a rank orders of magnitude above the input width. At rank 256 on CLIP ViT-B/16, random projection costs 0.55% of separability and 38.2% of Recall@1 (Table 8).

5 Forecasting the Cost from a Retrieval Log

Both factors of the account can be read before a transformation is applied. We calibrate $\hat{\delta^*}$ on reference pairs, read the target's untransformed margins against it, and test the resulting forecast on loss size, ranking, single queries and the choice of a compression under a predefined tolerance.

Threshold calibration. For a transformation T we estimate $\hat{\delta^*}$ on a small set of reference encoder pairs, to which T is applied so that both ambient and transformed similarities are available, and take the median. The estimate reads only whether candidate orderings survive T, never which candidate is correct. One threshold is estimated per transformation, so the cost of screening a target grows with candidate transformations and not with references. A single reference pair retains 96% of the full-calibration correlation for image–text compression, and five come within 5% of calibrating on all of them in every setting at Recall@1 (Appendices G and H). Every forecast below calibrates on references that exclude the target and never applies the transformation to the target.

A risk score from ambient margins. For a target we compute only its ambient similarities and the top-1 decision margin $g(x)$ of Equation 4, which we read against $\hat{\delta^*}$. The risk score is

$$r = \Pr\big(g(x) < \hat{\delta^*}\big), \tag{5}$$

the fraction of queries whose top-1 decision lies within the transformation's order-instability scale, which we use to predict the relative Recall@1 loss $(R_{\mathrm{amb}} - R_T)/R_{\mathrm{amb}}$. No relevance label enters at any stage, and g reads only two similarities per query, so a retrieval log suffices for every candidate transformation at once. Counting only the queries a target answers correctly, the only ones that can lose Recall@1, would require its labels, and the two counts rank configurations alike, at $\rho = 0.971$ under compression and 0.924 under alignment (Appendix G).

Table 1: Every within-dataset setting, the target's own group excluded from the references. Risk: rank correlation of the risk score with the relative Recall@1 loss, pooled and within each transformation. Severity (the reference average) and Both (Equation 6): MAE / ρ at Recall@1, with their MAE at Recall@5. Label-free: Both with the references' top-1 change in place of their loss. Transformed: the target itself compressed and its changed top-1 decisions mapped to a loss, undefined for alignment maps without the target's pairs. Other calibrations and the query-performance baselines in Appendix I.

		Risk ρ		Recall@1 MAE / ρ		Recall@5 MAE	Label-free	Target
Setting	n	Pooled	Within	Severity	Both	Severity / Both	Both ρ	transformed
Compres., MSCOCO	348	0.887	0.594	0.111/0.851	0.100/0.904	0.096/0.077	0.903	0.091/0.983
Compres., Flickr30k	324	0.741	0.176	0.081/0.902	0.088/0.904	0.061/0.054	0.899	0.090/0.924
Compres., MSR-VTT	348	0.804	0.532	0.156/0.776	0.143/0.861	0.144/0.098	0.863	0.124/0.927
Align., MSCOCO	116	0.500	0.176	0.054/0.789	0.060/0.892	–	–	–
Align., Flickr30k	108	0.451	0.272	0.055/0.802	0.057/0.846	–	–	–
Align., MSR-VTT	116	0.301	0.411	0.181/0.138	0.178/0.341	–	–	–

Table 2: Rank correlation with the relative Recall@1 loss of the threshold alone (Thr.), the target's margins read against each transformation's threshold (Per-T, the risk score) or one per family (Fam.), and their median (Margin), over the settings of Table 1.

Setting	Thr.	Per-T	Fam.	Margin
Compres., MSCOCO	0.830	0.887	0.252	0.240
Compres., Flickr30k	0.896	0.741	−0.030	−0.031
Compres., MSR-VTT	0.719	0.804	0.295	0.295
Align., MSCOCO	0.672	0.500	0.120	0.123
Align., Flickr30k	0.730	0.451	0.073	0.078
Align., MSR-VTT	0.138	0.301	0.331	0.289

Table 3: Mean AUROC for the queries a compression breaks, among those answered correctly before it, scored by small g, low NQC or low WIG.

Dataset	n	g	NQC	WIG
MSCOCO	348	**0.723**	0.633	0.525
Flickr30k	322	**0.746**	0.672	0.567
MSR-VTT	329	**0.751**	0.665	0.597

Predicting the magnitude of loss. Sizing the loss takes two terms. A transformation's *severity* is the model-wise mean relative Recall@1 loss $\bar{L}_T$ it causes on the reference pairs (Xia et al., 2020). The *exposure* of a target is its position within the reference risk distribution, and the predicted loss is

$$\hat{L}_T \;=\; \bar{L}_T \;+\; \beta \left(\hat{F}_T(r) - 0.5 \right), \tag{6}$$

where $\hat{F}_T$ is the empirical distribution function of the reference risks under T and β is the least-squares slope of the reference losses on the centred percentile, fitted within each transformation. A percentile makes the correction invariant to any monotone rescaling of the risk (Appendix I compares the raw difference $r - \bar{r}_T$). Calibration uses measured losses on the references, never on the target.

What the target's margins add. Pooled over transformations, the loss is dominated by which transformation is applied, which the reference average reads directly, and the threshold alone, one label-free value per transformation, already orders the compressions at $\rho = 0.830$, 0.896 and 0.719 (Table 2). Neither can separate targets. Within a single transformation both give every target the same value, while the risk score orders the targets of a compression at a mean ρ of 0.594 on MSCOCO, 0.532 on MSR-VTT and 0.176 on Flickr30k, where the transformation fixes 85% of the loss variance (Table 1), and over the 29 MSCOCO pairs it orders their mean loss under compression at $\rho = 0.755$, where their ambient Recall@1, separability and width reach 0.31, 0.33 and -0.10 (Appendix E). The ViT-B/32 and ViT-B/16 checkpoints trained on LAION-2B lose 52.3% and 50.8% of Recall@1 on average, against 67.6% and 64.3% for their OpenAI counterparts of identical architecture and width. The transformation's scale is what makes the margins informative: read against one threshold per family, the same margins lose most of their correlation with the loss in five of the six settings. Added to the reference average, the risk score raises its rank correlation in 19 of the 26 settings, most where the transformation explains least of the loss variance ($\rho = -0.51$

over the 26, a share computable from the reference runs alone), and lowers its Recall@1 error in 9 of them, though at Recall@5 it lowers the error in every compression setting (Section 5).

Query-performance predictors. Scores of the ambient system alone cannot tell transformations apart. Mapped to a loss by a line fitted on the references, NQC and WIG reach errors of 0.27 to 0.33 in the three within-dataset compression settings and the two perturbation-based predictors 0.28 to 0.33, against 0.09 to 0.14 for the two terms, and the risk score ranks the losses above NQC and WIG in 23 of the 26 settings (Table 16, Appendix I).

Alignment maps. Alignment maps are the case the forecast is built for, since a map cannot be fitted on a target before its pairs exist. Pooled over the four maps, the risk score ranks their losses at $\rho = 0.500$, 0.451 and 0.301 on MSCOCO, Flickr30k and MSR-VTT, and added to the reference average it raises the rank correlation in all three, from 0.789 to 0.892, from 0.802 to 0.846 and from 0.138 to 0.341, at errors within 0.006 of it (Table 1). Within a single map, the margins order targets under orthogonal Procrustes, at $\rho = 0.648$, and not under least squares, ASIF or CCA, whose losses the reference average sizes to within 0.032 to 0.060 (Appendix G).

Transfer. Thresholds calibrated on the 29 MSCOCO pairs rank compressions of the 27 Flickr30k pairs and of 19 MSR-VTT models at $\rho = 0.772$ and 0.849, and compressions of the ten independently pretrained image–text pairs after alignment at 0.902 (Table 11, with all 26 settings in Table 16). Rg transfers more readily than the size. Across datasets and provenances, the target's risks sit on a different scale from the references', the Flickr30k targets at a mean risk of 0.232 against 0.381 for the MSCOCO references, and exposure inherits that offset. The two terms then size the Recall@1 loss less closely than the reference average in most cross-dataset settings, as for Flickr30k compressions calibrated on MSCOCO, at 0.145 against 0.100, and least-squares alignment followed by compression, at 0.320 against 0.217, but references from the target's own corpus, which need no target label, bring the error below the reference average's on MSCOCO and MSR-VTT.

Product quantization. The forecast extends to the codecs of production vector search. Four product-quantized galleries from 256 to 32 bits and two with an optimized rotation, trained with FAISS (Johnson et al., 2019) and searched asymmetrically, remove 42% to 89% of Recall@1 from the 29 MSCOCO pairs. The risk score orders the targets of a single code at a mean ρ of 0.545, close to the 0.594 of the twelve compressions, where the reference average gives every target the same loss, and orders all six codes at 0.722, against 0.777 for the reference average, which reads the ordering of the codes directly (Appendix H).

Without any label. The severity term uses measured losses on the reference pairs. The references are transformed anyway to estimate the threshold, so the share of their top-1 decisions that a transformation changes can stand in for those losses, and the exposure slope can be fitted against it. The resulting predictor uses no label at any stage and never transforms the target, and it ranks the compression losses at $\rho = 0.903$, 0.899 and 0.863, against 0.904, 0.904 and 0.861 for the labelled one (Appendix I). The reference labels then serve only to size the loss.

Reading the margin at other depths. Recall@K is lost when a transformation reorders candidates across the K-th boundary, so the same argument reads the gap between the K-th and $(K{+}1)$-th highest untransformed similarities, with the threshold unchanged. At $K = 5$ the two terms lower the error of the reference average from 0.096 to 0.077 on MSCOCO, from 0.061 to 0.054 on Flickr30k and from 0.144 to 0.098 on MSR-VTT, and they lower its error and raise its rank correlation in all five compression settings (Table 21).

Identifying the queries a compression breaks. Each query carries its own flag. Among the queries answered correctly before a compression, a small g identifies those it breaks above NQC and WIG on all three datasets (Table 3), and above both perturbation-based predictors, of which the stability of the top-1 candidate under noise (Arabzadeh et al., 2023) approaches it from below as the noise grows, since it becomes a sampled reading of the same gap (Appendix I).

Transforming the target. When the target's embeddings are at hand, a compression can be applied to them and the share of top-1 decisions it changes mapped to a loss, the last column of

Table 1. That reading ranks the losses more closely, at $\rho = 0.983$, 0.924 and 0.927, at the price of one transformed retrieval over the target per candidate. The two terms, which never transform the target, come within 0.009 and 0.019 of its error on MSCOCO and MSR-VTT and below it on Flickr30k. On the full $5{,}000$-item MSCOCO corpus, applying a compression to a random 500-item subsample and mapping its share of changed top-1 decisions through a line fitted on the reference pairs predicts the full loss at an error of 0.071 and $\rho = 0.969$ in 9 ms per candidate, against 0.104 and 0.903 for the two terms, which read the whole log in one 61 ms pass shared by every candidate (Appendix I). A transformed subsample is therefore the more accurate estimate, but only where the target's embeddings can be transformed, and the two terms serve every setting where they cannot.

Choosing a compression under a tolerable loss. Tolerating a relative Recall@K loss, the predictor selects the most aggressive map of each compression family whose predicted loss stays within it, and an oracle that knows the target's losses picks the most aggressive admissible one. Over 423 choices, 3 families for each of 141 target and calibration pairings, the predicted loss holds to the tolerance more often than the reference average while stopping nearer the oracle at Recall@1 from a tolerance of 20%, at Recall@5 from 7.5%, and at Recall@10 at 4%, 5% and 10%. Tolerating 5% at Recall@10, it exceeds the tolerance in 6% of choices against 9% and stops 0.37 steps short of the oracle against 0.51. At tighter Recall@1 tolerances, the reference average, which forecasts one loss for every target and mostly declines to compress, exceeds them less often (Appendices I and J).

6 Related Work

Structure of cross-modal spaces. The modality gap (Liang et al., 2022), its relation to downstream performance (Yi et al., 2026) and post-hoc maps that close it or re-shape the joint space (Choi et al., 2023; Norelli et al., 2023; Moschella et al., 2022; Li et al., 2025; Maniparambil et al., 2025) are studied through aggregate measures such as CKA (Kornblith et al., 2019) and canonical correlation. We show these measures miss the comparisons that decide retrieval, and identify the margin-level quantity that does register them. **Predicting retrieval performance.** Query-performance predictors score a fixed system from its score distribution (Cronen-Townsend et al., 2002; Zhou & Croft, 2007; Shtok et al., 2012) or its stability under perturbations of the query or the space (Arabzadeh et al., 2023; Datta et al., 2025), so before a transformation they give every candidate the same value. Reference averages (Xia et al., 2020; Khramtsova et al., 2023; 2024) estimate a candidate's effect from reference systems and give every target the same value, and transferability estimators (Dai et al., 2024) and the compression literature (Jegou et al., 2010; Johnson et al., 2019) measure the cost on a labelled or transformed target. Our forecast reads the interaction of a particular transformation with a particular target, from the target's untransformed margins, and the gap it reads is the margin used in active learning (Scheffer et al., 2001). Appendix A extends the comparison.

7 Conclusions

Standard measures of cross-modal structure do not register what post-hoc transformations cost retrieval, because this is decided by few comparisons on top of each ranking. Read through those comparisons, the loss is governed by a transformation severity, its order-preservation threshold, and a target exposure, its top-two similarity gaps. The threshold behaves as a transformation property that orders its losses without labels, the gaps order targets and encoder pairs where ambient accuracy and separability do not, and CKA and canonical correlation are blind to re-orientations that remove two fifths of Recall@1. Read together from an unlabelled retrieval log, the two factors forecast a transformation's cost before it is applied, size it more closely than the reference average at Recall@5 in every compression setting, and rank it with no label at any stage.

Scope and limitations. The forecast serves settings in which target's embeddings cannot be transformed: a retrieval log, a scoring interface, or an alignment map chosen before the target's pairs exist. Where they can, a target's transformed subsample estimates a compression's loss more accurately, though at one transformed pass per candidate. The account is first-order: the pairs that decide Recall@1 follow its symmetric law at about half the random-pair scale, a factor that varies across configurations. Fitting transformations to preserve decisive margins, rather than aggregate agreement, is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools for drafting text. We did not use generative AI tools for defining the research topic or questions, conducting the literature review, planning the project, designing the methodology, collecting or analyzing data, or running experiments. All AI-assisted work was manually reviewed and verified by the authors. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, medical data, or sensitive personal information. All datasets and models used are publicly available, and we adhered to their respective licenses and intended research-only usage. The study does not raise concerns related to conflicts of interest, discrimination, fairness, privacy, security, or legal compliance. No potentially harmful applications or insights are introduced, and no integrity issues (such as IRB requirements or research-ethics risks) apply. To the best of our knowledge, this work poses no ethical concerns under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Every number in this paper comes from public datasets and publicly released encoder checkpoints, together with two ResNet-50 pairs trained from scratch with publicly released code and configuration, under a single fixed protocol. Section 2 states the datasets, encoder pairs and transformations, and Appendix B gives the splits, the map-fitting details and the full inventory of checkpoints. Appendix B.1 collects the measures taken to make the results reproducible, namely the fixed seeds and split fractions, the deterministic double-precision estimation of every transformation and threshold, and the resampling procedures used wherever randomness enters a reported quantity. The threshold aggregation used by the predictor is specified in Appendix G, its insensitivity to the order-preservation criterion in Appendix G, and every baseline in Appendix I. Proposition 1 is stated with its assumption made explicit and proved in full in Appendix F. Appendix K lists the license of every dataset and checkpoint we use. Anonymized code covering feature extraction, all sixteen transformations, the threshold estimator, the baselines and every figure and table in the paper accompanies the submission as supplementary material.

REFERENCES

Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. BERT-QPP: contextualized pre-trained transformers for query performance prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2857–2861, 2021.

Negar Arabzadeh, Radin Hamidi Rad, Maryam Khodabakhsh, and Ebrahim Bagheri. Noisy perturbations for estimating query difficulty in dense retrievers. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 3722–3727, 2023.

J Douglas Carroll. Generalization of canonical correlation analysis to three or more sets of variables. In *APA 76th Annual Convention, San Francisco, CA, August 30-September 3, 1968*, 1968.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Hyunjin Choi, Hyunjae Lee, Seongho Joe, and Youngjune Gwon. Is cross-modal information retrieval possible without training? In *European Conference on Information Retrieval*, pp. 377–385. Springer, 2023.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations*, volume 2025, pp. 42128–42149, 2025.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. *Advances in Neural Information Processing Systems*, 38: 112004–112022, 2026.

Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 299–306, 2002.

Mengyu Dai, Amir Hossein Raffiee, Aashish Jain, and Joshua Correa. Evaluating transferability in retrieval tasks: an approach using mmd and kernel methods. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22390–22400. IEEE, 2024.

Suchana Datta, Guglielmo Faggioli, Nicola Ferro, Debasis Ganguly, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonellotto. Projection-displacement-based query performance prediction for embedded space of dense retrievers. *ACM Transactions on Information Systems*, 44(1):1–30, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sedigheh Eslami and Gerard De Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024.

Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, and Benjamin Piwowarski. Query performance prediction for neural IR: Are we there yet? In *European Conference on Information Retrieval*, pp. 232–248. Springer, 2023.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, pp. 36221–36237, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 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.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

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.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019.

William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, Xi Wang, and Guido Zuccon. Selecting which dense retriever to use for zero-shot search. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 223–233, 2023.

Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, and Guido Zuccon. Leveraging llms for unsupervised dense retriever ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1307–1317, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Po-han Li, Sandeep Chinchali, et al. Csa: Data-efficient mapping of unimodal features to multimodal features. In *International Conference on Learning Representations*, volume 2025, pp. 99459–99478, 2025.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E O'Connor. Harnessing frozen unimodal encoders for flexible multimodal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29847–29857, 2025.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pp. 1–18. Springer, 2022.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *Advances in Neural Information Processing Systems*, 36:15303–15319, 2023.

David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4055–4063, 2022.

Longtian Qiu, Shan Ning, and Xuming He. Mining fine-grained image-text alignment for zero-shot captioning via text-only training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4605–4613, 2024.

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.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden Markov models for information extraction. In *International symposium on intelligent data analysis*, pp. 309–318. Springer, 2001.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems (TOIS)*, 30(2): 1–35, 2012.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Alex H Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. *bioRxiv*, pp. 2024–10, 2024.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8625–8646, 2020.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference on Learning Representations*, volume 2024, pp. 47812–47831, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Lingjie Yi, Raphael Douady, and Chao Chen. The convergent representation of contrastive vision-language models: Geometry, modality gap and shared space alignment. In *International Conference on Machine Learning*, volume 306. PMLR, 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.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Yun Zhou and W Bruce Croft. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 543–550, 2007.

A Extended Related Works

Performance prediction for retrieval. Query-performance prediction estimates how well a fixed retrieval system answers each query, from the query itself or from its retrieved scores, with pre-retrieval predictors and post-retrieval predictors such as Clarity (Cronen-Townsend et al., 2002), WIG (Zhou & Croft, 2007) and NQC (Shtok et al., 2012), and is evaluated by the correlation between predicted and measured per-query effectiveness. Supervised estimators such as BERT-QPP (Arabzadeh et al., 2021) extend the task to neural rankers, and Faggioli et al. (2023) evaluate predictors on neural rankers, dense retrievers included. For dense retrievers, perturbation-based predictors estimate a query's difficulty from how stable its ranking stays when the query embedding is perturbed (Arabzadeh et al., 2023), or when queries and retrieved documents are projected onto subspaces of pseudo-relevant documents (Datta et al., 2025). Those perturbations are set by the predictor and are the same for every system change. Our threshold is the perturbation scale that a specific candidate transformation induces, measured on reference pairs, so one target receives a different risk under each transformation. We use NQC and WIG at that native, per-query level and as configuration-level scores. The gap between the two highest scores is a classical confidence signal, which margin sampling uses to select uncertain examples in active learning (Scheffer et al., 2001). Margin sampling ranks a model's own predictions by that gap to decide which examples to label. We read the gap of an untransformed retrieval system against the perturbation scale a candidate transformation induces, estimated on reference pairs, to forecast which decisions that transformation will reverse before it is applied. Read on its own, the gap orders the losses poorly, at $\rho = 0.240$ on MSCOCO compression against 0.887 once read at the transformation's scale (Table 2), so the scale is what turns a confidence signal into a predictor. Transferability estimation ranks pretrained models by their expected retrieval performance after fine-tuning on a labelled target task (Dai et al., 2024). Performance prediction from records of other runs estimates the score of an unseen configuration, and Xia et al. (2020) compare their predictor with mean value baselines, among them the model-wise mean, the average score of the same setting over all other models, which is our severity term. For selecting a dense retriever for a collection without relevance labels, ranking retrievers by their labelled effectiveness on a reference collection is the strongest of the earlier selection methods (Khramtsova et al., 2023; 2024). Our setting fixes the encoder pair and the corpus and varies a post-hoc transformation that has not yet been applied to the target, which the margin-based risk score addresses from the target's ambient similarities.

Post-hoc cross-modal alignment. A line of work establishes cross-modal retrieval between frozen, independently pretrained encoders without training either of them. Choi et al. (2023) obtain retrieval from raw BERT and RoBERTa features using least-squares and orthogonal Procrustes maps, and Li et al. (2025) solve canonical correlation analysis over frozen encoders. ASIF (Norelli et al., 2023) builds a common space non-parametrically by representing each input through its similarities to a set of anchor pairs, retaining the anchor set at inference, and Maniparambil et al. (2025) align frozen unimodal encoders with a learned mapping. Relative representations (Moschella et al., 2022) use a related anchor-based device for communication between latent spaces of the same modality. These methods supply the transformations we analyse. Post-hoc transformations are applied to jointly pretrained spaces as well, for purposes other than retrieval. Closing the modality gap of CLIP lets captioning and question-answering models trained on text alone run on images (Nukrai et al., 2022; Qiu et al., 2024), residual adapters on frozen CLIP features improve few-shot classification (Gao et al., 2024), and aligning a further modality into a frozen jointly pretrained space extends it (Girdhar et al., 2023). Each changes the cross-modal geometry, and our contribution is to characterize and predict the retrieval loss they induce from the target's margin distribution read at the scale each transformation sets.

Representation compression. Reducing embedding dimensionality or numerical precision to control memory and latency is long-standing practice in retrieval systems (Jegou et al., 2010; Johnson et al., 2019), and the accuracy cost is well quantified. Matryoshka representations train an embedding so that its leading coordinates remain usable at several widths (Kusupati et al., 2022). Gaussian random projection carries distortion guarantees in the sense of Johnson et al. (1984). That literature measures the cost of compression after it is applied, and we predict which similarity orderings survive it before it is applied to the target, finding the same signature under compression as under maps fitted for alignment.

Representation similarity and the modality gap. Centred kernel alignment and related indices (Kornblith et al., 2019) are widely used to compare representations, and linear centred kernel alignment is the mean squared canonical correlation under a whitening change of variables (Williams, 2024), so an improvement in that statistic is not independent of the correlation structure an alignment map already manipulates. A parallel line studies the modality gap in contrastive embeddings. Shifting the gap after training changes zero-shot classification in a direction that depends on the task (Liang et al., 2022), and methods that close it by retraining the encoders report retrieval gains (Eslami & De Melo, 2024), or marginal retrieval gains together with marked gains on group-level tasks such as clustering (Grassucci et al., 2026). The latter act on training and lie outside the post-hoc setting studied here. Yi et al. (2026) is our closest point of contact: they identify shared-space alignment as the quantity governing downstream performance, and report that post-hoc transforms which reduce the gap do not improve retrieval. Section 4 supplies a mechanistic account of this observation.

Retrieval evaluation. Recall@K is the standard measure for cross-modal retrieval (Radford et al., 2021; Girdhar et al., 2023). Our analysis introduces the decisive margin as the quantity that determines Recall@1, separates it from whole-ranking separability, and reads it against a threshold calibrated for the candidate transformation, so the prediction turns on the interaction between a transformation's scale and a target's margins, before the transformation is applied and without ground truth on the target.

B Detailed experimental setup

Protocol. Every reported number comes from a single implementation with a three-way split of the evaluation corpus at fractions $0.40/0.20/0.40$ under a fixed seed. Maps are fitted on the first part, every configuration choice is selected on the second by validation Recall@1, and the reported number is measured once on the third, which selection never touches. Two controls accompany the experiments. The *do-nothing* control applies no transformation, so its Recall@K is the ambient value by construction. It establishes that the split, the cached features and the evaluation code introduce no change of their own, and it fixes the reference against which every loss is measured. The *shuffled* control permutes the pairing while preserving all marginals, destroying the cross-modal correspondence and leaving every marginal statistic intact. It gives the level each quantity reaches

with no correspondence to read, separating a relation carried by the pairing from one a measure would report from the marginals alone.

Datasets. MSCOCO (Lin et al., 2014) contributes $5{,}000$ paired image–text items, the $5{,}000$ images of its 2017 validation split each with its first caption, which the $0.40/0.20/0.40$ split turns into $2{,}000$ fitting, $1{,}000$ selection and $2{,}000$ evaluation items, so its retrieval gallery is the $2{,}000$ evaluation candidates. Flickr30k (Young et al., 2014) contributes the $2{,}014$ images of the validation and test splits of Karpathy & Fei-Fei (2015), each with its first caption, giving 805, 402 and 807 items, so its gallery is 807 candidates. MSR-VTT (Xu et al., 2016) contributes 884 items, giving 353, 176 and 355, so its gallery is 355 candidates. Chance Recall@K therefore differs across the three and is quoted explicitly where necessary.

Jointly pretrained image–text pairs. The 29 MSCOCO pairs comprise 27 publicly released checkpoints and two ResNet-50 pairs we train from scratch, one under a plain contrastive objective and one under a modality-gap-regularized variant (Grassucci et al., 2026). The public checkpoints are the OpenAI CLIP models RN50, RN101, RN50x4, ViT-B/32, ViT-B/16 and ViT-L/14 (Radford et al., 2021), with ResNet (He et al., 2016) and ViT (Dosovitskiy, 2020) backbones, and OpenCLIP models (Cherti et al., 2023) trained on LAION-2B (Schuhmann et al., 2022) (ViT-B/32, ViT-B/16, ViT-L/14, ConvNeXt-B-w (Liu et al., 2022) and CoCa ViT-B/32 (Yu et al., 2022)), LAION-400M (Schuhmann et al., 2021) (ViT-B/32 and ViT-B/16), YFCC-15M (Thomee et al., 2016) (RN50 and RN101), CC12M (Changpinyo et al., 2021) (RN50) and three DataComp pools (Gadre et al., 2023), namely DataComp-XL (ViT-B/32, ViT-B/16 and ViT-L/14), DataComp-M and CommonPool-S (one ViT-B/32 each). They are completed by MetaCLIP ViT-B/32 (Xu et al., 2024), DFN-2B ViT-B/16 and ViT-L/14 (Fang et al., 2024), EVA02-CLIP B/16 (Sun et al., 2023), and SigLIP ViT-B/16 (Zhai et al., 2023) and SigLIP 2 ViT-B/16 (Tschannen et al., 2025), both trained on WebLI. The public checkpoints span twelve pretraining corpora, namely OpenAI's WIT, LAION-2B, LAION-400M, YFCC-15M, CC12M, DataComp-XL, DataComp-M, CommonPool-S, MetaCLIP-400M, DFN-2B, WebLI and the merged corpus of EVA02-CLIP. Ambient Recall@1 ranges from 2.2% to 61.2% on MSCOCO. Flickr30k uses the 27 public checkpoints, with ambient Recall@1 from 2.0% to 87.7%.

Jointly pretrained tri-modal checkpoints. On MSR-VTT we use the video, text, and audio towers of five jointly pretrained tri-modal checkpoints: the three GRAM variants (Cicchetti et al., 2025), TRIANGLE (Cicchetti et al., 2026) and VAST (Chen et al., 2023). GRAM's three variants are its five-modality and four-modality pretrained checkpoints and a version fine-tuned on MSR-VTT, all trained under its Gramian objective on a 150k subset of VAST-27M. TRIANGLE and VAST share the same tower architecture and were trained under different objectives, and VAST predates GRAM. Except for the fine-tuned variant, all are evaluated zero-shot on MSR-VTT without task-specific fine-tuning.

All five are evaluated on the same 884 MSR-VTT test videos in the same order, namely the subset carrying an audio track. The released VAST configuration otherwise returns all $1{,}000$ test videos and assigns an identical placeholder audio feature to the 116 without one, which would place 116 copies of a single vector into every audio-side result. Because the three retrieval directions differ by roughly a factor of four in ambient Recall@1, every aggregate over these five checkpoints is a mean over the three directions with equal weight per direction.

Additional MSR-VTT models. The characterization and the predictor both cover fourteen further models on the same 884 videos, each paired with its retrieval caption. Eight image–text models, OpenAI CLIP ViT-B/32, ViT-B/16 and ViT-L/14, OpenCLIP ViT-B/32 and ViT-B/16 on LAION-2B, ViT-L/14 on DataComp-XL, SigLIP ViT-B/16 and EVA02-CLIP B/16, embed eight uniformly spaced frames of each video, and the video embedding is the mean of the normalized frame embeddings. Two X-CLIP models (Ni et al., 2022), the base model with 32-pixel patches on eight frames and the zero-shot base model with 16-pixel patches on 32 frames, embed the video directly. These ten models retrieve video from text. Four CLAP models (Wu et al., 2023), the HTS-AT models with and without feature fusion and the two larger general and music-and-speech models, retrieve the audio track, resampled to 48 kHz, from text. With the five tri-modal checkpoints in three directions, the MSR-VTT evaluation comprises 29 units from 19 models, with ambient Recall@1 from 3.9% to 55.2%. Every unit is transformed by the twelve compressions.

Independently pretrained encoder pairs. The sixteen encoder pairs for which we first create cross-modal structure are grouped as follows. Four vary the vision training objective against a fixed MPNet text encoder: DINOv2-B (self-distillation), ViT-MAE (masked autoencoding), BEiT (masked image modelling) and DeiT (supervised classification). Two pair DINOv2-B with raw language encoders never fine-tuned toward sentence-level similarity, BERT-base and DistilRoBERTa-base. Four have incompatible native widths, namely DINOv2-B with MiniLM (768:384), DINOv2-L with MiniLM (1024:384), Swin with MPNet (1024:768) and ResNet-50 with MPNet (2048:768). Ambient similarity is undefined for this group because the two spaces share no inner product. The remaining six are on MSR-VTT using three encoders that never shared a training signal, DINOv2, MPNet and BEATs, evaluated both as the three pairwise combinations and as a single joint three-modality solve, again read in three directions. This tri-modal group uses independent encoders and is therefore distinct from the five jointly pretrained checkpoints used for the Recall@1 degradation experiments.

Transformations. The twelve generic compressions are principal component analysis and Gaussian random projection (Johnson et al., 1984) at output ranks 256, 128, 64, 32 and 16, together with per-tensor symmetric quantization to 8 and 4 bits. The four alignment maps are canonical correlation analysis and its multi-way generalization (Hotelling, 1992; Carroll, 1968; Kettenring, 1971), orthogonal Procrustes and least-squares regression (Choi et al., 2023), and ASIF (Norelli et al., 2023). The canonical correlation condition is fitted as a generalized canonical correlation solve whose per-modality projections are used either directly or after replacement by their nearest semi-orthonormal factor, at an output width of 128 or 256. These four options are selected per encoder pair on the selection split by validation Recall@1, which returns the semi-orthonormal variant for all 29 MSCOCO pairs, at width 256 for 23 pairs and 128 for six: the two trained from scratch, RN50 and RN101 on YFCC-15M, RN50 on CC12M and ViT-B/32 on DataComp-M. Projected embeddings are rescaled to unit length, so candidates in the shared space are compared by normalized correlation, the quantity canonical correlation maximizes (Hardoon et al., 2004; Gong et al., 2014), which is also the cosine similarity used for every other transformation. This solve is the only one of the four maps that extends to a joint solve over three modalities, and it is therefore the map applied in every setting we study and the map used by the figures that illustrate a single map. For ASIF we use the anchor budget and exponent reported by Norelli et al. (2023). ASIF retains its anchor set at inference, which places it in a different deployment cost class from the other maps, fitted on the same data.

Measures of cross-modal structure. Stack unit-normalised embeddings of N paired items as X and Y, write $\bar{x}$ and $\bar{y}$ for their column means, $\tilde{X}$ and $\tilde{Y}$ for column-centred matrices, and $S = XY^\top$ for the cross-modal similarity matrix, so that $S_{ij} = s(x_i, y_j)$ and the true partner of query i is y_i. Five common measures, tracked in Section 3, judge this pair of spaces:

$$\widehat{\mathrm{Sep}} = \frac{1}{N}\sum_{i=1}^{N}\Big(1 - \frac{r_i}{N-1}\Big), \qquad r_i = \big|\{\, j \neq i \;:\; S_{ij} > S_{ii} \,\}\big|, \tag{7a}$$

$$\mathrm{Gap} = \big\|\bar{x} - \bar{y}\big\|_2, \tag{7b}$$

$$\mathrm{Cos} = \frac{1}{N}\sum_{i=1}^{N} S_{ii}, \tag{7c}$$

$$\mathrm{CKA} = \frac{\|\tilde{X}^\top \tilde{Y}\|_F^2}{\|\tilde{X}^\top \tilde{X}\|_F \, \|\tilde{Y}^\top \tilde{Y}\|_F}, \tag{7d}$$

$$\mathrm{MSC} = \frac{1}{k}\sum_{j=1}^{k} c_j^2. \tag{7e}$$

MSC averages the squares of the k canonical correlations c_j, written c so that ρ stays the rank correlation used throughout. Each quantity is computed on the evaluation split from the two modality matrices. The modality gap and the mean true-pair cosine use the implementation released with Grassucci et al. (2026), and linear centred kernel alignment its feature-space form on column-centred matrices (Kornblith et al., 2019). The mean squared canonical correlation reduces each side to its top $k = 128$ principal components, with ridge 10^{-6} on each covariance, which keeps the solve

well conditioned at every output width we evaluate. Every transformation we study returns unit-normalised embeddings in both spaces, so each measure is on the same scale before and after it.

B.1 REPRODUCIBILITY

Public models and public data. Every encoder we evaluate is a publicly released checkpoint, obtained either from the Hugging Face Hub or from the authors' official repository, and is listed by name above. The only exceptions are the two ResNet-50 pairs we train from scratch, whose training configuration ships with the code. The three evaluation corpora, MSCOCO, Flickr30k and MSR-VTT, are public and are used in their standard test splits, Flickr30k in the split of Karpathy & Fei-Fei (2015). Appendix K gives the license of each.

Seeding. All randomness is drawn from explicitly seeded generators. The three-way split is a single permutation under seed 0 at fractions $0.40/0.20/0.40$, shared by every script through one common protocol routine, so all reported numbers of a dataset are measured on the same evaluation rows. The Gaussian random projection draws its matrix from a generator seeded at 0 on the host, making the matrix independent of the device the experiment runs on, and the same matrix is applied to both modalities. The shuffled control permutes the pairing under a fixed seed while preserving all marginals.

Deterministic estimation. No reported transformation or threshold is fitted by stochastic optimization. Principal component analysis, quantization, orthogonal Procrustes, least-squares regression, ASIF and canonical correlation analysis are each a deterministic, non-iterative function of the fitting split, ASIF taking that split itself as its anchor set, and the order-preservation threshold is a deterministic function of the two similarity matrices. Wherever a covariance is inverted, by the canonical correlation solves and by least-squares regression, it is regularized by a ridge of 10^{-3}. Similarity matrices, margins and thresholds are computed in double precision throughout, and feature extraction is a single forward pass with the model in evaluation mode, gradients disabled and the checkpoint's own released preprocessing, so a rerun reproduces the cached features.

Randomness that remains is reported as a distribution. Where a quantity depends on a sample, we report its spread. Correlations between predicted risk and measured loss carry bootstrap 95% confidence intervals over queries and over encoder pairs (Appendix G), and comparisons between predictors carry paired bootstrap intervals over encoder pairs or models (Appendix I). The calibration-efficiency curve of Figure 9 is a mean over 500 resampled reference sets at each size, with the range across draws shown as a band. The order-preservation test of Section 4 is reported at several margin cuts, each expressed in units of the encoder pair's own $\hat{\delta}^*$ (Appendix F), and the sensitivity of the predictor to the order-preservation criterion is reported over a grid of criteria in Appendix G.

Selection hygiene. Configuration choices are made on the selection split by validation Recall@1 and the reported number is measured once on the evaluation split, which selection never touches. Each reported number carries the configuration that produced it, so a reader can see whether a result depends on one corner of the grid.

Code. Anonymized code accompanies the submission as supplementary material. It covers feature extraction for every encoder, all sixteen transformations, the threshold estimator, the predictor and the baselines, and the scripts that emit each figure and table in this paper from the cached features. The experiments run under Python 3.10 with PyTorch $2.13.0$ built against CUDA 13.0, NumPy 2.2 and SciPy 1.15, on one Intel Xeon Gold 5218 host with eight threads and one Quadro RTX 8000 with 47 GiB. The GPU carries two steps, the forward passes that extract features from each checkpoint, one per checkpoint and corpus, and the similarity-matrix products. Everything else runs on the CPU in double precision from the cached features: the sixteen transformations, the threshold estimator, the margins and risk scores, the predictors and the baselines, the resampling, and the scripts that emit every figure and table.

C THE LOSS IN ABSOLUTE RECALL POINTS

The main text reports the loss relative to the target's own ambient Recall@K, which ranges from 2.0% to 87.7% across the pairs we study. This section repeats the phenomenon and the predictions in absolute Recall points, $R_{\mathrm{amb}} - R_T$, which has no denominator and therefore admits every configuration, including the two aligned spaces that Appendix D sets aside.

The phenomenon does not depend on the normalisation. Counted in absolute points over all $1{,}956$ configurations of both provenances, Recall@1 fails to improve in $1{,}815$, against $1{,}804$ of $1{,}932$ in relative terms, and the two counts agree configuration by configuration because the two quantities share a sign. The median drop is 20.6 points on the jointly pretrained pairs, 2.2 points on the compressions of aligned spaces, whose ambient accuracy is an order of magnitude lower, and 9.7 points pooled, with a largest drop of 86.4 points. The two set-aside aligned spaces lose a median of -0.07 points and at most 0.25 points, so in absolute terms their configurations sit inside the measurement noise, which is why a ratio taken against their 0.25 to 0.30% ambient Recall@1 is unstable rather than informative.

The predictions transfer to absolute points through the ambient accuracy. Table 4 repeats the 26 settings with both targets. Against the relative loss the risk score reaches a median ρ of 0.665 and the predicted loss 0.865. Against the absolute loss, fitted directly on it, they reach 0.551 and 0.707, because the absolute loss is the relative loss multiplied by the target's ambient Recall@1, a quantity that the margins do not carry and that a target cannot supply without labels. Supplying it recovers the absolute loss: multiplying the predicted relative loss by the target's ambient Recall@1 orders the absolute losses at a median ρ of 0.911, above the 0.707 of a predictor fitted on absolute losses directly, with a median error of 3.7 Recall points. The relative loss is therefore the scale-free quantity the margins predict, and the absolute loss follows from it by a factor the practitioner already knows.

Table 4: The same 26 settings with the loss measured relative to the target's ambient Recall@1 and in absolute Recall points. Risk: rank correlation of the risk score with that target. Predicted: rank correlation of Equation 6 with that target, fitted on it. Composite: the predicted relative loss multiplied by the target's ambient Recall@1, against the absolute loss. Median over the settings, with the interquartile range.

	Risk score	Predicted loss	Composite
Relative loss	0.665 (0.50–0.82)	0.865 (0.72–0.90)	–
Absolute loss	0.551 (0.24–0.62)	0.707 (0.58–0.78)	$\mathbf{0.911}$ (0.87–0.95)

D THE FIVE MEASURES, BY DATASET AND BY PROVENANCE

Table 5 reports every measure on each dataset, on the aligned spaces of the second provenance and pooled, and Table 6 the four alignment maps, which improve the measures they optimize while removing Recall@1, every one of them closing the modality gap and all but one raising the true-pair cosine. The *After alignment* block holds the twelve compressions applied to the 50 aligned spaces built from the 13 two-view units of the sixteen independently pretrained pairs, the three-modality joint solve excluded (Appendix G), and the pooled block adds it to the jointly pretrained grid. Two further aligned spaces, the DINOv2-B and DistilRoBERTa-base pair under orthogonal Procrustes and under ASIF, align to a Recall@1 of 0.25 and 0.30%, so a single query changing rank moves their relative loss by tens of per cent and their 24 configurations range from a 400% loss to an 83% gain. They are excluded from the phenomenon statistics quoted in the main paper and reported only here. The prediction grid of Appendix G keeps them, since a near-zero denominator disturbs a rank correlation far less than it disturbs a median. Figure 4 is the complete form of Figure 1, where every measure straddles zero while Recall@1 lies to one side, and Figure 6 plots the loss against the change in separability with each compression family marked and each alignment map labelled at its mean, where the four maps sit at a large loss and a separability change near zero. Figure 5 plots each measure's change against the loss, where the compression families separate: random projection lowers every measure, and principal component analysis raises the modality gap and the canonical

correlation while removing most of Recall@1. Separability is a probability with chance level 0.5, so its change can also be read against its distance from chance, $|\Delta\text{sep}|/(\text{sep} - 0.5)$. Read this way, the relative Recall@1 loss still exceeds the change in 99.8%, 99.5% and 97.3% of the losing configurations on MSCOCO, Flickr30k and MSR-VTT, in 94.2% after alignment and in 97.6% of all 1,760, by median factors of 24.9, 13.8, 3.8, 5.2 and 8.5.

Neither of those two measures registers a re-orientation. Over the 85 orthogonal Procrustes configurations, 29 on MSCOCO, 27 on Flickr30k and 29 on MSR-VTT, the relative change of centred kernel alignment is at most $6.9 \times 10^{-7}\%$ and that of the mean squared canonical correlation at most $5.4 \times 10^{-7}\%$. The other three maps are not orthogonal and move both quantities freely, by a median of 11.8% and 13.2% under the canonical correlation solve and 12.1% and 4.4% under ASIF. Recall@1 meanwhile falls by a mean of 39.9% over those same Procrustes configurations, so the transformation that these two measures cannot see at all is one that removes two fifths of the retrieval they are consulted about. Appendix F proves both invariances.

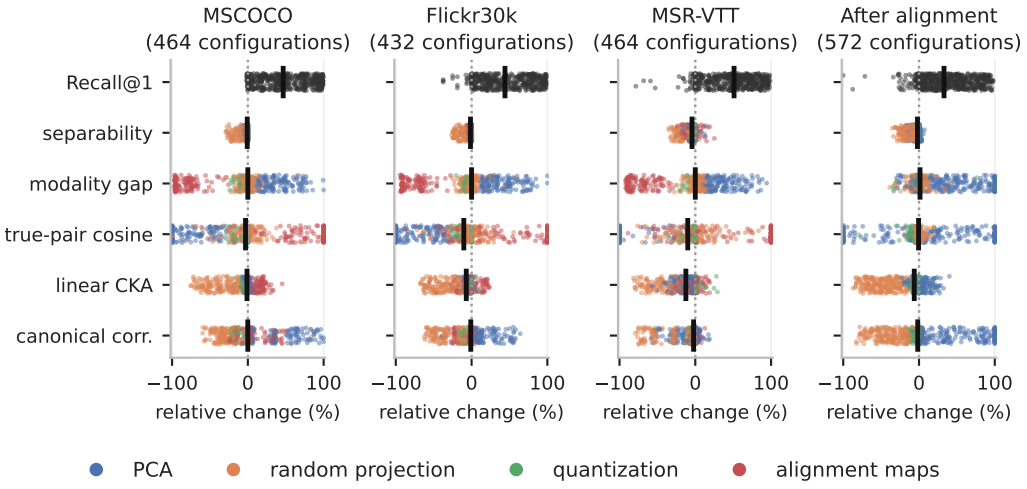


Figure 4: Relative change of Recall@1 and of the five measures, one point per configuration, for each dataset and for the compressions of the post-hoc aligned independently pretrained pairs. Bars mark medians and values are clipped to ±100%. Recall@1 is plotted as a loss, so its points lie to the right.

E Complete transformation results

Table 7 reports the four post-hoc alignment maps at four retrieval depths on MSCOCO and MSR-VTT, expanding the result summarised in Section 3. The loss concentrates at the top of the ranking: the relative loss falls with depth for every map on both datasets, and on MSR-VTT least-squares regression raises Recall@50 by 10.1% while removing 30.5% of Recall@1, improving the coarse ranking and destroying the decisive comparison at once. Table 8 reports the full compression sweep for one pair, and Table 9 the complete MSR-VTT grid: all sixteen transformations on each of the five jointly pretrained tri-modal checkpoints, in all three retrieval directions.

Dependence on the encoder pair. Over the twelve compressions on MSCOCO, a pair's mean Recall@1 loss correlates with its ambient Recall@1 at $\rho = 0.31$, its ambient separability at 0.33 and its embedding width at -0.10, none significantly ($p \geq 0.08$ over the 29 pairs), while the risk score of Section 5 orders the same pairs by mean loss at $\rho = 0.755$ ($p < 10^{-5}$). The ViT-B/32 and ViT-B/16 checkpoints trained on LAION-2B lose 52.3% and 50.8% of Recall@1 on average, against 67.6% and 64.3% for their OpenAI counterparts of identical architecture and width.

Figures 7 and 8 show the relative Recall@1 loss of each encoder pair under each of the sixteen transformations. On MSCOCO the same transformation can leave one pair's Recall@1 intact and

Table 5: Each quantity's relative change under the sixteen transformations. Improves: the share of configurations in which the quantity improves, rising or, for the modality gap, shrinking. Improves as loss: the share in which it improves while Recall@1 falls. ρ: Spearman correlation between the quantity's relative change and the relative Recall@1 change, so a measure that falls with retrieval correlates positively. Loss exceeds: among the configurations that lose Recall@1, the share in which the relative loss exceeds the quantity's own relative change, with the median factor by which it does. The Recall@1 row reports the share of configurations that lose it and the median loss.

Dataset	Quantity	Median	Improves	Improves as loss	ρ	Loss exceeds	Factor
MSCOCO (464)	Recall@1 loss	$+48.7\%$	0.95	–	–	–	–
	separability	-0.8%	0.03	0.02	0.90	1.00	50
	modality gap	-0.0%	0.52	0.48	-0.13	0.65	1.4
	true-pair cosine	-3.1%	0.37	0.36	0.09	0.52	1.1
	linear CKA	-1.1%	0.30	0.29	0.39	0.97	4.1
	canonical corr.	-0.1%	0.43	0.42	0.22	0.91	2.5
Flickr30k (432)	Recall@1 loss	$+49.3\%$	0.94	–	–	–	–
	separability	-1.5%	0.05	0.01	0.94	1.00	28
	modality gap	-0.0%	0.54	0.50	0.06	0.58	1.1
	true-pair cosine	-10.2%	0.32	0.31	-0.06	0.54	1.4
	linear CKA	-7.0%	0.19	0.17	0.66	0.97	3.0
	canonical corr.	-1.1%	0.32	0.32	0.41	0.97	2.8
MSR-VTT (464)	Recall@1 loss	$+55.8\%$	0.89	–	–	–	–
	separability	-4.3%	0.12	0.06	0.83	1.00	8.8
	modality gap	$+0.0\%$	0.50	0.42	-0.03	0.68	1.7
	true-pair cosine	-9.9%	0.34	0.32	-0.00	0.48	0.9
	linear CKA	-12.4%	0.22	0.17	0.65	0.96	3.0
	canonical corr.	-2.2%	0.25	0.22	0.42	0.97	8.2
Jointly pretrained (1,360)	Recall@1 loss	$+52.1\%$	0.93	–	–	–	–
	separability	-1.9%	0.07	0.03	0.84	1.00	21
	modality gap	-0.0%	0.52	0.47	-0.04	0.64	1.4
	true-pair cosine	-7.5%	0.34	0.33	0.01	0.51	1.1
	linear CKA	-5.9%	0.24	0.21	0.55	0.97	3.2
	canonical corr.	-1.2%	0.34	0.32	0.35	0.95	3.0
After alignment (572)	Recall@1 loss	$+41.1\%$	0.88	–	–	–	–
	separability	-2.0%	0.17	0.11	0.75	0.97	11.0
	modality gap	$+1.3\%$	0.40	0.35	-0.04	0.68	3.8
	true-pair cosine	-0.6%	0.37	0.33	0.10	0.65	5.9
	linear CKA	-6.3%	0.24	0.21	0.45	0.87	2.2
	canonical corr.	-1.7%	0.38	0.33	0.25	0.69	1.5
All (1,932)	Recall@1 loss	$+48.9\%$	0.91	–	–	–	–
	separability	-1.9%	0.10	0.06	0.80	0.99	18
	modality gap	$+0.1\%$	0.49	0.43	-0.01	0.65	1.6
	true-pair cosine	-2.7%	0.35	0.33	0.06	0.55	1.4
	linear CKA	-6.0%	0.24	0.21	0.51	0.94	3.0
	canonical corr.	-1.3%	0.35	0.32	0.32	0.88	2.5

remove most of another's, as for PCA to 64 dimensions, where the change ranges from a 1.0% gain to an 83.4% loss. The pairs also keep a similar order across transformations, with a mean rank correlation of 0.58 between the columns, excluding 8-bit quantization, which leaves every pair essentially unchanged. On MSR-VTT each marker is one checkpoint in one retrieval direction, so each column mixes modality pairs whose ambient Recall@1 differs by a factor of four, and post-hoc alignment raises Recall@1 on the weakest audio directions.

F Proofs, assumptions and empirical tests

Proof of Proposition 1. Fix a query x and two candidates y_i, y_j with positive ambient margin $m_{ij} = s(x, y_i) - s(x, y_j)$, as in Equation 1. Write the post-transform margin as $m_{ij} + \Delta m_{ij}$, where Δm_{ij} is

Table 6: The four post-hoc alignment maps: mean relative Recall@1 loss and mean relative change of the modality gap, the mean true-pair cosine and separability, per dataset. On MSR-VTT the mean is over all 29 units, where Table 7 averages the five tri-modal checkpoints alone.

Dataset	Map	Recall@1 loss	Modality gap	True-pair cosine	Separability
MSCOCO	Orthogonal Procrustes	$+37.7\%$	-84.6%	$+128.0\%$	-0.78%
	Least-squares regression	$+18.4\%$	-70.1%	$+169.3\%$	-0.11%
	ASIF	$+72.7\%$	-63.3%	$+59.0\%$	-1.12%
	Canonical correlation analysis	$+47.2\%$	-91.2%	$+56.6\%$	-0.75%
Flickr30k	Orthogonal Procrustes	$+32.9\%$	-86.9%	$+100.5\%$	-1.10%
	Least-squares regression	$+21.0\%$	-66.1%	$+138.1\%$	-0.62%
	ASIF	$+78.1\%$	-59.3%	$+30.0\%$	-4.26%
	Canonical correlation analysis	$+63.2\%$	-91.5%	-22.2%	-4.65%
MSR-VTT	Orthogonal Procrustes	$+48.5\%$	-80.4%	$+314.7\%$	-4.19%
	Least-squares regression	$+40.6\%$	-62.1%	$+408.8\%$	$+1.39\%$
	ASIF	$+60.8\%$	-53.6%	$+147.4\%$	-5.63%
	Canonical correlation analysis	$+49.5\%$	-86.3%	$+30.1\%$	-4.93%

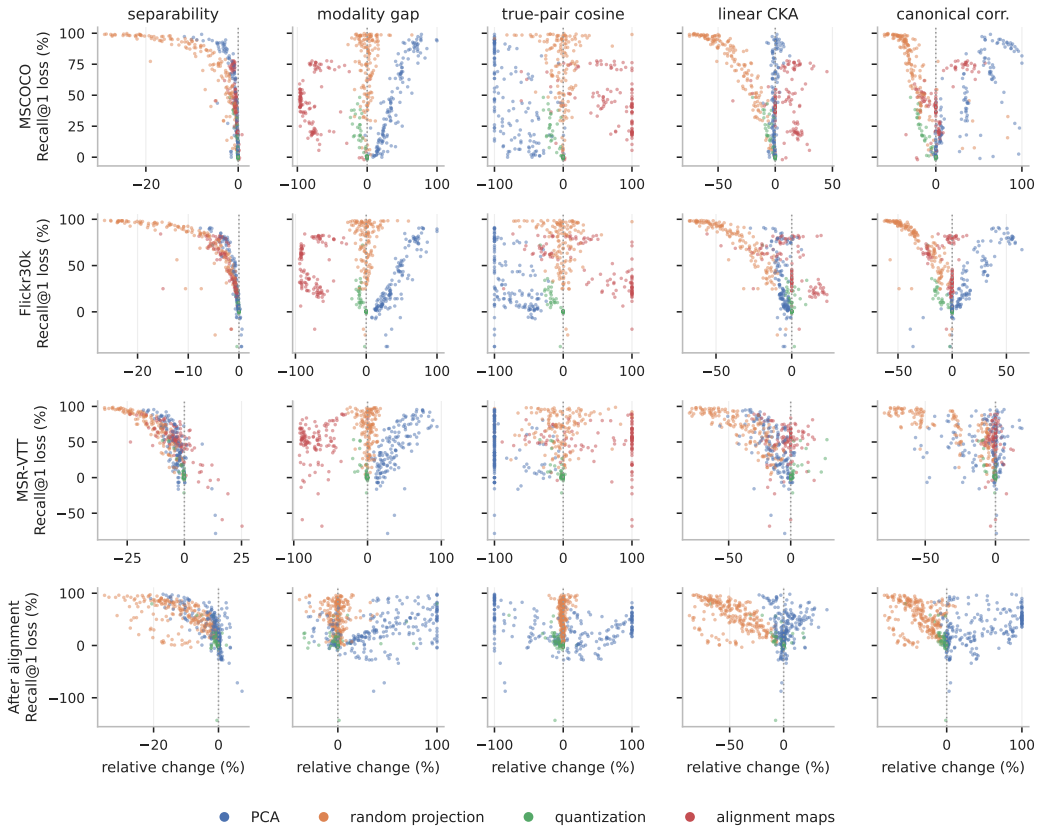


Figure 5: Each measure's relative change against the relative Recall@1 loss, one column per measure and one row per dataset, with a fourth row for the compressions of the post-hoc aligned independently pretrained pairs. Values are clipped to $\pm 100\%$.

the perturbation the transformation induces on the difference of the two similarities. By assumption Δm_{ij} is symmetric about zero with scale σ_T. Let F denote the cumulative distribution function of $\Delta m_{ij}/\sigma_T$, which is continuous and strictly increasing by assumption, and satisfies $F(0) = 0.5$ by

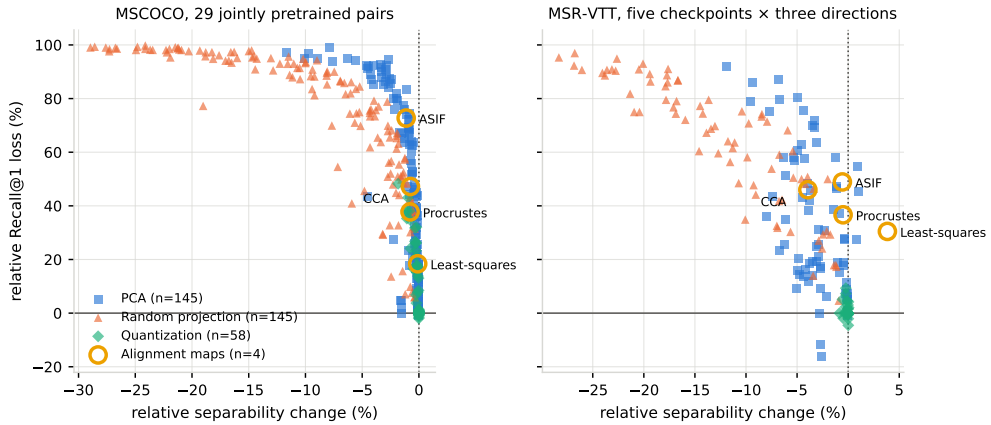


Figure 6: Relative Recall@1 loss against relative change in separability, the measure of the five that orders it best. **Left**: the 29 jointly pretrained MSCOCO pairs. **Right**: the five jointly pretrained MSR-VTT checkpoints in all three retrieval directions. Marker shape denotes the compression family and open circles the four alignment maps, labelled at their mean. Loss is positive upwards, so points below the horizontal line gained Recall@1, and points right of the vertical dotted line improved separability.

Table 7: Post-hoc alignment at depth on jointly pretrained encoder pairs. Relative Recall@K loss in percent, positive for a loss, together with the absolute change in match-versus-mismatch separability. MSCOCO is the mean over the 29 image–text pairs and MSR-VTT the mean over five tri-modal checkpoints in all three retrieval directions. Negative entries mark the cases in which recall improved.

	Relative Recall@K loss (%)				
Map	$K=1$	$K=5$	$K=10$	$K=50$	Δ Separability
MSCOCO, 29 jointly pretrained image–text pairs					
Orthogonal Procrustes	37.7	23.9	17.8	5.9	−0.0077
Least-squares regression	18.4	9.0	5.1	0.3	−0.0011
ASIF	72.7	50.9	38.1	10.2	−0.0111
Canonical correlation analysis	47.2	29.8	21.7	5.4	−0.0072
MSR-VTT, five jointly pretrained tri-modal checkpoints × three directions					
Orthogonal Procrustes	36.7	18.6	11.2	−0.6	−0.0065
Least-squares regression	30.5	2.7	−2.4	−10.1	+0.0287
ASIF	48.9	34.1	23.7	0.7	−0.0091
Canonical correlation analysis	46.0	32.7	24.7	10.4	−0.0364

symmetry. The ambient order is preserved if and only if $m_{ij} + \Delta m_{ij} > 0$, so

$$\Pr[\text{order preserved}] \;=\; \Pr\left[\frac{\Delta m_{ij}}{\sigma_T} > -\frac{m_{ij}}{\sigma_T}\right] \;=\; 1 - F\left(-\frac{m_{ij}}{\sigma_T}\right) \;=\; F\left(\frac{m_{ij}}{\sigma_T}\right),$$

the second equality by continuity of F and the third by its symmetry about zero, which gives $1 - F(-u) = F(u)$. This is monotone in m_{ij}/σ_T, tends to $F(0) = 0.5$ as $m_{ij} \to 0$, and tends to 1 as $m_{ij} \to \infty$, which is the first part of the claim. For the second, Definition 1 sets δ^* as the margin at which the preservation probability first exceeds a fixed criterion, which Section 4 fixes at $0.5 + \kappa$ with $\kappa = 0.05$. Solving $F(\delta^*/\sigma_T) = 0.5 + \kappa$ gives

$$\delta^* \;=\; \sigma_T \, F^{-1}(0.5 + \kappa)\,,$$

so δ^* is proportional to the perturbation scale σ_T with a constant of proportionality depending only on the shape of F and on the fixed κ. Within the model, σ_T and F are properties of T alone, so δ^* carries no dependence on the encoder pair. Its estimate $\hat{\delta}^*$ is read from the preservation curve one

Table 8: Generic compression on CLIP ViT-B/16 (LAION-2B), MSCOCO, whose embeddings have width $d = 512$. Values are relative changes from the ambient embedding.

Transformation	Δ Recall@1	Δ Separability
8-bit quantization	-0.1%	0.00%
4-bit quantization	-23.3%	-0.25%
PCA-256	-4.5%	-0.03%
PCA-128	-13.2%	-0.05%
PCA-64	-36.5%	-0.15%
PCA-32	-67.2%	-0.80%
PCA-16	-88.9%	-2.88%
JL-256	-38.2%	-0.55%
JL-128	-67.2%	-1.92%
JL-64	-79.2%	-3.96%
JL-32	-93.2%	-10.01%
JL-16	-98.5%	-21.99%

Table 9: Complete MSR-VTT transformation results. Each cell reports ΔRecall@1 / ΔSeparability, relative to the ambient embedding and averaged over the three retrieval directions.

Transformation	GRAM-5	GRAM-4	GRAM-ft	TRIANGLE	VAST
PCA-256	-12.4 / -2.49	-10.1 / -3.41	-2.6 / -1.93	-22.8 / -2.80	-12.9 / -3.49
PCA-128	-15.6 / -2.33	-14.1 / -3.28	-2.0 / -1.84	-26.0 / -2.23	-16.8 / -3.12
PCA-64	-20.3 / -2.04	-16.3 / -2.98	-17.3 / -1.92	-41.5 / -1.84	-31.6 / -3.50
PCA-32	-40.0 / -2.41	-42.7 / -3.17	-32.3 / -2.68	-64.0 / -2.31	-60.7 / -4.89
PCA-16	-63.5 / -4.13	-61.4 / -4.63	-69.0 / -5.33	-77.1 / -4.97	-85.6 / -10.30
JL-256	-30.9 / -4.95	-30.3 / -4.95	-26.3 / -3.26	-40.6 / -3.57	-38.5 / -6.13
JL-128	-45.4 / -7.48	-44.5 / -7.52	-44.3 / -8.65	-52.1 / -6.75	-62.9 / -10.24
JL-64	-60.6 / -11.07	-65.9 / -10.50	-58.9 / -10.34	-60.6 / -9.27	-68.7 / -12.11
JL-32	-75.3 / -15.32	-78.2 / -15.88	-76.4 / -16.05	-77.6 / -14.04	-90.2 / -19.08
JL-16	-90.7 / -21.33	-88.8 / -21.20	-91.4 / -23.93	-89.1 / -21.31	-94.2 / -25.15
8-bit quant.	0.2 / -0.00	0.0 / -0.01	0.7 / -0.01	1.5 / 0.01	-0.8 / -0.02
4-bit quant.	0.1 / -0.46	-2.2 / -0.24	-2.9 / -0.06	-3.7 / -0.21	-3.1 / -0.25
Orth. Procrustes	-41.8 / 0.29	-46.5 / -0.04	-34.3 / -2.91	-14.1 / 4.28	-46.6 / -4.10
Least squares	-39.1 / 2.97	-37.0 / 2.51	-22.0 / 1.72	-13.5 / 8.76	-40.8 / 3.24
ASIF	-51.4 / -1.10	-53.4 / -2.16	-53.4 / -1.86	-34.0 / 3.03	-52.3 / -0.76
CCA	-54.0 / -5.44	-49.9 / -6.59	-50.8 / -4.45	-34.9 / -0.92	-40.5 / -2.33

encoder pair produces, so it varies across pairs, and Appendix G takes the median over the reference pairs as the value used throughout. □

Proof of Corollaries 1 and 2. Fix a query answered correctly before the transformation, with decisive margin $m > 0$, and let F be the distribution function of $\Delta m_{ij}/\sigma_T$ as above. The query is answered incorrectly afterwards exactly when $m + \Delta m_{ij} \le 0$, an event of probability $1 - F(m/\sigma_T)$ by the computation in the proof of Proposition 1. Averaging over those queries gives

$$\ell(\sigma_T, M) \;=\; \int_{(0,\infty)} \Big(1 - F\big(m/\sigma_T\big)\Big)\, dM(m).$$

The pairwise statement of Corollary 1 is the case of a single m: for $\sigma'_T \ge \sigma_T$ and every $m > 0$ we have $m/\sigma'_T \le m/\sigma_T$, and F is increasing, so $1 - F(m/\sigma'_T) \ge 1 - F(m/\sigma_T)$ and the reversal probability is non-decreasing in σ_T. Its aggregate statement follows by integrating that pointwise inequality against a fixed M. For Corollary 2, fix σ_T and let M_2 first-order stochastically dominate M_1, so $M_2(t) \le M_1(t)$ for every t and M_2 places its decisive margins systematically higher. The map $m \mapsto 1 - F(m/\sigma_T)$ is non-increasing, and the expectation of a non-increasing function is no smaller under a stochastically dominated distribution, which gives $\ell(\sigma_T, M_1) \ge \ell(\sigma_T, M_2)$. □

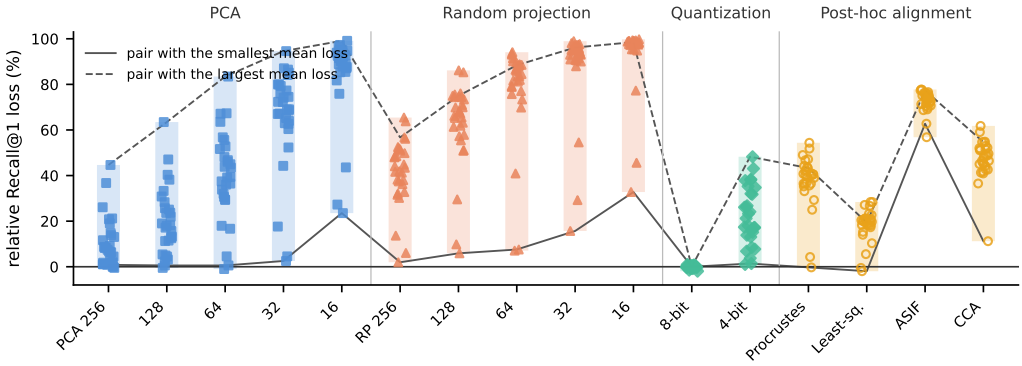


Figure 7: Relative Recall@1 loss of each of the 29 jointly pretrained MSCOCO encoder pairs under each transformation, one marker per pair. Bars span the range across pairs. The two lines trace the single pair with the smallest and the single pair with the largest mean loss over all sixteen columns, so they follow two individual pairs and not the bounds of each column. Loss is positive upwards.

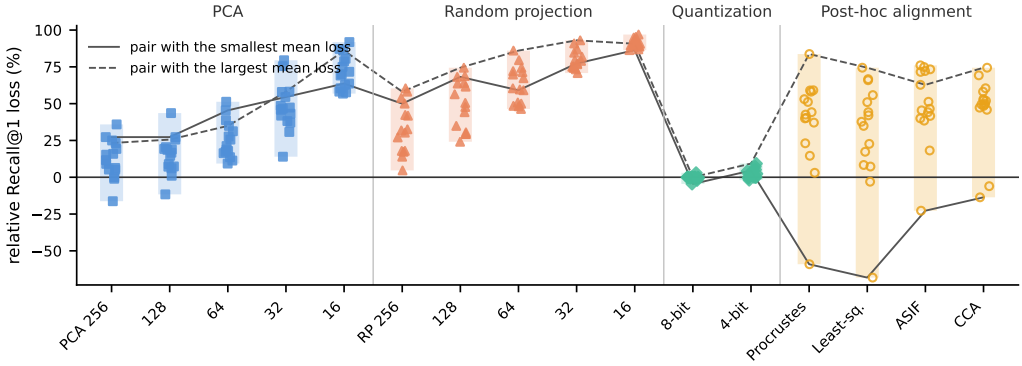


Figure 8: As Figure 7, for the five jointly pretrained MSR-VTT checkpoints in each of the three retrieval directions, one marker per checkpoint and direction.

Exact invariances of two of the measures. Both statements below are exact invariances under a group action. Let Q be orthogonal, $Q^\top Q = I$, applied to one side of an encoder pair, $Y \mapsto YQ$, an action of the orthogonal group $O(d)$. Centring commutes with it, since the column means of YQ are $Q^\top \bar{y}$ and the centred matrix is therefore $\tilde{Y}Q$. For linear centred kernel alignment (Equation 7d) the numerator obeys

$$\|\tilde{X}^\top \tilde{Y} Q\|_F^2 = \mathrm{tr}\big(Q^\top \tilde{Y}^\top \tilde{X} \tilde{X}^\top \tilde{Y} Q\big) = \mathrm{tr}\big(\tilde{Y}^\top \tilde{X} \tilde{X}^\top \tilde{Y}\big) = \|\tilde{X}^\top \tilde{Y}\|_F^2,$$

by the cyclic property of the trace and $QQ^\top = I$, and the denominator obeys $\|Q^\top \tilde{Y}^\top \tilde{Y} Q\|_F = \|\tilde{Y}^\top \tilde{Y}\|_F$, the Frobenius norm being invariant under orthogonal conjugation. The statistic is therefore identical before and after, and the same argument applies to a map of the other side. The canonical correlations are invariant under the wider action of the general linear group, any invertible linear map M of either side, since they are the stationary values of $\mathrm{corr}(Xa, Yb)$ over a and b, and $Y \mapsto YM$ reparametrises $b \mapsto M^{-1}b$, a bijection of the search space that leaves the attainable correlations unchanged, and with them Equation 7e. Our estimator first reduces each side to its top 128 principal components, which preserves the orthogonal case exactly, the principal subspace rotating with the data, and leaves the statement for a general M one about the population quantity.

On the assumption. Symmetry of Δm_{ij} about zero is a modelling hypothesis. It fixes the value of the curve at the origin: preservation must approach 0.5 as the margin vanishes. Pooled over

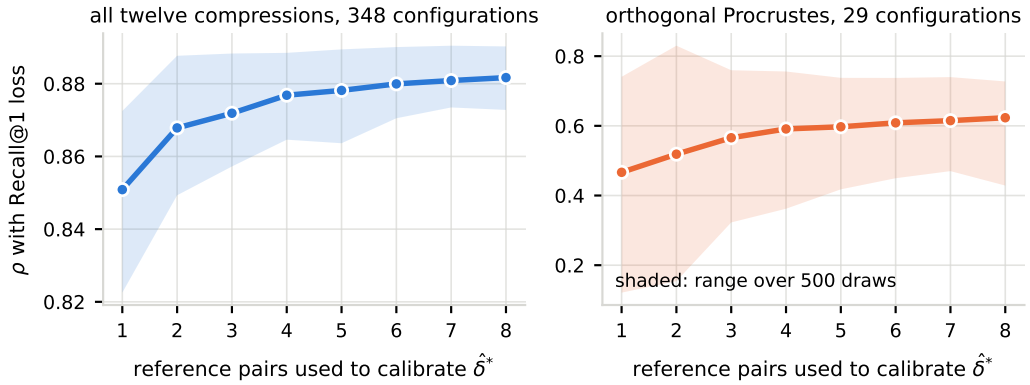


Figure 9: Calibration efficiency on MSCOCO: rank correlation between risk and measured Recall@1 loss against the number of reference pairs the threshold is calibrated on, a mean over 500 draws from the other 28 pairs, with the range across draws shaded. Compression saturates at one reference and orthogonal Procrustes is still rising at eight. A reference costs one public checkpoint and one run of the transformation, with no relevance label on either side (Section 5).

the 29 jointly pretrained encoder pairs of Section 4, with each candidate pair's ambient margin expressed in units of its encoder pair's own $\hat{\delta}^*$, measured preservation is 0.5050 below $\hat{\delta}^*/100$ on 36,376 sampled candidate pairs (query-clustered 95% interval $[0.5002, 0.5100]$), 0.5037 below $\hat{\delta}^*/50$ (72,276), 0.5037 below $\hat{\delta}^*/20$ (181,136) and 0.5048 below $\hat{\delta}^*/10$ (364,882). Preservation therefore stays within 0.005 of 0.5 over the smallest margins, against the rise to 0.55 at $\hat{\delta}^*$ that defines the threshold and the approach to 1 at wide margins (Figure 2). The same holds on the pairs that decide Recall@1. Over the twelve compressions of the 29 pairs, the order of each query's two highest ambient candidates survives with probability 0.513 when their gap lies below a quarter of the configuration's $\hat{\delta}^*$ (94,985 queries), and with probability 0.977 beyond twelve times it. The gap at which it reaches the 0.55 criterion is a median 0.47 of the random-pair $\hat{\delta}^*$ (interquartile range 0.28–0.68 over 348 configurations), so the top of the ranking is perturbed with the same symmetry at about half the scale that random candidate pairs register. The assumption concerns the marginal law of a single perturbation. The magnitude of the loss depends on the joint law of $\{\Delta m_{ij}\}_j$ across candidates, and Section 5 therefore treats it as a calibrated quantity.

G Predictor calibration

Estimating one threshold from several reference pairs. For a given transformation, $\hat{\delta}^*$ is estimated independently on each reference pair from that pair's own ambient and transformed similarities, and the reference value is the median of those estimates, taken separately for each transformation. There is therefore one threshold per transformation, and with a single reference pair the median reduces to that pair's own estimate. The figures quoted in Section 5 take the median over the 28 pairs other than the target. On Flickr30k the median runs over the 26 other pairs, and on MSR-VTT over the units of the 18 models other than the target's, so that no direction of the target's model enters its calibration. The calibration-size curve of Figure 9 resamples subsets of the remaining 28 pairs and reports the mean correlation over 500 draws at each size.

How many reference pairs calibration needs. Table 10 draws k reference groups at random from those other than the target's, takes their median threshold per transformation and re-evaluates the risk score, over 200 draws at each size. The threshold and the margin do not depend on the retrieval depth, so only the measured loss changes between the three blocks, and the $k =$ all column reproduces every leave-one-group-out correlation of Table 11 exactly. One reference pair carries most of the signal where the targets are image–text pairs under compression, retaining 96% of the full-calibration correlation on MSCOCO and on Flickr30k, and much less elsewhere, 80% on MSR-VTT compression and between 62% and 88% on the alignment maps. Five pairs come within 5%

of full calibration in all six settings at Recall@1, which is the figure quoted in Section 5, and within 11% at the deeper ranks, where the setting that needs the most references, MSR-VTT alignment, is also the one whose correlation rises most with depth, from 0.301 to 0.444. One reference suffices in the mean rather than reliably. Figure 9 shows the spread the means may hide. Under compression, one reference gives 0.851 in the mean and 0.822 on the unluckiest of 500 draws, a range of 0.050 that closes to 0.018 by eight references. Under orthogonal Procrustes the same draw runs from 0.121 to 0.741, and eight references still leave a standard deviation of 0.044. Table 10 separates the settings the same way, at a standard deviation across draws of 0.010 on MSCOCO compression against 0.111 on MSR-VTT alignment, where it exceeds half the mean. Five references cut those spreads by half or better under compression and by more than a third on the alignment maps, which is the reason to prefer five where the references are heterogeneous. A reference is another publicly released checkpoint run through the same transformation, with no relevance labels and no access to the target, so calibrating on every available reference costs only those transformation runs, and that is what the rest of the paper does.

Table 10: Mean rank correlation between predicted risk and the measured relative Recall@K loss when the threshold is calibrated on k reference groups drawn at random from those other than the target's, a mean over 200 draws. Groups are encoder pairs, and models on MSR-VTT. The last column calibrates on every other group, as the rest of the paper does. In brackets, the standard deviation across draws at the two sizes where it decides how many references to use. The score is read at the top-1 margin throughout, as in Table 11.

Family	Dataset	$k=1$	$k=2$	$k=3$	$k=5$	$k=8$	all
Recall@1							
Compression	MSCOCO	0.851 (0.010)	0.868	0.872	0.878 (0.005)	0.882	0.887
	Flickr30k	0.713 (0.020)	0.722	0.731	0.735 (0.007)	0.738	0.741
	MSR-VTT	0.646 (0.048)	0.686	0.749	0.779 (0.021)	0.795	0.804
Alignment	MSCOCO	0.413 (0.040)	0.469	0.462	0.482 (0.025)	0.496	0.500
	Flickr30k	0.396 (0.029)	0.406	0.431	0.438 (0.016)	0.443	0.451
	MSR-VTT	0.186 (0.111)	0.204	0.264	0.287 (0.067)	0.298	0.301
Recall@5							
Compression	MSCOCO	0.877 (0.010)	0.895	0.901	0.907 (0.004)	0.910	0.915
	Flickr30k	0.764 (0.018)	0.773	0.786	0.791 (0.007)	0.793	0.796
	MSR-VTT	0.669 (0.051)	0.702	0.779	0.810 (0.019)	0.827	0.838
Alignment	MSCOCO	0.410 (0.044)	0.463	0.463	0.481 (0.023)	0.493	0.498
	Flickr30k	0.521 (0.028)	0.537	0.562	0.570 (0.014)	0.578	0.586
	MSR-VTT	0.260 (0.107)	0.294	0.325	0.378 (0.062)	0.409	0.420
Recall@10							
Compression	MSCOCO	0.878 (0.010)	0.897	0.902	0.908 (0.004)	0.912	0.917
	Flickr30k	0.795 (0.020)	0.803	0.817	0.821 (0.007)	0.825	0.827
	MSR-VTT	0.668 (0.052)	0.702	0.775	0.810 (0.020)	0.827	0.841
Alignment	MSCOCO	0.406 (0.040)	0.460	0.463	0.481 (0.026)	0.493	0.499
	Flickr30k	0.563 (0.029)	0.581	0.601	0.612 (0.014)	0.617	0.628
	MSR-VTT	0.289 (0.107)	0.314	0.340	0.398 (0.068)	0.427	0.444

Confidence intervals. Table 11 gives the rank correlation of the risk score with the loss, with a cluster bootstrap over encoder pairs, for every target and calibration of the within-dataset and transfer analyses, at the three retrieval depths, over 2,000 resamples. The configurations of one encoder pair share that pair's margin distribution, so the pair, and on MSR-VTT the model, is the resampling unit a ranking claim needs. Every interval lies above zero at every depth, and the weakest setting, the MSR-VTT alignment maps at Recall@1, keeps a lower bound of 0.147. Resampling evaluation queries instead of pairs gives $[0.865, 0.895]$, $[0.901, 0.919]$ and $[0.908, 0.927]$ on the 348 MSCOCO compression configurations, narrower than the pair interval because a query enters one configuration where a pair enters twelve. At every depth the correlations lie between 0.741 and 0.940 in the compression settings and between 0.301 and 0.628 on the alignment maps, all significant at $p < 0.005$. Appendix J instead reads the margin at the K-th boundary, which is the form the account prescribes for Recall@K.

Table 11: Rank correlation between predicted risk and measured relative Recall@K loss with a 95% cluster bootstrap over encoder pairs, and over models on MSR-VTT, at $2{,}000$ resamples. References: the other groups of the target's own dataset, or the 29 jointly pretrained MSCOCO pairs, or for compression after alignment the other aligned units of the same map. Resampling units: 29 MSCOCO pairs, 27 Flickr30k pairs, 19 MSR-VTT models and 10 aligned units. n: configurations. The risk score is read at the top-1 margin at every depth.

Target	References	n	Recall@1	Recall@5	Recall@10
Compression, MSCOCO	28 other pairs	348	0.887 $[0.83, 0.92]$	0.915 $[0.88, 0.94]$	0.917 $[0.87, 0.94]$
Alignment, MSCOCO	28 other pairs	116	0.500 $[0.41, 0.60]$	0.498 $[0.41, 0.60]$	0.499 $[0.41, 0.58]$
Compression, Flickr30k	26 other pairs	324	0.741 $[0.55, 0.87]$	0.796 $[0.67, 0.88]$	0.827 $[0.74, 0.90]$
	29 MSCOCO pairs	324	0.772 $[0.59, 0.89]$	0.828 $[0.71, 0.91]$	0.857 $[0.77, 0.92]$
Alignment, Flickr30k	26 other pairs	108	0.451 $[0.29, 0.59]$	0.586 $[0.51, 0.66]$	0.628 $[0.56, 0.70]$
	29 MSCOCO pairs	108	0.392 $[0.20, 0.56]$	0.537 $[0.45, 0.63]$	0.572 $[0.49, 0.66]$
Compression, MSR-VTT	18 other models	348	0.804 $[0.74, 0.87]$	0.838 $[0.77, 0.89]$	0.841 $[0.76, 0.89]$
	29 MSCOCO pairs	348	0.849 $[0.81, 0.89]$	0.888 $[0.85, 0.92]$	0.899 $[0.85, 0.93]$
Alignment, MSR-VTT	18 other models	116	0.301 $[0.14, 0.43]$	0.420 $[0.25, 0.55]$	0.444 $[0.27, 0.56]$
	29 MSCOCO pairs	116	0.367 $[0.19, 0.52]$	0.470 $[0.29, 0.62]$	0.498 $[0.33, 0.64]$
Compr. after align.	9 other aligned	98	0.927 $[0.91, 0.94]$	0.939 $[0.92, 0.95]$	0.940 $[0.92, 0.95]$
	29 MSCOCO pairs	98	0.902 $[0.87, 0.93]$	0.894 $[0.87, 0.92]$	0.893 $[0.87, 0.92]$

Reading the risk without the target's labels. The risk score counts, over all queries, the top-1 decision margins g that fall below $\hat{\delta}^*$ (Equation 5). On every correctly answered query g equals the decisive margin $m^\star$, so restricting the count to those queries gives the labelled variant $\Pr(m^\star < \hat{\delta}^* \mid m^\star > 0)$, which requires the true partner of each query. The two scores rank the configurations almost identically, their Spearman correlation with each other being 0.971 over the 348 compression cells and 0.924 over the 116 alignment cells, with each threshold taken from the other 28 encoder pairs. The two differ in level, at a mean risk of 0.381 against 0.225 under compression, since the count over every query includes those already answered incorrectly, which is why Equation 6 corrects by the target's percentile among the reference risks rather than by the risk itself. The level can also be recovered without the target's labels. Learning on the reference units, whose labels the severity term already uses, how often a query with gap g is answered correctly, as a function of g over the spread of the unit's ambient similarities, and weighting each target query by it, brings the count within a mean absolute 0.018 of the labelled risk on MSCOCO compression, against 0.156 for the unweighted count, and within 0.017 to 0.056 in the other five settings. Counting every query also ranks the losses more closely than the labelled count, at $\rho = 0.887$ against 0.825 under compression on MSCOCO, 0.741 against 0.712 on Flickr30k and 0.804 against 0.719 on MSR-VTT, and at 0.500 against 0.489, 0.451 against 0.442 and 0.301 against 0.235 on the alignment maps, so the form that counts every query is used throughout.

Correlation within each alignment map. The alignment correlations of Section 5 are taken over the 116 configurations that mix all four maps. Table 12 reads them one map at a time, where the threshold is effectively constant across targets, taking three distinct values within a map, so the score varies through the target's margins. The correlation is 0.648 under orthogonal Procrustes, whose losses range most widely, and 0.636 on Flickr30k and 0.671 on MSR-VTT for the same map, the latter two at $p < 0.001$. Under ASIF, every pair loses nearly the same amount, a standard deviation of 4.4 points, so the quantity to predict is the common level of the loss, which the level calibrated on the other 28 pairs of the same map recovers with a mean absolute error of 0.032. Under least-squares regression and the canonical correlation solve, the losses do differ across pairs and the score does not order them. This is the case Corollary 2 leaves open: it ranks two targets only when one's decisive margins stochastically dominate the other's, and the risk score replaces that distribution with the mass below a single threshold, which two targets can share while losing different amounts. Calibrating each pair on its own threshold gives 0.516, -0.147, -0.388 and -0.118 for the four maps, recovering the ordering under Procrustes alone. Two things carry the forecast through this case. The level is still recovered, the reference mean predicting a target's loss to within 0.050 under least-squares regression and 0.060 under the canonical correlation solve, which is the severity term of Equation 6 doing the work exposure cannot. And the case announces itself before the target is

touched, since the share of the loss variance a transformation fixes is computable from the reference runs alone and is largest exactly where exposure adds least (Section 5 and Appendix I).

Table 12: Correlation between risk and measured Recall@1 loss within each alignment map, over its 29 encoder pairs, with the range and standard deviation of that loss in percent. The last two columns give the mean absolute error of each pair's predicted loss when calibrated on the other 28 pairs of the same map, using the map's mean loss level and using the risk-to-loss relation fitted within the map. Risk uses thresholds from the other 28 encoder pairs.

Map	Loss range	s.d.	ρ	p-value	MAE, level	MAE, relation
Orthogonal Procrustes	-0.3 to 54.4	11.5	0.648	< 0.001	0.076	0.066
Least-squares regression	-2.0 to 28.4	7.1	0.005	0.978	0.050	0.060
ASIF	56.8 to 77.8	4.4	-0.037	0.849	0.032	0.034
CCA	11.2 to 61.8	8.8	0.088	0.649	0.060	0.064

Every first-stage map, on both datasets. Section 5 reports compression after the canonical correlation solve. Table 13 gives the complete grid: each of the four alignment maps as the first stage, on all thirteen independently pretrained units, namely the ten image–text pairs on MSCOCO and the three MSR-VTT directions built from three encoders that never shared a training signal. The three-modality joint solve is excluded because the other three maps have no two-view analogue for it. The number of cells differs by map because the compressions available to a unit are those whose output rank is below its aligned width, and the canonical correlation solve emits 256 or 128 dimensions where the other maps emit the encoder's own width or, for ASIF, the anchor count. Two of the thirteen units fall close to the floor under Procrustes and ASIF: the DINOv2-B and DistilRoBERTa-base pair aligns to a Recall@1 of 0.2–0.3%, so its relative loss has a near-zero denominator and ranges from -0.83 to $+4.0$ across its 24 cells. Those cells are kept in the table, and excluding the units below 1% moves the pooled correlations from 0.701, 0.592 and 0.619 to 0.753, 0.654 and 0.672.

Table 13: Compression after alignment, every first-stage map and both datasets. Rank correlation between predicted risk and measured relative Recall@1 loss under three calibrations: the unit's own threshold, the median threshold of the other aligned units of the same map, and the median over the 29 jointly pretrained MSCOCO pairs. Aligned Recall@1 is the mean over the units of the row, and is the ambient accuracy that the compressions then act on.

First stage	Dataset	n	Units	Aligned Recall@1	Own	Other units	MSCOCO
Orthogonal Procrustes	MSCOCO	120	10	7.5%	0.571	0.612	0.655
	MSR-VTT	36	3	3.9%	0.848	0.895	0.923
	both	156	13	6.7%	0.641	0.657	0.700
Least-squares regression	MSCOCO	120	10	14.1%	0.848	0.634	0.661
	MSR-VTT	36	3	4.7%	0.928	0.847	0.847
	both	156	13	11.9%	0.854	0.672	0.694
ASIF	MSCOCO	120	10	4.6%	0.626	0.535	0.561
	MSR-VTT	36	3	5.4%	0.837	0.560	0.689
	both	156	13	4.8%	0.652	0.509	0.548
Canonical correlation analysis	MSCOCO	98	10	7.1%	0.932	0.927	0.902
	MSR-VTT	30	3	4.8%	0.935	0.904	0.914
	both	128	13	6.5%	0.925	0.902	0.893
All four maps	both	596	13	7.5%	0.701	0.592	0.619

Sensitivity to the order-preservation criterion. Section 4 estimates $\hat{\delta^*}$ by reading the measured preservation curve at a fixed criterion of 0.55. The criterion sets the scale of $\hat{\delta^*}$: raising it moves the read-off point further along the curve, so the resulting estimates grow, and the median $\hat{\delta^*}$ over all 464 MSCOCO configurations runs from 0.0090 at 0.55 to 0.0273 at 0.65, a factor of three. The criterion leaves the ordering the predictor induces essentially unchanged. Repeating the risk analysis under each criterion, with thresholds from the other 28 encoder pairs, Table 14 gives the correlation

between predicted risk and measured Recall@1 loss on both transformation families. Across a threefold change in $\hat{\delta^*}$ it varies by 0.016 on the compression cells and by 0.037 on the alignment cells. Risk is a quantile of each pair's own margin distribution, and a criterion shift relocates that quantile for every configuration at once, which is why the ranking survives it.

Table 14: The risk analysis repeated under three order-preservation criteria. The median $\hat{\delta^*}$ is taken over all 464 configurations of the 29 jointly pretrained MSCOCO pairs. Correlations are Spearman between predicted risk, with thresholds from the other 28 encoder pairs, and measured Recall@1 loss, over the 348 compression configurations and the 116 post-hoc alignment configurations. The paper uses 0.55 throughout.

Criterion	Median $\hat{\delta^*}$	ρ, compressions (n=348)	ρ, alignment (n=116)
0.55	0.0090	0.887	0.500
0.60	0.0180	0.882	0.522
0.65	0.0273	0.871	0.485

H SUPPORTING MEASUREMENTS FOR SECTIONS 3, 4 AND 5

The margins and thresholds of every pair. Figure 10 shows the ambient top-1 decision margins of all 29 MSCOCO pairs against their thresholds (Section 4 tests the interaction on the two pairs with the widest margins). Where $\hat{\delta^*}$ reaches the margins of every pair, as under ASIF, every target loses heavily and the risk score has little left to order, which is the ASIF row of Table 12, and there the severity term carries the level to a mean absolute error of 0.032.

Compression families differ in how broadly they perturb the ranking. Gaussian random projection eventually disrupts the whole ranking as well as its top. On CLIP ViT-B/16 (LAION-2B) at rank 16 it costs 21.99% of separability, against 2.88% for principal component analysis at the same rank, which removes 88.9% of Recall@1. The transformation therefore determines how broadly the ranking is perturbed, and substantial top-ranked loss arises well before whole-ranking separability shows a comparable change. This is the family dependence that the magnitude calibration of Section 5 absorbs. Separability falls appreciably only as $\hat{\delta^*}$ grows: across the twelve compressions the median $\hat{\delta^*}$ and the mean separability loss are rank-correlated at $\rho = 0.96$, rising together from 0.0017 and 0.00% for 8-bit quantization to 0.037 and 21.0% for random projection to 16 dimensions.

The threshold does not depend on the gallery size. Recall@1 is harder over a larger gallery, since more candidates compete with the true partner and the top-1 margins shrink. The threshold, a property of the transformation, should not move. For the 29 MSCOCO pairs and the twelve compressions, galleries of 500, 1,000, 2,000 and 5,000 items drawn from all MSCOCO image-caption pairs, each compression fitted on the gallery it compresses, give thresholds within 2% of those estimated at the evaluation split (median ratio 1.01 to 1.02), while the median top-1 margin falls from 2.88 to 1.74 and 1.35 times the threshold at 500, 2,000 and 5,000 items. The risk score read against the one threshold calibrated at the evaluation split therefore rises with the gallery, from a mean of 0.276 to 0.450, as the measured loss rises from 0.424 to 0.545. It ranks the losses at $\rho = 0.882$, 0.879, 0.895 and 0.885 at the four sizes and 0.874 over all of them together, and within each configuration it follows the loss across sizes at a mean ρ of 0.688, in the right direction in 91% of the 348.

Product quantization. Four product-quantized galleries (PQ with 64, 32, 16 and 8 sub-quantizers of 16 centroids, from 256 to 32 bits) and two with an optimized rotation (OPQ at 128 and 64 bits) extend the compressions on the 29 MSCOCO pairs. The codebooks are trained with FAISS on each pair's fit and selection splits and the query stays exact, as in product-quantization search. Recall@1 is taken in expectation over tied scores and the threshold counts a tied candidate pair as half preserved, since coarse codes produce exact ties. PQ removes 42%, 65%, 81% and 89% of Recall@1 on average from 256 to 32 bits, OPQ 45% and 59% at 128 and 64 bits. Over the six product-quantized codes the risk score ranks the losses at $\rho = 0.722$, against 0.777 for severity, and within a single code at a mean ρ of 0.545, close to the 0.594 of the twelve compressions.

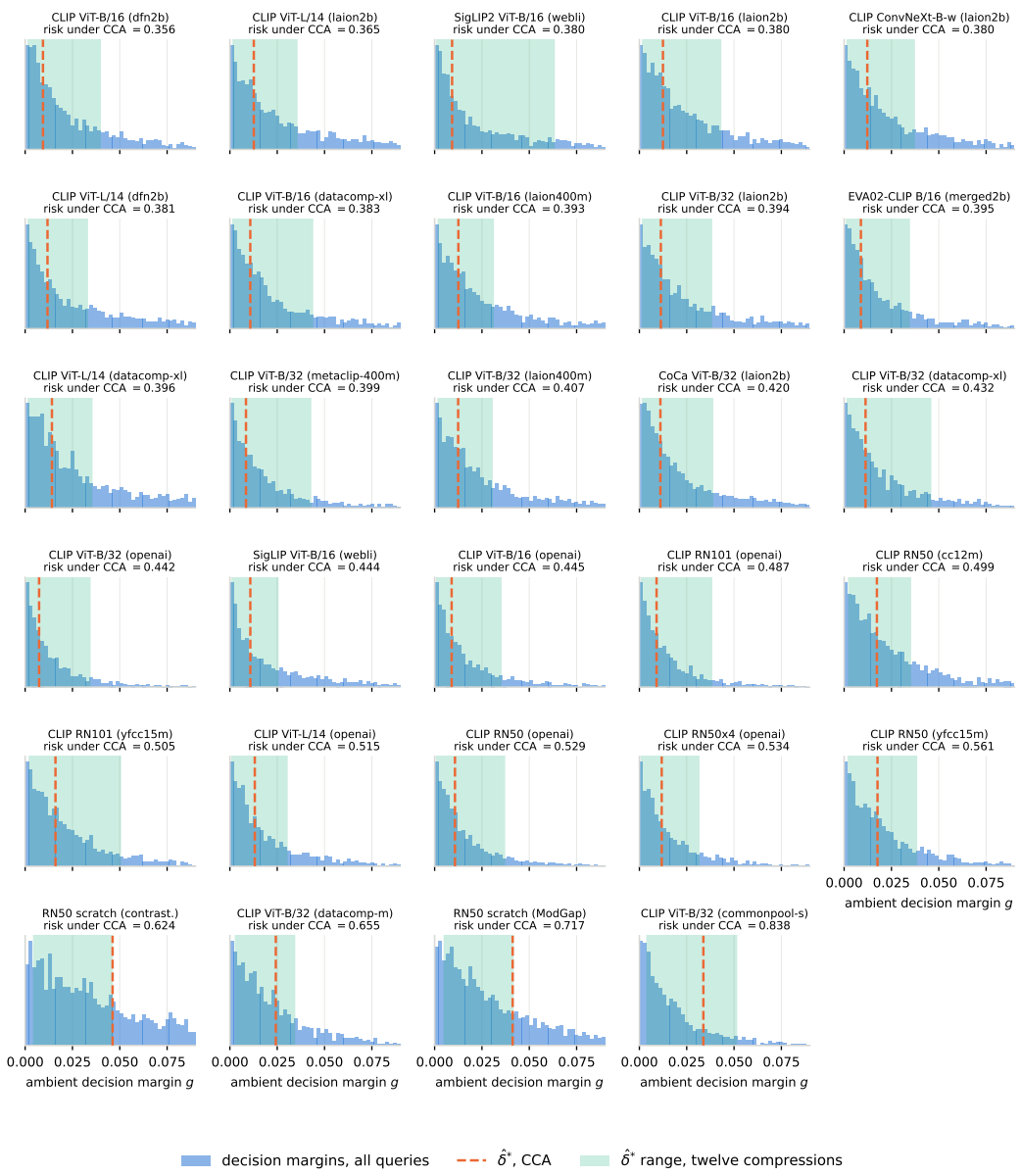


Figure 10: Ambient top-1 decision-margin distribution, over all queries, of every jointly pretrained MSCOCO encoder pair, ordered by risk under CCA. This is the complete form of Figure 3, which shows three of the 29. The distribution is a property of the encoder pair alone, and the transformation supplies the threshold. Dashed lines mark each pair's $\hat{\delta}^*$ for CCA, and the shaded band the range of $\hat{\delta}^*$ that the twelve compressions produce on that pair. Margins are drawn to 0.09, which truncates the upper tail of the widest-margin pairs.

The thresholds, by transformation and dataset. Table 15 gives the median $\hat{\delta}^*$ each transformation produces on each dataset, with the range over that dataset's encoder pairs. The scale is a property of the transformation: 8-bit quantization sits at 0.0017 to 0.0020 on all three datasets, random projection grows from 0.008 at output rank 256 to between 0.033 and 0.037 at rank 16, and on the image–text datasets it places a larger threshold than principal component analysis at every rank, by up to a factor of 3.4. The estimate also varies across the pairs of one dataset, by up to a factor of eight and a half on the image–text datasets and more widely on MSR-VTT, whose units mix three retrieval directions, which is why calibration takes the median over references (Appendix G). Under

the alignment maps the variation follows each pair's similarity spread. Within each map, $\log \hat{\delta}^*$ rises with the log standard deviation of the pair's off-diagonal ambient similarities at a median slope of 0.64, 0.84 and 1.01 on MSCOCO, Flickr30k and MSR-VTT, and dividing by that spread narrows the median ratio between the largest and the smallest threshold from 4.95, 3.47 and 30.3 to 4.45, 2.05 and 11.7. Under compression the slopes are 0.15, 0.21 and -0.05. Scaling each target's threshold by its own spread lowers the pooled rank correlation of the risk score with the loss in all six settings, from 0.887 to 0.692 on MSCOCO compression, since the spread already shapes the target's margins, and the median over references is kept.

Table 15: Median order-preservation threshold $\hat{\delta}^*$ of every transformation on every dataset, with the range over the encoder pairs of that dataset in brackets, at the 0.55 criterion. Estimated on each pair from its own ambient and transformed similarities, with no relevance label. Two cells of the $1{,}360$, both on MSR-VTT, reverse orderings at every margin and so admit no threshold; they are omitted here and are still scored as targets by the reference threshold.

Transformation	MSCOCO	Flickr30k	MSR-VTT
PCA-256	0.0057 [0.0030, 0.0120]	0.0064 [0.0047, 0.0158]	0.0097 [0.0044, 0.0752]
PCA-128	0.0056 [0.0035, 0.0121]	0.0065 [0.0048, 0.0158]	0.0105 [0.0046, 0.0741]
PCA-64	0.0064 [0.0046, 0.0122]	0.0070 [0.0052, 0.0162]	0.0106 [0.0048, 0.0706]
PCA-32	0.0080 [0.0052, 0.0128]	0.0085 [0.0065, 0.0167]	0.0118 [0.0055, 0.0670]
PCA-16	0.0107 [0.0079, 0.0199]	0.0106 [0.0070, 0.0166]	0.0132 [0.0068, 0.0669]
RP-256	0.0085 [0.0065, 0.0112]	0.0081 [0.0062, 0.0101]	0.0085 [0.0050, 0.0101]
RP-128	0.0121 [0.0083, 0.0158]	0.0121 [0.0084, 0.0154]	0.0105 [0.0076, 0.0159]
RP-64	0.0176 [0.0123, 0.0245]	0.0163 [0.0130, 0.0213]	0.0168 [0.0084, 0.0242]
RP-32	0.0262 [0.0167, 0.0435]	0.0247 [0.0174, 0.0324]	0.0239 [0.0113, 0.0303]
RP-16	0.0368 [0.0179, 0.0630]	0.0338 [0.0241, 0.0501]	0.0333 [0.0203, 0.0661]
8-bit quant.	0.0017 [0.0010, 0.0052]	0.0020 [0.0012, 0.0036]	0.0017 [0.0010, 0.0039]
4-bit quant.	0.0049 [0.0019, 0.0065]	0.0046 [0.0017, 0.0066]	0.0023 [0.0016, 0.0067]
Orth. Procrustes	0.0083 [0.0046, 0.0169]	0.0087 [0.0056, 0.0155]	0.0109 [0.0047, 0.1666]
Least squares	0.0076 [0.0058, 0.0131]	0.0083 [0.0054, 0.0226]	0.0122 [0.0053, 0.1091]
ASIF	0.0108 [0.0068, 0.0576]	0.0120 [0.0088, 0.0209]	0.0192 [0.0074, 0.1885]
CCA	0.0118 [0.0074, 0.0462]	0.0217 [0.0151, 0.0861]	0.0280 [0.0105, 0.4024]

The range the estimator resolves. Reading $\hat{\delta}^*$ off a preservation curve requires the criterion to be crossed inside the range of margins the bins resolve. A transformation that already preserves ordering at the smallest resolvable margins returns that bin's margin, which is where 8-bit quantization sits for about a third of the pairs of each dataset, at a mean Recall@1 change within 0.7% of ambient. A transformation that reverses orderings at every margin never reaches the criterion, which happens in two of the $1{,}360$ jointly pretrained cells, an orthogonal Procrustes map that raises MSR-VTT Recall@1 by 59.1% and a rank-16 random projection that removes 92.2% of it. Both are scored as targets by the reference threshold and enter no reference median.

H.1 Recovery after aligning independently pretrained encoders

Across the sixteen independently pretrained pairs, post-hoc GCCA raises match-versus-mismatch separability to 0.756–0.975. For the ten image–text pairs, Recall@1 averages 7.0%, rising to 35.7% at Recall@10 and 73.9% at Recall@50 (Figure 11). These aligned embeddings are the targets of the after-alignment setting of Section 5.

I Baselines and the two-term predictor

Definitions. Every predictor is calibrated on reference pairs that exclude the target, as in Appendix G, and none applies the transformation to the target or uses its labels. The *constant* baseline predicts the mean loss of the reference cells over all transformations. The query-performance predictors NQC (Shtok et al., 2012) and WIG (Zhou & Croft, 2007) are computed from the ambient similarities of each target query with $k = 10$, NQC as the standard deviation of the top-k similarities divided by the absolute mean similarity over the gallery, and WIG as the mean top-k similarity minus the mean over the gallery, the gallery mean taking the role of the corpus score of the original

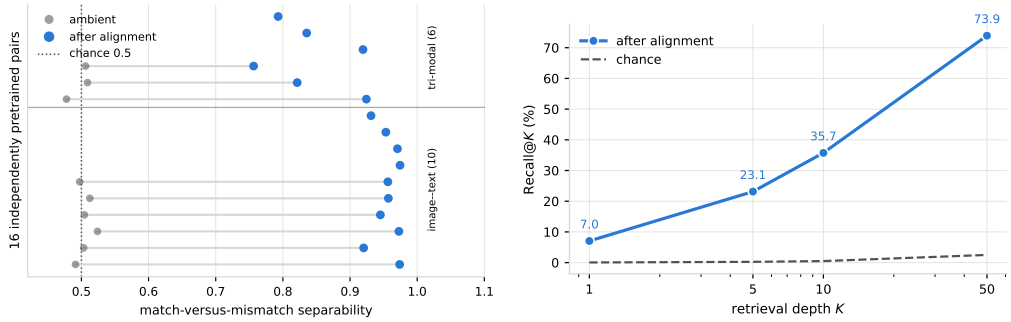


Figure 11: Recovery after aligning independently pretrained encoders. Left: separability after alignment for all sixteen pairs, grouped into ten image–text and six tri-modal pairs. Grey connectors mark ambient separability where defined. Seven pairs have no ambient value, owing to missing shared inner products or tri-modal training. Right: Recall@K for the image–text block, which shares a 2,000-candidate gallery and a common chance rate.

definitions. At the configuration level, the target's mean NQC or WIG over its queries is mapped to loss by a line fitted on the references. At the query level, among the queries a configuration answers correctly before the transformation, each score ranks the queries by their risk of being answered incorrectly after it, low NQC, low WIG or small g indicating risk, and the ranking is evaluated by AUROC. The *severity* term $\bar{L}_T$ is the mean relative Recall@1 loss of transformation T over the references, the model-wise mean value baseline of Xia et al. (2020). The *exposure* term alone is a line from risk to loss fitted on the references per compression family or per alignment map. The two-term predictor is Equation 6, whose correction is the target's centred percentile among the reference risks under the same transformation, with β the least-squares slope of the reference losses on that quantity after centring within each transformation, pooled over the transformations of the setting.

The form of the exposure correction. A correction driven by the raw difference $r - \bar{r}_T$ carries any offset between the target's risk scale and the references', which a percentile removes by construction. Over the 26 settings of every reference and target pairing of the two families and the four first-stage maps under both provenance calibrations, the percentile form ranks the losses better than the raw difference in 19 settings and predicts them more accurately in 18, and its largest error in any setting is 0.39 against 0.52. Measured against the severity term alone, it raises the rank correlation in 19 of the 26 settings, against 13 for the raw difference, and never falls more than 0.037 below it. Both forms are available in the released code.

The role of the threshold. The risk score combines two inputs, the target's margins and a threshold calibrated for the candidate transformation. Table 2 separates their contributions. Replacing the per-transformation threshold with a single threshold for the whole family, which leaves the score varying with the target alone, costs most of the correlation in five of the six settings: under compression from 0.887 to 0.252 on MSCOCO, from 0.741 to −0.030 on Flickr30k and from 0.804 to 0.295 on MSR-VTT, and on the alignment maps from 0.500 to 0.120 on MSCOCO and from 0.451 to 0.073 on Flickr30k. The target's median margin on its own behaves the same way. On MSR-VTT alignment, the three readings are stable and lie within 0.043 of one another. The signal is therefore the margin distribution read at the scale the transformation sets. The two inputs play distinct roles. The threshold is a property of the transformation and not of the target, so on its own it orders transformations, at $\rho = 0.830$, 0.896 and 0.719 under compression and 0.672, 0.730 and 0.138 under alignment, and leaves the targets of any one transformation indistinguishable. The margins carry the differences between targets and lose most of their correlation with the loss once the transformation's scale is taken out of the reading. Only the two together order both, each target's margins read at each transformation's scale.

Reading the whole margin distribution. Proposition 1 gives the expected share of queries whose top-1 decision reverses as the mean of $1 - F(g/\sigma_T)$ over the target's margins, of which the risk score keeps the mass below one threshold. Reading every margin through the preservation curve of Fig-

ure 2, in units of the reference threshold, gives rank correlations with the relative Recall@1 loss of 0.869, 0.764 and 0.792 under compression on MSCOCO, Flickr30k and MSR-VTT, against 0.887, 0.741 and 0.804 for the risk score, and 0.484, 0.475 and 0.244 on the alignment maps, against 0.500, 0.451 and 0.301. Within each transformation the mean correlations under compression are 0.519, 0.185 and 0.529, against 0.594, 0.176 and 0.532, and a Gaussian F matched to the 0.55 criterion agrees with the measured curve to within 0.003. Calibrating instead, on the reference pairs, how often the actual top-1 decision survives as a function of $g/\hat{\delta}^*$, and reading the untransformed target through it, gives 0.902 on the MSCOCO compressions against 0.887, with a within-transformation mean of 0.421 against 0.594. The single threshold therefore carries the ranking information of the full curve, and the risk score keeps its simpler form.

All twenty-six settings. Table 16 gives every setting the paper counts: each of the three calibrations on each of the three datasets, separately for the compressions and the alignment maps, and each of the four first-stage maps after alignment under both provenance calibrations. Across them, the risk score reaches a median rank correlation of 0.665 with the measured relative Recall@1 loss, with an interquartile range of 0.50 to 0.82, and the predicted loss a median of 0.865 from 0.72 to 0.90. The risk score ranks above both query-performance predictors in 23 of the 26, and the two terms together have a lower error than every outside baseline in 22 and raise the rank correlation over severity alone in 19. The settings where they add least are those where the transformation already fixes most of the loss variance, the last column, the two being rank-correlated at -0.51.

Every setting and calibration. Table 17 extends Table 1 with the 27 public MSCOCO pairs on their own, with calibration on MSCOCO for Flickr30k and MSR-VTT, and with compression after alignment calibrated on MSCOCO. Table 18 gives paired bootstrap intervals for the two-term predictor against the severity term. The rank-correlation gain has an interval above zero in five of the eight settings. It includes zero on Flickr30k, whose losses are almost fixed by the transformation alone, and after alignment, where severity alone already reaches $\rho = 0.924$.

What each term captures. The severity term carries the part of the loss that the transformation fixes, and the exposure term the part that varies across targets under the same transformation. Their shares differ across settings. The transformation accounts for 65% of the variance of the loss on MSR-VTT, 78% on MSCOCO and 85% on Flickr30k, and the alignment map for 85% of the variance of the alignment losses. The error gain of the exposure term follows that remaining, target-specific share: 0.014 on MSR-VTT and 0.011 on MSCOCO, where the transformation leaves the most to explain, and -0.007 on Flickr30k, where it leaves the least. Within a single transformation, where the severity term assigns every target the same loss, the loss still varies across targets with a standard deviation of 0.157 on MSCOCO, 0.187 on MSR-VTT and 0.131 on Flickr30k, which the risk score orders as Section 5 reports. Where the transformation's share is largest, the severity term already matches the two terms together, at $\rho = 0.902$ against 0.904 on Flickr30k, and that share is available from the reference runs alone, so which term carries a setting is known before either is applied. The exposure term reads a target's risk against the references' mean risk, so it carries across datasets when targets and references share a risk scale. Under the MSCOCO thresholds, the MSR-VTT targets have a mean risk of 0.353 against 0.381 for the MSCOCO references, at mean losses of 0.462 and 0.504, and the calibration transfers. The Flickr30k targets, ranked against 807 candidates at a mean ambient Recall@1 of 68.5%, against $2{,}000$ candidates and 40.9% on MSCOCO, have a mean risk of 0.232 at a mean loss of 0.454, and their within-dataset calibration matches that scale. After post-hoc alignment of independently pretrained encoders, the aligned targets lose 0.341 on average under the twelve compressions, against 0.553 for the MSCOCO pairs, and the exposure term, which reads the level from the targets' own margins, predicts their loss with an error of 0.164. The two terms are complementary by construction: severity is the term that carries the transformation and exposure the term that separates targets under it, so a predictor asked to rank transformations and to size a particular target's loss needs both.

Transforming the target. Section 5 compares the two-term predictor with the share of top-1 decisions a compression changes on the transformed target, mapped to a loss by a line fitted on the reference pairs with the target excluded. Across corpus sizes the transformed reading and the two terms reach errors of 0.087 and 0.098, and on the three MSR-VTT directions of the GRAM five-modality checkpoint 0.116 and 0.146. On the 27 public MSCOCO pairs alone the two terms size

Table 16: Every one of the twenty-six settings, with the references that calibrate it. Risk ρ and QPP ρ: rank correlation with the measured relative Recall@1 loss of the risk score and of NQC and WIG mapped to a loss by a line fitted on the references, bold where the risk score exceeds both. MAE: mean absolute error of the constant reference mean, of NQC and WIG, of the severity term, and of Equation 6, bold where the two terms beat every outside baseline. Var.: the share of the loss variance the transformation itself fixes, computable from the reference runs alone. The after-alignment rows pool both datasets, so their risk correlations are those of the *both* rows of Table 13.

				QPP ρ		MAE					
Target	References	n	Risk ρ	NQC	WIG	Const.	NQC	WIG	Severity	Both	Var.
Compression											
MSCOCO	MSCOCO	348	**0.887**	+0.17	+0.32	0.320	0.305	0.294	0.111/0.851	**0.100**/0.904	0.78
Flickr30k	MSCOCO	324	**0.772**	+0.01	+0.24	0.327	0.330	0.331	0.100/0.920	**0.145**/0.892	0.85
MSR-VTT	MSCOCO	348	**0.849**	+0.22	+0.42	0.300	0.300	0.293	0.159/0.806	**0.140**/0.868	0.65
MSCOCO	Flickr30k	348	**0.878**	+0.24	+0.33	0.319	0.314	0.371	0.126/0.878	**0.140**/0.857	0.78
Flickr30k	Flickr30k	324	**0.741**	−0.16	+0.22	0.325	0.325	0.316	0.081/0.902	**0.088**/0.904	0.85
MSR-VTT	Flickr30k	348	**0.846**	+0.22	+0.42	0.301	0.300	0.267	0.161/0.801	**0.166**/0.776	0.65
MSCOCO	MSR-VTT	348	**0.822**	+0.24	+0.33	0.319	0.304	0.386	0.143/0.874	**0.122**/0.917	0.78
Flickr30k	MSR-VTT	324	**0.635**	+0.01	+0.24	0.324	0.325	0.315	0.117/0.908	**0.170**/0.871	0.85
MSR-VTT	MSR-VTT	348	**0.804**	−0.16	+0.41	0.303	0.305	0.269	0.156/0.776	**0.143**/0.861	0.65
Alignment											
MSCOCO	MSCOCO	116	**0.500**	+0.11	+0.15	0.174	0.170	0.169	0.054/0.789	**0.060**/0.892	0.85
Flickr30k	MSCOCO	108	**0.392**	+0.03	+0.06	0.223	0.223	0.222	0.095/0.920	**0.122**/0.906	0.88
MSR-VTT	MSCOCO	116	0.367	+0.26	+0.40	0.202	0.191	0.181	0.221/0.281	0.221/0.351	0.09
MSCOCO	Flickr30k	116	**0.492**	+0.15	+0.16	0.179	0.177	0.167	0.090/0.909	**0.093**/0.913	0.85
Flickr30k	Flickr30k	108	**0.451**	−0.15	−0.04	0.223	0.223	0.224	0.055/0.802	**0.057**/0.846	0.88
MSR-VTT	Flickr30k	116	0.295	+0.26	+0.40	0.184	0.183	0.174	0.234/0.281	0.233/0.345	0.09
MSCOCO	MSR-VTT	116	**0.545**	+0.15	+0.16	0.182	0.206	0.174	0.138/0.775	**0.146**/0.783	0.85
Flickr30k	MSR-VTT	108	**0.497**	+0.03	+0.06	0.223	0.225	0.222	0.181/0.703	**0.182**/0.706	0.88
MSR-VTT	MSR-VTT	116	0.301	+0.04	+0.36	0.185	0.173	0.178	0.181/0.138	0.178/0.341	0.09
After alignment											
Procrustes	other aligned	156	**0.657**	−0.33	−0.06	0.417	0.424	0.419	0.269/0.724	**0.294**/0.703	0.32
Least squares	other aligned	156	**0.672**	+0.10	+0.05	0.221	0.219	0.220	0.066/0.857	**0.064**/0.884	0.87
ASIF	other aligned	156	**0.509**	+0.20	+0.31	0.389	0.363	0.356	0.316/0.404	**0.305**/0.497	0.19
CCA	other aligned	128	**0.902**	−0.04	+0.22	0.269	0.268	0.261	0.084/0.882	**0.090**/0.897	0.88
Procrustes	29 MSCOCO	156	**0.700**	+0.13	+0.19	0.426	0.443	0.423	0.308/0.789	**0.394**/0.807	0.32
Least squares	29 MSCOCO	156	**0.694**	+0.11	+0.11	0.288	0.314	0.244	0.217/0.903	0.320/0.912	0.87
ASIF	29 MSCOCO	156	**0.548**	+0.35	+0.42	0.378	0.335	0.335	0.311/0.502	**0.291**/0.614	0.19
CCA	29 MSCOCO	128	**0.893**	+0.08	+0.23	0.303	1.085	0.265	0.191/0.913	**0.191**/0.911	0.88

Table 17: Mean absolute error and rank correlation (MAE / ρ) of the predicted relative Recall@1 loss for every setting and calibration. Constant: the mean reference loss over all transformations. Exposure: the risk score mapped to a loss by a line fitted on the references per compression family or alignment map. Severity and Both as in Table 1. Dashes: not evaluated.

Setting	References	n	Constant	NQC	WIG	Severity	Exposure	Both
Compression, MSCOCO	28 other pairs	348	0.320	0.305 / 0.17	0.294 / 0.32	0.111 / 0.851	0.128 / 0.890	0.100 / 0.904
27 public pairs	26 other pairs	324	0.311	0.311 / 0.01	0.301 / 0.22	0.077 / 0.916	0.123 / 0.892	0.070 / 0.945
Alignment, MSCOCO	28 other pairs	116	0.174	0.170 / 0.11	0.169 / 0.15	0.054 / 0.789	0.056 / 0.886	0.060 / 0.892
Compression, Flickr30k	26 other pairs	324	0.325	0.325 / −0.16	0.316 / 0.22	0.081 / 0.902	0.187 / 0.741	0.088 / 0.904
	29 MSCOCO pairs	324	0.327	0.330 / 0.01	0.331 / 0.24	0.100 / 0.920	0.277 / 0.752	0.145 / 0.892
Compression, MSR-VTT	18 other models	348	0.303	0.305 / −0.16	0.269 / 0.41	0.156 / 0.776	0.153 / 0.807	0.143 / 0.861
	29 MSCOCO pairs	348	0.300	0.300 / 0.22	0.293 / 0.42	0.159 / 0.806	0.142 / 0.849	0.140 / 0.868
After alignment	29 MSCOCO pairs	98	0.293	–	–	0.218 / 0.924	0.164 / 0.914	0.216 / 0.915

Table 18: Paired bootstrap over encoder pairs (over models on MSR-VTT, 2,000 resamples) of the two-term predictor against the severity term: mean gain in MAE (severity minus both) and in ρ (both minus severity), with 95% intervals.

Setting	References	MAE gain	ρ gain
Compression, MSCOCO	28 other pairs	$+0.011$ $[-0.009, 0.031]$	$+0.051$ $[0.006, 0.101]$
27 public pairs	26 other pairs	$+0.007$ $[-0.001, 0.015]$	$+0.028$ $[0.008, 0.046]$
Alignment, MSCOCO	28 other pairs	-0.005 $[-0.012, 0.003]$	$+0.101$ $[0.069, 0.137]$
Compression, Flickr30k	26 other pairs	-0.007 $[-0.011, -0.003]$	$+0.002$ $[-0.010, 0.011]$
	29 MSCOCO pairs	-0.045 $[-0.066, -0.024]$	-0.027 $[-0.075, 0.005]$
Compression, MSR-VTT	18 other models	$+0.014$ $[-0.004, 0.036]$	$+0.084$ $[0.030, 0.142]$
	29 MSCOCO pairs	$+0.019$ $[-0.000, 0.041]$	$+0.062$ $[0.016, 0.112]$
After alignment	29 MSCOCO pairs	$+0.002$ $[-0.013, 0.021]$	-0.009 $[-0.028, 0.010]$

the loss more closely, at 0.070 against 0.084, and after alignment they do so again, at 0.216 against 0.235, with the exposure term alone at 0.164. The transformed reading does not extend to the alignment maps, whose fitting on the target needs the target's pairs.

A transformed subsample. The transformed pass also works on part of the target. On the full 5,000-item MSCOCO test corpus, for each of the 348 configurations of the twelve compressions on the 29 pairs, we apply the compression to a random subsample of B queries and their true items, count the share of top-1 decisions it changes within the subsample, and map that share to a loss by a line fitted on the other 28 pairs, averaging three draws. Fitted against the references' losses on their own subsamples, the line inherits the smaller gallery, where each query has fewer rivals and the mean loss is 0.424 at $B = 500$ against 0.545 on the full corpus, and it predicts the full loss at an error of 0.141, 0.113 and 0.094 for $B = 500$, 1,000 and 2,000. Fitted against the references' full-corpus losses, it reaches 0.071, 0.069 and 0.082, at $\rho = 0.969$, 0.974 and 0.983. The same line on the share of changed decisions over the whole corpus reaches 0.108 at $\rho = 0.988$, and the two-term predictor 0.104 at $\rho = 0.903$, with the risk score alone at $\rho = 0.885$. Per configuration on one GPU, the transformed pass takes 8.6, 14.8 and 33.5 ms at the three budgets and 150 ms on the full corpus, and the risk score takes 0.05 ms after one 61 ms pass that extracts the target's margins for every candidate.

Both terms without labels. The label-free predictor of Section 5 takes the references' share of changed top-1 decisions as the severity term and fits the exposure slope on those shares against the centred percentile of the references' risk. The label-free severity term alone ranks the compression losses at $\rho = 0.845$ on MSCOCO, 0.913 on Flickr30k and 0.782 on MSR-VTT. Read directly as a loss, with no mapping, the label-free predictor has a mean absolute error of 0.168, 0.133 and 0.204, so the reference labels serve the sizing of the loss and the ranking needs none.

Identifying the queries that a compression breaks. Table 3 is computed over every compression configuration in which at least one query answered correctly before the transformation is answered incorrectly after it, namely all 348 on MSCOCO, 322 of 324 on Flickr30k and 329 of 348 on MSR-VTT. The top-1 decision margin g needs no threshold and no reference run for this use, so a target's queries at risk can be listed before any calibration exists. The number of those configurations in which a score reads above chance is 335, 324 and 215 on MSCOCO, 319, 313 and 284 on Flickr30k and 305, 298 and 268 on MSR-VTT, for g, NQC and WIG respectively.

The two perturbation-based predictors for dense retrieval, re-implemented on the ambient similarities, read the compression rows as follows. Noisy perturbation (Arabzadeh et al., 2023) in its original form, the overlap of the top-10 sets after ten isotropic perturbations of norm 0.1 of the query embedding, scores 0.510, 0.512 and 0.512 on MSCOCO, Flickr30k and MSR-VTT, and projection displacement (Datta et al., 2025), the change of the top-10 similarities after projection onto the span of the top five candidates, scores 0.493, 0.507 and 0.477. The stability of the top-1 candidate under larger noise suits the question better, at 0.667, 0.660 and 0.693 for norm 0.4 and 0.703, 0.711 and 0.738 for norm 0.8. As the noise grows that stability becomes a sampled reading of the top-two gap itself, and it approaches g from below. At the configuration level, mapped to a loss by a line fitted on

the references as NQC and WIG are, the four readings reach errors between 0.276 and 0.327 under compression, close to the 0.269 to 0.325 of NQC and WIG.

Table 19 carries the axis to the other seven settings, the four alignment maps pooled per dataset and compression after each first-stage map. With the three compression rows of Table 3, the margin reads above both query-performance predictors in nine of the ten settings, the exception being the alignment maps on MSR-VTT, where NQC leads it by 0.025.

Table 19: Mean AUROC on the query axis for the settings beyond Table 3. Among the queries a configuration answers correctly before the transformation, each score ranks them by the risk of being answered incorrectly after it. g is the top-1 decision margin, which needs no threshold, and NQC and WIG are the query-performance baselines. n counts the configurations in which at least one such query fails. Bold marks the settings in which g reads above both baselines.

Family	Target	n	g	NQC	WIG
Alignment	MSCOCO	114	**0.652**	0.611	0.512
	Flickr30k	106	**0.648**	0.626	0.555
	MSR-VTT	110	0.561	0.586	0.527
After alignment	Orthogonal Procrustes	147	**0.684**	0.613	0.561
	Least-squares regression	153	**0.727**	0.613	0.561
	ASIF	139	**0.706**	0.676	0.594
	Canonical correlation analysis	127	**0.699**	0.597	0.573

Choosing a compression under a tolerable loss. For each target, compression family and tolerance $\mathcal{T}$ on the relative Recall@1 loss, each predictor selects the most aggressive setting of the family whose predicted loss is at most $\mathcal{T}$, and the realised loss of that choice is measured. Each family is a ladder of settings ordered by aggressiveness, PCA and random projection at ranks 256 down to 16 and quantization at 8 then 4 bits, so a choice is one step on one ladder and the shortfall against the oracle is counted in steps. An oracle with the target's true losses selects the most aggressive admissible setting. Table 20 reports the share of choices that exceed the tolerance and the mean number of family steps left unused relative to the oracle. The two quantities trade against each other, since a predictor that forecasts larger losses selects milder settings, with fewer violations and more unused compression. The severity term assigns every target of a family the same forecast and sits on the conservative side of this trade-off on MSCOCO, where it leaves more compression unused at the 10% and 20% tolerances (0.49 and 0.44 steps against 0.39 and 0.31) and slightly less at 5% (0.30 against 0.33). With calibration on MSCOCO, the choices on Flickr30k and MSR-VTT inherit the risk scale of the references described above. At the 20% tolerance on MSCOCO the predicted loss is the better of the two on both counts at once, exceeding the tolerance in 8% of choices against 14% while leaving 0.31 steps unused against 0.44.

J Recall@K at other depths, and the choice under a tolerance

Section 5 reads Recall@K at the gap between the K-th and the $(K+1)$-th highest ambient similarities, which reduces to g at $K = 1$. The threshold is estimated from randomly drawn candidate pairs, so the same threshold serves every depth and only the margin statistic and the measured loss change. Table 21 repeats the risk score and the predicted loss of Section 5 at the three depths under both calibrations. At $K = 1$ the numbers reproduce those of Table 17, and at $K = 5$ and $K = 10$ both quantities improve, the predicted loss beating the reference average in every setting at $K = 5$ and in all but one at $K = 10$. Deeper ranks have more correctly answered queries per target, so the same reordering costs a smaller share of them, the mean relative loss falling from 0.504 to 0.360 on MSCOCO between Recall@1 and Recall@10 (Table 21, mean loss columns).

The choice under a tolerance, at multiple depths. Table 22 gives the selection experiment of Section 5 over the 423 choices formed by the five target-and-calibration settings above, three compression families each, at nine tolerances. The predicted loss is better than the severity term on both counts, exceeding the tolerance less often and leaving less compression unused, from 20% at Recall@1, from 7.5% at Recall@5 and at the 4%, 5% and 10% tolerances at Recall@10. Below

Table 20: Compression choice under a tolerance $\mathcal{T}$ on the relative Recall@1 loss, target excluded from calibration: share of choices exceeding the tolerance / mean family steps left unused relative to an oracle, over 3 families per target (87 choices per tolerance on MSCOCO and MSR-VTT, 81 on Flickr30k).

Dataset	References	Predictor	$\mathcal{T} = 5\%$	$\mathcal{T} = 10\%$	$\mathcal{T} = 20\%$
MSCOCO	28 other pairs	Both	0.02 / 0.33	0.06 / 0.39	0.08 / 0.31
		Severity	0.00 / 0.30	0.07 / 0.49	0.14 / 0.44
		Exposure	0.01 / 0.30	0.02 / 0.39	0.10 / 0.45
Flickr30k	26 other pairs	Both	0.11 / 0.30	0.14 / 0.24	0.21 / 0.25
		Severity	0.15 / 0.20	0.14 / 0.16	0.20 / 0.22
		Exposure	0.00 / 0.49	0.10 / 0.58	0.14 / 0.79
	29 MSCOCO pairs	Both	0.22 / 0.19	0.32 / 0.14	0.20 / 0.24
		Severity	0.00 / 0.37	0.00 / 0.59	0.06 / 0.42
		Exposure	0.40 / 0.21	0.44 / 0.21	0.43 / 0.14
MSR-VTT	18 other models	Both	0.05 / 0.40	0.06 / 0.48	0.10 / 0.43
		Severity	0.01 / 0.37	0.05 / 0.55	0.21 / 0.57
		Exposure	0.05 / 0.39	0.05 / 0.46	0.09 / 0.52
	29 MSCOCO pairs	Both	0.15 / 0.41	0.13 / 0.35	0.11 / 0.36
		Severity	0.01 / 0.37	0.01 / 0.55	0.17 / 0.52
		Exposure	0.08 / 0.31	0.08 / 0.31	0.14 / 0.36

those tolerances, the severity term, which assigns every target of a family the same forecast and so selects milder settings, exceeds the tolerance less often at the cost of more unused compression. The tight end of the grid is also where the task itself leaves little room: at a 1% tolerance the oracle declines to compress at all in 64% of the choices at Recall@1 and 58% at Recall@5 and Recall@10, against 37%, 27% and 20% at 20%, so a forecast that systematically under-compresses is correct by default there.

K LICENSES

All datasets and checkpoints we use are publicly available. Table 23 lists each one with its license as published by its authors. We use each artefact within the terms of its license and for non-commercial academic research.

Datasets. MSCOCO, Flickr30k and MSR-VTT are all available and usable for research under their respective licenses and terms of use, which Table 23 lists. We access them through their standard public distributions and redistribute none of them. The pretraining corpora behind the checkpoints we evaluate are never used and enter this paper only through the released weights of those checkpoints.

Table 21: The forecast at three retrieval depths. Risk ρ: rank correlation between the risk score, read at the K-th boundary, and the measured relative Recall@K loss. Predicted and Reference: mean absolute error and rank correlation (MAE / ρ) of the predicted loss of Equation 6 and of the severity term alone, bold where the predicted loss is the better of the two. Mean loss: the mean measured relative Recall@K loss over the configurations of the row. Within: references are the other encoder pairs of the dataset, the target's group excluded. MSCOCO: references are the 29 jointly pretrained MSCOCO pairs. Twelve compressions per target.

			Recall@1			Recall@5			Recall@10			Mean loss		
Target	References	n	Risk ρ	Predicted	Reference	Risk ρ	Predicted	Reference	Risk ρ	Predicted	Reference	K=1	5	10
MSCOCO	within	348	0.887	**0.100 / 0.904**	0.111 / 0.851	0.937	**0.077 / 0.937**	0.096 / 0.871	0.936	**0.071 / 0.934**	0.087 / 0.875	0.504	0.410	0.360
Flickr30k	within	324	0.741	0.088 / **0.904**	0.081 / 0.902	0.914	**0.054 / 0.954**	0.061 / 0.921	0.928	**0.047 / 0.957**	0.053 / 0.920	0.454	0.341	0.288
	MSCOCO	324	0.772	0.145 / 0.892	0.100 / 0.920	0.933	**0.088 / 0.946**	0.092 / 0.943	0.946	0.105 / **0.954**	0.088 / 0.941	0.454	0.341	0.288
MSR-VTT	within	348	0.804	**0.143 / 0.861**	0.156 / 0.776	0.878	**0.098 / 0.914**	0.144 / 0.766	0.854	**0.092 / 0.903**	0.127 / 0.773	0.462	0.366	0.318
	MSCOCO	348	0.849	**0.140 / 0.868**	0.159 / 0.806	0.920	**0.118 / 0.890**	0.149 / 0.794	0.915	**0.115 / 0.876**	0.138 / 0.792	0.462	0.366	0.318

Table 22: Choosing a compression under a tolerance on the relative Recall@K loss: share of choices exceeding the tolerance / mean family steps left unused relative to an oracle, over 423 choices per cell. Predicted: Equation 6. Severity: the reference average. Bold marks the predictor that is better on both counts.

	Recall@1		Recall@5		Recall@10	
Tolerance	Predicted	Severity	Predicted	Severity	Predicted	Severity
1%	0.12 / 0.27	**0.04 / 0.19**	0.13 / 0.23	**0.02 / 0.23**	0.11 / 0.23	**0.01 / 0.25**
2%	0.10 / 0.30	**0.02 / 0.25**	0.11 / 0.25	**0.03 / 0.27**	0.11 / 0.30	**0.03 / 0.38**
3%	0.10 / 0.30	**0.02 / 0.28**	0.11 / 0.26	0.02 / 0.36	0.10 / 0.36	0.02 / 0.46
4%	0.10 / 0.31	**0.01 / 0.30**	0.10 / 0.26	0.02 / 0.38	**0.08 / 0.40**	0.10 / 0.47
5%	0.11 / 0.33	**0.03 / 0.32**	0.10 / 0.29	0.02 / 0.44	**0.06 / 0.37**	0.09 / 0.51
7.5%	0.12 / 0.33	0.02 / 0.42	**0.07 / 0.36**	0.09 / 0.53	0.07 / 0.38	0.13 / 0.36
10%	0.14 / 0.32	0.05 / 0.47	**0.07 / 0.33**	0.13 / 0.39	**0.06 / 0.39**	0.09 / 0.46
15%	0.13 / 0.29	0.10 / 0.43	**0.06 / 0.28**	0.13 / 0.34	0.07 / 0.36	0.14 / 0.35
20%	**0.14 / 0.32**	0.16 / 0.44	**0.09 / 0.27**	0.12 / 0.42	0.09 / 0.31	0.13 / 0.31

Table 23: Licenses of the datasets and checkpoints used in this paper. Checkpoint licenses are those published with the released weights.

Artefact	Source	License
Datasets		
MSCOCO annotations	Lin et al. (2014)	CC BY 4.0
MSCOCO and Flickr30k images	Flickr	Flickr Terms of Use
Flickr30k captions	Young et al. (2014)	Research use, see text
MSR-VTT	Xu et al. (2016)	Research use, see text
Jointly pretrained image–text checkpoints		
CLIP, OpenAI weights	Radford et al. (2021)	MIT
OpenCLIP weights (LAION, YFCC-15M, CC12M, DataComp, CommonPool, CoCa, ConvNeXt)	`laion/*`, `timm/*`	MIT
MetaCLIP ViT-B/32	`timm/vit_base_patch32_clip_224`	CC BY-NC 4.0
DFN-2B ViT-B/16, ViT-L/14	`apple/DFN2B-CLIP-ViT-*`	Apple AMLR
EVA02-CLIP B/16	`timm/eva02_base_patch16_clip_224`	MIT
SigLIP, SigLIP 2 ViT-B/16	`timm/ViT-B-16-SigLIP*`	Apache-2.0
ResNet-50 pairs	trained here	Released with our code
Video, audio and tri-modal checkpoints		
GRAM, three variants	Cicchetti et al. (2025)	MIT
TRIANGLE	Cicchetti et al. (2026)	MIT
VAST	Chen et al. (2023)	MIT
X-CLIP	`microsoft/xclip-base-patch{32,16-zero-shot}`	MIT
CLAP	`laion/clap-htsat-*`, `laion/larger_clap_*`	Apache-2.0
Independently pretrained encoders		
DINOv2-B, DINOv2-L	`facebook/dinov2-{base,large}`	Apache-2.0
ViT-MAE	`facebook/vit-mae-base`	Apache-2.0
BEiT	`microsoft/beit-base-patch16-224`	Apache-2.0
DeiT	`facebook/deit-base-patch16-224`	Apache-2.0
Swin	`microsoft/swin-base-patch4-window7-224`	Apache-2.0
ResNet-50	`microsoft/resnet-50`	Apache-2.0
MPNet	`sentence-transformers/all-mpnet-base-v2`	Apache-2.0
MiniLM	`sentence-transformers/all-MiniLM-L6-v2`	Apache-2.0
BERT-base	`bert-base-uncased`	Apache-2.0
DistilRoBERTa-base	`distilroberta-base`	Apache-2.0
BEATs	`microsoft/unilm`	MIT