ALIGNMENT WITH DIFFERENT PAIRS: CROSS-MODAL SIMILARITY AND RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

We distinguish three questions about cross-modal alignment: what structure each modality has, whether the two modalities organize designated partners alike, and whether the deployed scorer ranks those partners highly. Sensitivity to random shuffling does not establish that relational agreement orders pairings by original-pair fidelity, and correlation with recall does not establish its value for checkpoint selection. We introduce hard negative re-pairing: an external sentence encoder guides a one-to-one reassignment to semantically similar captions from other images, while representations and, in retrieval experiments, all image–text scores remain fixed. On COCO, hard negative re-pairing with no original partners retained yields higher centered kernel alignment (CKA) than the mean under random re-pairing retaining most original partners. Across 22 global image–text checkpoints, neighborhood agreement retains a larger fraction of its original-minus-random contrast than assigned-pair recall. On Flickr30k, selecting checkpoints by validation agreement rather than validation recall costs 5.75 percentage points of held-out bidirectional Recall@1, despite a positive agreement–recall correlation. Relational agreement alone does not establish original-pair fidelity, and checkpoint-selection criteria should be judged by the held-out retrieval performance of their choices.

1 INTRODUCTION

Cross-modal diagnostics such as centered kernel alignment (CKA) (Kornblith et al., 2019) and neighborhood agreement are used to ask whether vision and language representations organize corresponding examples similarly (Huh et al., 2024). CKA falls toward zero when image–caption partners are randomly shuffled (Maniparambil et al., 2024). This drop shows that the diagnostic is sensitive to the designated image–caption pairing. We ask what such relational agreement establishes about the original pairing and retrieval performance.

Interpreting this evidence requires distinguishing three questions:

- **Q1.** What structure does each modality have on its own?
- **Q2.** Do the two modalities organize designated partners alike?
- **Q3.** Does the deployed scorer rank those partners highly?

We measure marginal structure using local intrinsic dimensionality (LID) (Levina & Bickel, 2005), neighbor radii, and effective rank of the covariance spectrum (Roy & Vetterli, 2007); relational agreement using CKA and neighborhood agreement; and retrieval using Recall@K. Exact controls demonstrate how these diagnostics separate. We then evaluate hard negative re-pairing on COCO (Lin et al., 2014) and test whether agreement guides retriever selection on held-out Flickr30k images (Young et al., 2014).

The decrease in CKA under random shuffling motivates recovering image–caption partners by maximizing CKA (Maniparambil et al., 2024). That control does not establish an ordering by original-pair fidelity—the fraction of original pairs retained—because it leaves semantically similar replacements untested. Separately, relational agreement is reported to rise with model capability (Huh et al., 2024; but see Koepke et al., 2026). A positive trend across models does not show that agreement selects well. We test selection among trained retrievers, where each choice can be checked on held-out data.

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.

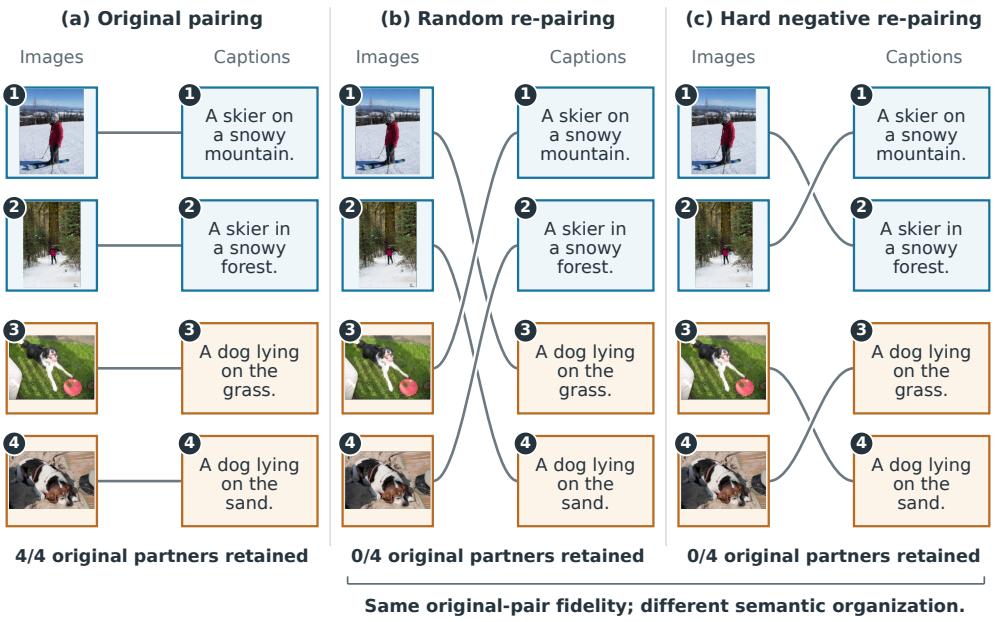


Figure 1: Hard negative re-pairing vs random re-pairing. All three panels use the same images and captions. Both re-pairings use every caption once and retain no original partners. Hard negative re-pairing selects replacements similar to the original partners; random re-pairing ignores semantic similarity. Colors indicate semantic categories, and numbered IDs identify original pairs. Representations remain fixed throughout; in retrieval experiments, every image–text score also remains fixed.

As shown in Figure 1, we introduce *hard negative re-pairing* to test the missing case. An external sentence encoder guides a one-to-one reassignment to semantically similar captions from other images, without consulting the evaluated models. Comparing hard negative and random re-pairing at matched original-pair fidelity isolates the effect of replacement choice. Both representation collections and, in the retrieval studies, all image–text scores remain fixed; only the designated pairing changes.

Our contributions are:

- **A diagnostic distinction** that separates marginal, relational, and retrieval questions by the information each consumes, with exact controls demonstrating their separation.
- **Hard negative re-pairing**, an externally selected control that changes designated partners based on semantic information.
- **Three findings**: higher CKA does not imply higher original-pair fidelity; hard negative re-pairing preserves a larger fraction of the original-minus-random contrast in neighborhood agreement than in assigned-pair recall; and selecting checkpoints by agreement rather than recall incurs a held-out retrieval cost despite a positive agreement–recall correlation.

2 Related Work

Testing what a diagnostic tracks. Kornblith et al. (2019) study which transformations a similarity index should ignore. Section 4 uses invariances of this kind as exact controls. Ding et al. (2021) test whether similarity measures follow model behavior. Davari et al. (2023) show that CKA can change substantially when function does not. We show a converse: CKA can stay high when every original partner is gone. For cross-modal pairs, the usual reference is a random pairing. CKA falls as partners are shuffled, which motivates recovering pairs by maximizing CKA (Maniparambil et al., 2024). A random shuffle, however, removes original partners and scrambles semantic organization at the same time. So a shuffling curve does not show which of the two a diagnostic follows. Hard negative re-pairing separates them (Figure 1). Since relational diagnostics compare structure within

each modality, close substitutes are expected to keep some of their value. How much they keep on real data is an empirical question, which Section 5 measures.

What cross-modal agreement consists of. Huh et al. (2024) report that cross-modal agreement rises with model capability. Two recent studies narrow this reading. Gröger et al. (2026) calibrate similarity against random pairings. After calibration, the convergence seen in global measures mostly vanishes, while local neighborhood agreement stays significant across modalities. Koepke et al. (2026) find that agreement drops as the dataset grows. They argue that what remains is shared coarse semantics rather than shared fine structure. Hard negative re-pairing asks what this remaining agreement consists of. It intervenes on the pairing itself. Each re-paired image receives a semantically close caption from another image. The representations and the retrieval scores stay fixed. The name comes from retrieval, where a hard negative is a non-matching caption that resembles the true one (Faghri et al., 2018). Hard negative benchmarks use such captions to test Q3: whether the scorer ranks the true caption above them (Yuksekgonul et al., 2023; Hsieh et al., 2023). We make them the designated partners and measure Q2 and Q3 on the same pairing.

Relational agreement as a selection proxy. Representational similarity already guides choices before training: which layers to align (Gröger et al., 2025) and which encoders to pair (Maniparambil et al., 2025). Model-selection research judges a proxy by the downstream performance of the model it selects (Renggli et al., 2022). We apply this test to neighborhood agreement among trained retrievers, where each choice can be checked on held-out retrieval. Our test does not cover selection before training. It asks whether the reported link between agreement and capability (Huh et al., 2024) is strong enough to pick a model when the pick can be verified.

3 Setup and hard negative re-pairing

3.1 Evaluation setup

We describe the evaluation of an encoder by the state $E = (I, T, p, s)$, where $I = (v_1, \ldots, v_n)$ and $T = (t_1, \ldots, t_n)$ are the embedding representations of the n images in the index and the n text queries, respectively, the bijection $p : [n] \to [n]$ encodes the relevance judgments by mapping each image index i to its relevant text index $p(i)$, and $s : T \times I \to \mathbb{R}$ is the scoring function. For each query t_i, the images are ranked by $s(t_i, v_j)$, and we report Recall@K (R@K), the fraction of queries whose relevant image $v_{p^{-1}(i)}$ appears among the top K results.

Here $[n] = \{1, \ldots, n\}$; a text unit is a single caption in the retrieval studies or a caption group in the CKA study. We report both retrieval directions and their arithmetic mean. The score matrix $S_{ij} = s(t_j, v_i)$ has image rows and text columns.

The original dataset pairing is $p_0(i) = i$. Its retained fraction under p, the *original-pair fidelity*, is

$$F(p) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[p(i) = i]. \tag{1}$$

This counts retained dataset partners; it does not assess whether a replacement is also a valid description.

Re-pairing replaces p_0 with another one-to-one assignment p of the same text units to images, while holding I, T, and s fixed. Retrieval scores and candidate rankings therefore remain unchanged.

3.2 Hard negative re-pairing

Definition. Hard negative re-pairing chooses replacement text partners that are semantically similar to each image's original partner while still being non-matching (i.e., not the original partner). Comparing it with random re-pairing at the same replacement rate tests whether a diagnostic responds to *which* partners replace the originals, beyond how many original pairs are retained.

Feasible assignments. At replacement rate $q > 0$, fix a subset $\mathcal{I}_q$ of image identities, with $q = |\mathcal{I}_q|/n$. A feasible assignment permutes the text partners within $\mathcal{I}_q$ and leaves all other partners

fixed. Every text partner is used exactly once. Within $\mathcal{I}_q$, each image must receive another image's text partner, and replacements identical to the original after text normalization are excluded. Let $\mathcal{P}_q$ be the set of feasible assignments. Every $p \in \mathcal{P}_q$ has original-pair fidelity $F(p) = 1 - q$ (Equation 1); at $q = 1$, every feasible assignment is a *derangement*: a bijection with no fixed point ($p(i) \neq i$ for all i), so $F(p) = 0$.

Semantic assignment. We choose replacements using an external sentence encoder: MPNet (Song et al., 2020), through Sentence-BERT (Reimers & Gurevych, 2019). Let z_i be its unit-normalized representation of image i's original text partner. These selector vectors are separate from the evaluated representations t_i; no evaluated representations, scores, or outcomes enter the assignment. We choose a feasible assignment minimizing the total cosine dissimilarity between original and replacement text partners,

$$p_q^\star \in \arg\min_{p \in \mathcal{P}_q} \sum_{i \in \mathcal{I}_q} \left(1 - z_i^\top z_{p(i)}\right). \tag{2}$$

The matching is solved jointly over $\mathcal{I}_q$ to enforce the one-to-one constraint.

Section 5.1 and Section 5.2 also test alternative matchings with nearly the same total selector cost, and assignments chosen using COCO object annotations instead of MPNet. Construction is described in Appendix A.4 (Tables A.1 and A.2).

3.3 Diagnostics

Marginal diagnostics. For the exact controls, we compute local intrinsic dimensionality (LID) (Levina & Bickel, 2005), neighbor radii, and effective rank of the covariance spectrum (Roy & Vetterli, 2007) separately within each modality. Their definitions and implementation details are given in Appendix A.2.

Neighborhood agreement. Let $N_k^I(i)$ and $N_k^T(j)$ be the k nearest Euclidean neighbors of image i and text unit j within their respective modalities, excluding the item itself and breaking distance ties by index. Under pairing p,

$$A_k(p) = \frac{1}{nk} \sum_{i=1}^{n} \left| N_k^I(i) \cap p^{-1}\left(N_k^T(p(i))\right) \right|. \tag{3}$$

This is the mean fraction of an image's neighbors whose text partners are also neighbors of its own text partner. We use $k = 50$ as the primary scale and $k \in \{10, 20, 100\}$ as secondary scales. For an unrestricted uniformly random bijection, the expected agreement is $c_k = k/(n-1)$; Lemma A.1 gives the proof, and Equation A.4 the chance-corrected reporting convention.

CKA. Centered kernel alignment (Kornblith et al., 2019) compares within-modality kernels. We build an RBF kernel for each modality, with its bandwidth set by that modality's median nonzero squared distance, and center both kernels. With K_c^I and K_c^T the centered kernels,

$$\mathrm{CKA}(p) = \frac{\sum_{i,j} (K_c^I)_{ij} (K_c^T)_{p(i),p(j)}}{\|K_c^I\|_F \|K_c^T\|_F}, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. Re-pairing relabels the rows and columns of the text kernel; neither kernel is recomputed, and both norms remain unchanged.

3.4 Random reference and retention

The random assignments are generated from the same feasible set $\mathcal{P}_q$, so they change the same images and have the same fidelity as $p_q^\star$. For a diagnostic M, let $\overline{M}_{\mathrm{ref}}(q)$ be its mean over these seeded random assignments. The generator is not known to be uniform over $\mathcal{P}_q$, so this mean is a reference for the chosen procedure rather than a uniform null.

When the denominator is nonzero, define retention as

$$\rho_M(q) = \frac{M(p_q^\star) - \overline{M}_{\mathrm{ref}}(q)}{M(p_0) - \overline{M}_{\mathrm{ref}}(q)}. \tag{5}$$

$\rho_M(q) = 1$ means the semantic assignment matches the original-pairing score; $\rho_M(q) = 0$ means it matches the random-reference mean. Values are not clipped and may fall below zero or exceed one. For bidirectional recall, we average the two directional recalls within each condition before applying Equation 5.

4 WHY THE THREE QUESTIONS IN THE INTRODUCTION COME APART

We can formulate the introduction's three questions as follows:

- **Q1. What structure does each modality have on its own?** *Marginal diagnostics* use only I or only T, without the pairing p or the cross-modal score matrix S.
- **Q2. Do the two modalities organize designated partners alike?** *Relational diagnostics*, including neighborhood agreement $A_k(p)$ and CKA (Section 3.3), compare within-modality structure under p. They use I, T, and p, but not S.
- **Q3. Does the deployed scorer rank designated partners highly?** *Retrieval diagnostics*, such as assigned-pair Recall@K, use rankings from S and the relevance judgments specified by p.

The controls in Table 1 separate every pair of diagnostic families: for each pair, one diagnostic can remain fixed while the other changes.

We apply the controls to SigLIP B/16 224 (Zhai et al., 2023) on 5,000 COCO pairs (Lin et al., 2014), keeping the scoring rule s fixed. We assess marginal structure using local intrinsic dimension (LID), neighbor radii, and effective rank; relational agreement using neighborhood agreement A_{50}; and retrieval using bidirectional Recall@1. Table 1 reports changes at the maximum prespecified control strengths. The one-sided rotation separation also holds for CLIP ViT-B/16 (Radford et al., 2021) and SigLIP2 B/16-224 (Tschannen et al., 2025); shared rotation and radial scaling also preserve normalized retrieval in both models (Appendix B.4, Table B.3).

Proposition 1 (Orthogonal invariance)**.** *At fixed p, applying orthogonal transformations Q_I and Q_T to I and T, respectively, preserves within-modality Euclidean distances, covariance spectra, and neighbor graphs, and hence the marginal diagnostics used here and $A_k(p)$. The distance-based RBF-CKA in Section 3.3 is also unchanged. For equal-width unit-normalized representations with $s(t, v) = \alpha\langle t, v\rangle + \beta$ and fixed $\alpha > 0, \beta$, a shared rotation $Q_I = Q_T$ additionally preserves S and all candidate rankings; a one-sided rotation need not.*

Proposition 2 (Fixed marginals under re-pairing)**.** *Changing only p to another bijection preserves both representation collections and every marginal diagnostic. The score matrix S and candidate rankings remain fixed; only the relevance judgments encoded by p change.*

Proposition 3 (Positive radial cancellation)**.** *Let $t_i = \nu(\tilde{t}_i)$, where $\tilde{t}_i \neq 0$ and $\nu(x) = x/\|x\|_2$. For any $a_i > 0$, $\nu(a_i\tilde{t}_i) = \nu(\tilde{t}_i)$. Thus, at fixed I, p, s, positive sample-wise scaling preserves T, the score matrix S, candidate rankings, and Recall@K, while raw geometry can change.*

Proofs are in Appendix B.1; the control construction and measured responses are in Appendices B.2 to B.4.

Control	**Q1: Marginal Diagnostic**	**Q2: Relational Diagnostic (Neighborhood Agreement)**	**Q3: Retrieval Diagnostic (Recall)**
One-sided rotation (T)	Fixed	Fixed	-47.2 points
Random derangement (p)	Fixed	-96.2%	-47.2 points
Radial scaling (raw $\tilde{T}$)	Text LID $+45.8\%^{\dagger}$	$-48.7\%^{\dagger}$	Fixed
Shared rotation (I, T)	Fixed	Fixed	Fixed

Table 1: **Control responses separate the three diagnostic families.** Fixed entries follow from the corresponding invariance propositions. Percentages are relative changes; recall changes are percentage points. †Measured on raw projected vectors; normalized geometry remains fixed.

Under one-sided rotation, Proposition 1 guarantees that the marginal diagnostics (Q1) and relational agreement (Q2) remain fixed, while bidirectional Recall@1 (Q3) falls by 47.2 percentage points. This separates Q3 from both Q1 and Q2. Under random derangement, Proposition 2 guarantees that Q1 remains fixed, while A_{50} falls by 96.2%, separating Q1 from Q2. These two controls therefore cover all three pairs.

Proposition 3 explains why raw geometry can change while normalized retrieval remains fixed. Shared rotation provides a consistency check in which all three diagnostic families remain unchanged.

5 Experiments And Results

The three studies apply the diagnostics of Section 3.3 to the interventions of Section 3.2. Section 5.1 asks whether CKA orders pairings by original-pair fidelity, and Section 5.2 whether relational agreement and assigned-pair retrieval respond alike to the same re-pairing; both studies measure, on real encoders, the separation that Section 4 established by construction. Section 5.3 uses original pairs only, with no re-pairing, and asks whether selecting checkpoints by validation agreement rather than validation recall changes held-out retrieval; it tests a practical consequence of that separation.

Models. The CKA study uses DINOv2 image representations (Oquab et al., 2023) and All-RoBERTa caption representations (Sentence-Transformers contributors, 2021). The retrieval study evaluates 22 global image–text checkpoints: four CLIP, six SigLIP, six SigLIP2, E5-V (Jiang et al., 2024), Jina-CLIP-v2 (Koukounas et al., 2024), two LongCLIP variants (Zhang et al., 2024), and Qwen3-VL-Embedding-2B/8B (Li et al., 2026). Checkpoint selection uses the 16 CLIP, SigLIP, and SigLIP2 checkpoints. For these families, table labels identify model size, patch size, and input resolution: for example, B/16 224 denotes a base model with 16-pixel patches and 224 $\times$ 224 input images.

Evaluation and metrics. The re-pairing studies use 5,000 COCO validation images (Lin et al., 2014). We compare original, hard negative, and random pairings with representations fixed; retrieval scores and rankings also remain fixed. We report RBF-CKA in the CKA study and reference-adjusted retention of neighborhood agreement at $k = 50$ and bidirectional Recall@1 and Recall@10 in the retrieval study. Checkpoint selection uses original pairs and disjoint Flickr30k validation and test images (Young et al., 2014). Study-specific settings are given below, with full checkpoint identifiers, preprocessing, and implementation details in Appendices C.1, D.1 and E.1.

5.1 CKA and original-pair fidelity

Maniparambil et al. (2024) use CKA to match images with their original captions. They hold a set of known image–caption pairs fixed and search for a one-to-one pairing of the remaining images and captions that maximizes CKA between the image and text representations. This search is motivated by their observation that CKA between DINOv2 and All-RoBERTa falls as more image-caption pairs get randomly shuffled and fidelity (Equation 1) drops. We reproduce their random-shuffling curve using their released RBF-CKA settings. We use DINOv2's mean-pooled image representations and average All-RoBERTa's caption embeddings within each image's caption group. We likewise average the external selector's caption embeddings within each group.

This motivation fails under hard negative re-pairing. Figure 2 compares random and hard negative re-pairing, applied to the same images at each replacement rate. At 0.8 original-pair fidelity, hard negative re-pairing gives a CKA of 0.701, compared with 0.462 for random re-pairing. Even when no original pairs remain, hard negative re-pairing scores 0.659—higher than random re-pairing at a fidelity of 0.8. The original pairing still scores highest (0.719), but higher CKA need not indicate higher original-pair fidelity. The annotation-only assignment produces the same reversal, with a CKA of 0.547 when fidelity is zero. It selects replacement partners using COCO object labels rather than a text encoder, so the reversal does not depend on MPNet and All-RoBERTa sharing a training lineage. This reversal also holds across 128 alternative matchings, with linear CKA, and with row-normalized representations (Tables C.4 and C.5).

We next examine whether this disagreement also arises among matchings generated by optimizing CKA. (Appendix C.6 gives the setup.) For each of three sampled COCO query sets, we run FAQ

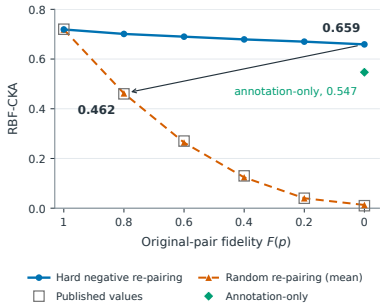


Figure 2: RBF-CKA versus original-pair fidelity for DINOv2/All-RoBERTa. Solid and dashed lines show hard negative re-pairing and mean random re-pairing, respectively, with the same images reassigned in both conditions at each fidelity level. Open squares show published values from Maniparambil et al. (2024).

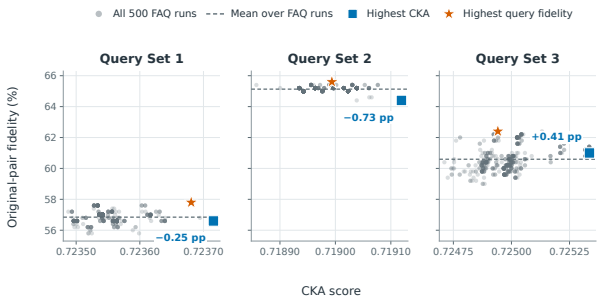


Figure 3: CKA-based selection and original-pair fidelity. Each panel shows one of three COCO query sets, each sampled with its own random seed. Each point is a candidate matching from one of 500 randomized FAQ starts, with 320 anchor correspondences and 500 queries held fixed within each panel. CKA is computed over all 820 nodes. Blue squares mark the highest-CKA candidates; orange stars mark the highest-fidelity candidates in the same pool. The dashed horizontal line in each panel marks the mean query fidelity over all 500 FAQ runs.

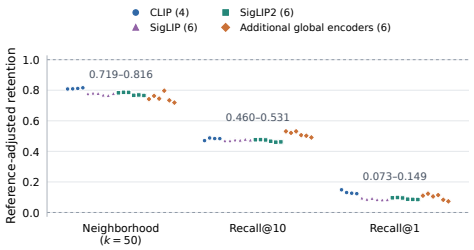


Figure 4: Retention ordering across 22 image–text checkpoints. Each point represents one checkpoint for one diagnostic under complete hard negative re-pairing on COCO (Section 3.4). All 22 checkpoints retain more neighborhood agreement ($k = 50$) than bidirectional R@10, and more R@10 than R@1. Horizontal offsets separate points; numerical labels give the observed ranges across checkpoints.

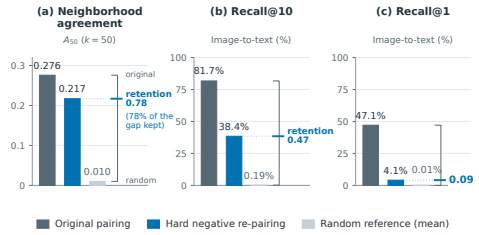


Figure 5: SigLIP B/16 224, image-to-text, under the original pairing, hard negative re-pairing, and the random reference; scores and rankings are identical across the three conditions, only the designated partners change. Brackets mark retention (Equation 5), the share of the random-to-original gap the hard negative condition keeps.

(Vogelstein et al., 2015), an approximate solver for the CKA-based quadratic assignment problem maniparambil2024vision, from 500 randomized initializations. Each problem contains 320 fixed anchor correspondences and 500 queries whose partners must be recovered (Figure 3).

Selecting the highest-CKA candidate yields query fidelity 0.25 and 0.73 percentage points below, and 0.41 points above, the mean over 500 FAQ runs in the three query sets, respectively. This mean is the expected fidelity of selecting a run uniformly at random. Thus, highest-CKA selection provides no consistent recovery advantage over this baseline across the three tested query sets.

5.2 Relational agreement versus assigned-pair retrieval

We apply hard negative re-pairing to 22 global image–text checkpoints. Each checkpoint uses its native processor. We select one caption per image using a fixed rule and evaluate each query against one designated partner. This differs from standard COCO evaluation with multiple captions per image; Appendix D.1 provides the protocol details.

Table 2: **Validation selection and held-out retrieval on Flickr30k.** A star marks a checkpoint selected by a validation criterion; boldface indicates the maximum in that validation column. Both recall columns report bidirectional R@1.

Checkpoint	Validation A_{50}	Validation R@1 (%)	Test R@1 (%)
CLIP B/32 224	0.2023	62.72	59.50
CLIP B/16 224	0.2072	66.37	61.40
CLIP L/14 224	0.2080	69.77	65.40
CLIP L/14 336	0.2142	71.94	67.95
SigLIP B/16 224	0.2311	75.49	74.35
SigLIP B/16 256*	**0.2330**	76.13	75.20
SigLIP B/16 384	0.2309	78.90	78.25
SigLIP L/16 256	0.2243	80.87	79.65
SigLIP L/16 384*	0.2220	**82.64**	80.95
SigLIP So400m/14 224	0.2061	76.38	75.65
SigLIP 2 B/16 224	0.2221	73.62	73.15
SigLIP 2 B/16 256	0.2216	75.74	74.70
SigLIP 2 B/16 384	0.2228	77.81	77.15
SigLIP 2 L/16 256	0.2211	78.30	77.35
SigLIP 2 L/16 384	0.2217	79.98	78.35
SigLIP 2 So400m/14 224	0.2258	78.94	77.85

Figure 4 shows that all 22 checkpoints, including the six additional encoders, share the same retention ordering. Under complete hard negative re-pairing, neighborhood agreement retains 72–82% of its original-minus-random gap, compared with 46–53% for bidirectional Recall@10 and 7–15% for bidirectional Recall@1. Figure 5 illustrates the separation for SigLIP B/16 224: neighborhood-agreement retention is 78%, whereas reassigned captions appear among the top ten candidates for 38.4% of image queries and rank first for 4.1% (per-checkpoint results in Table D.2). The new partners are close in object content. In 92% of images, the new caption comes from an image with at least one shared COCO object category, against 34% under one random re-pairing (Table D.8). Thus, the two modalities can continue to organize designated partners similarly even when the retriever rarely ranks those partners first. This separates Q2 and Q3 on real encoders. The ordering also holds at a matched scale: with $k = 10$, neighborhood agreement retains more than Recall@10 in all 22 checkpoints (Table D.10).

This ordering is empirical: although raw Recall@10 is always at least Recall@1, each metric has different original and random-reference values (Section 3.4). Additional retrieval controls and sensitivity analyses appear in Appendix D.6; Table D.9 lists the checkpoints each analysis covers. One may object that, with rankings fixed, a replacement partner cannot rank first when the original partner does. We therefore also examine queries whose original partner is not ranked first. For SigLIP B/16 224, assigned-pair Recall@1 remains only 7.75% image-to-text and 9.46% text-to-image within this subset (Appendix D.4). Thus, low assigned-pair Recall@1 persists even among queries where this constraint does not force failure. For the late-interaction image–text models ColQwen2.5 (Vidore, 2025a), ColQwen-Omni (Vidore, 2025b), and EVIE (Tencent, 2026), evaluated with native MaxSim scoring, Recall@10 retention exceeds Recall@1 retention in five of six retrieval directions and all three bidirectional aggregates; ColQwen2.5 image-to-text is the directional exception (Appendix F.2). The same ordering holds in five audio–text configurations on AudioCaps (Kim et al., 2019; Appendix F.3).

5.3 Checkpoint selection by neighborhood agreement or recall

We test whether neighborhood agreement is useful for selecting among already trained retrievers, assessing selection quality through the held-out retrieval performance of the chosen checkpoint (Renggli et al., 2022; Agostinelli et al., 2022).

We compare two checkpoint-selection rules over the same 16 CLIP, SigLIP, and SigLIP2 models: selecting the checkpoint with the highest validation neighborhood agreement A_{50}, or the highest validation bidirectional Recall@1. Both criteria are computed on the original image–caption pairs.

On Flickr30k (Young et al., 2014), selection uses 1,014 validation images and evaluation uses a disjoint set of 1,000 test images, with one deterministically selected caption per image.

Table 2 shows a positive correlation between agreement and recall across the validation panel ($\tau_b = +0.367$), but the two criteria select different checkpoints. Neighborhood agreement selects SigLIP B/16 256, while recall selects SigLIP L/16 384. On the held-out test, the two models reach 75.20% and 80.95% bidirectional Recall@1, respectively, a gap of 5.75 percentage points (95% paired bootstrap CI: $[3.80, 7.80]$; When any of an image's five captions counts as correct, the gap is 4.78 points (95% CI: $[3.62, 5.97]$; Table E.2). protocol and complete panel in Appendices E.1 to E.3).

Neighborhood agreement can correlate positively with recall and still favor a worse retriever. Its usefulness for checkpoint selection must therefore be judged by the held-out retrieval performance of its choice, not by correlation alone.

6 Discussion and Conclusion

CKA and neighborhood agreement are now used to match images to their captions (Maniparambil et al., 2024) and to pick encoders and layers for multimodal training (Maniparambil et al., 2025; Gröger et al., 2025), and their reported link to model capability (Huh et al., 2024) makes them tempting for choosing retrievers as well. Our exact controls show that the three questions from the introduction (Q1–Q3) can come apart: one diagnostic can stay fixed while another moves. Hard negative re-pairing shows the same thing on real encoders. A pairing with no original partners left can score higher CKA than one that keeps most of them. In other words, CKA rewards partners that are organized alike, whether or not they are the original ones. Neighborhood agreement does the same: it keeps most of its original-minus-random gap under hard negative re-pairing, while assigned-pair Recall@1 keeps little of it. The two modalities still organize their new partners alike, but the retriever rarely ranks those partners first. And among the 16 checkpoints we compared on Flickr30k, neighborhood agreement picked a worse retriever than validation recall did, even though the two were positively correlated.

None of this makes relational diagnostics useless. CKA drops steadily as random re-pairing lowers original-pair fidelity. Hard negative re-pairing is, after all, a constructed stress test, and original-pair fidelity does not tell us whether a swapped-in caption also describes its new image. Still, the underlying issue is what these diagnostics look at. They compare structure within each modality under a given pairing and never see the cross-modal scores a retriever ranks, so on their own they cannot vouch for original-pair fidelity or for retrieval. Our results bear on what the remaining cross-modal agreement consists of (Gröger et al., 2026; Koepke et al., 2026). With no original partner left, neighborhood agreement keeps most of its original-minus-random contrast. This holds between the towers of trained retrievers and between independently trained encoders (Figure 4; Table C.6). Top-1 retrieval keeps little of it. One question we leave open is whether our selection result carries over to CKA and to checkpoints from a single training run.

We suggest three things. Be explicit about which question a diagnostic answers (Q1, Q2, or Q3), and draw conclusions only about that one. When relational agreement is used to recover pairings or to argue for alignment, compare it against a structured alternative at matched fidelity, such as hard negative re-pairing, and not only against random shuffling or permutation nulls (Gröger et al., 2026). And when recall can be computed, select by validation recall. As a selection rule, relational agreement is worth using only before a retriever exists, and even then it should first be tested on trained retrievers by checking how the checkpoint it picks performs on held-out data (Renggli et al., 2022; Agostinelli et al., 2022). A positive correlation with recall is not enough, since selection only depends on which checkpoint comes out on top.

AI use statement

We used generative AI tools to assist with the proofs of Propositions 1–3 (Appendix B.1), to help write the code for our experiments, to research broadly about related work, and to help interpret the retention ordering reported in Section 5.2. We also used generative AI tools to review the manuscript for grammar, terminology consistency, and repeated wording, and to suggest edits; the authors reviewed each suggestion before applying it. The authors checked the AI-assisted proofs, and the

invariances the propositions establish were also confirmed numerically (Appendix B.3). All code, including AI-assisted code, was reviewed by the authors and validated with the checks described in Appendix E.3, Appendix C.7, and Appendix G.3. We take full responsibility for the content of this paper, including any text, proofs, or code produced with the aid of generative AI.

Reproducibility statement

All public inputs are identified by name and revision: the checkpoints in Tables C.1, D.1 and F.1, and the datasets and caption rules in Appendices C.1, D.1, E.1 and F. Appendix A defines the diagnostics, the semantic and random assignments, and the random references, and Table A.3 lists the seeds of every random re-pairing. Appendix B gives the proofs of Propositions 1–3 and the construction of the exact controls. Appendix G records what was fixed before each analysis, the software environments, and where the code and records for each result are found. The supplementary material contains the code and the compact records behind every table and figure; Table G.2 lists the additional public inputs needed for full recomputation.

References

Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *European Conference on Computer Vision*, volume 13694 of *Lecture Notes in Computer Science*, pp. 303–321, 2022. doi: 10.1007/978-3-031-19830-4_18. arXiv:2204.01403.

Sandra Arcos-Holzinger, Debashish Chakraborty, Rohita Mocharla, Will Walden, Andrew Yates, Reno Kriz, Sarah M. Erfani, James Bailey, Vishal M. Patel, and Sanjeev Khudanpur. Layerscope: A layerwise characterization of video and multimodal learned representations, 2026a. URL https://arxiv.org/abs/2609.28086.

Sandra Arcos-Holzinger, Sarah M. Erfani, James Bailey, and Sanjeev Khudanpur. GRIDS: Dimensionality-aware anomaly detection in learned representations of self-supervised speech models. *arXiv preprint arXiv:2605.02715*, 2026b.

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-Dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021. doi: 10.3390/e23101368.

Lorenzo Basile, Santiago Acevedo, Luca Bortolussi, Fabio Anselmi, and Alex Rodriguez. Intrinsic dimension correlation: Uncovering nonlinear connections in multimodal representations. In *International Conference on Learning Representations*, 2025.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.16156.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, 2018.

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

Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With limited data for multimodal alignment, let the STRUCTURE guide you. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2506.16895.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 36, 2023.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20617–20642, 2024. arXiv:2405.07987.

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, 2005.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

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, 2014.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E. O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14334–14343, 2024.

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*, 2025. arXiv:2409.19425.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

Cedric Renggli, André Susano Pinto, Luka Rimanic, Joan Puigcerver, Carlos Riquelme, Ce Zhang, and Mario Lučić. Which model to transfer? finding the needle in the growing haystack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9205–9214, 2022. arXiv:2010.06402.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, 2007.

Rickmer Schulte and David Rügamer. Rethinking intrinsic dimension estimation in neural representations. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, 2026.

Sentence-Transformers contributors. sentence-transformers/all-roberta-large-v1 model card. Hugging Face, 2021. URL `https://huggingface.co/sentence-transformers/all-roberta-large-v1`.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tencent. EVIE-Preview-4.5B. Hugging Face model repository, 2026. URL `https://huggingface.co/tencent/EVIE-Preview-4.5B`.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Vidore. ColQwen2.5-v0.2 model card. Hugging Face model repository, 2025a. URL `https://huggingface.co/vidore/colqwen2.5-v0.2`.

Vidore. ColQwen-Omni-v0.1 model card. Hugging Face model repository, 2025b. URL `https://huggingface.co/vidore/colqwen-omni-v0.1`.

Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Louis J. Podrazik, Steven G. Kratzer, Eric T. Harley, Donniell E. Fishkind, R. Jacob Vogelstein, and Carey E. Priebe. Fast approximate quadratic programming for graph matching. *PLOS ONE*, 10(4):e0121002, 2015. doi: 10.1371/journal.pone.0121002.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

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.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *European Conference on Computer Vision*, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *International Conference on Learning Representations*, 2024.

APPENDIX OVERVIEW

The appendices follow the order of the main text. Appendix A fixes conventions, marginal estimators, the chance correction, and the construction of assignments and random references. Appendix B proves Propositions 1–3 and reports the exact controls. Appendices C to E support the three studies of Section 5, Appendix F reports the operator and modality extensions, and Appendix G records protocols, environments, and artifacts. The table below links each main-text statement to its support. Throughout, *one random draw* denotes a single random re-pairing, and the *random reference* $\overline{M}_{\mathrm{ref}}$ denotes the mean over many draws that enters Equation 5.

Where each main-text statement is supported

Main text	Statement	Appendix support
Section 3.2	Alternative matchings; annotation-based selectors	Appendix A.4; Tables A.1 and A.2
Section 3.3	LID, neighbor radius, effective rank	Appendix A.2
Section 3.3	Chance level $k/(n-1)$; corrected reporting	Lemma A.1; Equation A.4; Corollary A.3
Section 3.4	Random reference and its draws	Appendix A.5; Table A.3
Section 4	Proofs of Propositions 1–3	Appendix B.1
Section 4, Table 1	Control construction; CLIP and SigLIP2 replications	Appendices B.2 and B.4; Tables B.2 and B.3
Section 5	Checkpoints, preprocessing, caption selection	Tables C.1 and D.1; Appendices D.1 and E.1
Section 5.1, Figure 2	Fidelity curve	Appendix C.2; Table C.2
Section 5.1	Annotation-only assignment	Appendix C.3
Section 5.1	128 alternatives; linear CKA; row normalization	Tables C.4 and C.5
Section 5.1, Figure 3	QAP setup and outcomes	Appendix C.6; Table C.7
Section 5.2	Protocol; one designated partner per query	Appendix D.1
Section 5.2, Figure 4	Per-checkpoint results for all 22	Tables D.2 and D.3
Section 5.2, Figure 5	Representative values; random reference	Tables D.4 and D.5
Section 5.2	Conditional R@1 of 7.75% and 9.46%	Appendix D.4; Table D.7
Section 5.2	Retrieval controls and sensitivity analyses	Appendix D.6; Table D.9
Section 5.2	Late interaction; audio	Tables F.2 to F.4
Section 5.3, Table 2	Flickr30k protocol, panel, bootstrap interval	Appendix E.1; Tables E.1 and E.2

A Additional diagnostic and assignment details

This appendix fixes the conventions used by all later appendices. It gives the marginal estimators and the chance correction for neighborhood agreement (Section 3.3), shows that the correction leaves retention unchanged (Section 3.4), and specifies how semantic assignments and random references are constructed (Section 3.2).

A.1 Notation, representation surfaces, and retrieval conventions

We use the evaluation state $\mathcal{E} = (I, T, p, s)$ of Section 3.1. The image and text representations are $I = (v_1, \ldots, v_n)$ and $T = (t_1, \ldots, t_n)$; the bijection $p : [n] \to [n]$ assigns image i its designated text partner $p(i)$; and s is the scoring function. The score matrix $S_{ij} = s(t_j, v_i)$ has image rows and text columns. The original pairing is $p_0(i) = i$, and original-pair fidelity $F(p)$ is defined in Equation 1.

Retrieval directions. Image-to-text retrieval ranks row i of S for image query v_i; its designated partner is $t_{p(i)}$. Text-to-image retrieval ranks column j for text query t_j; its designated partner is $v_{p^{-1}(j)}$. Recall@K (R@K) is the fraction of queries whose designated partner is among the first K candidates, and score ties are broken by candidate index. Bidirectional R@K is the arithmetic mean of the two directions. Because re-pairing leaves S unchanged, every query has the same candidate ranking under every pairing; only its designated partner changes. For bidirectional recall, the two directional recalls are averaged within each condition before Equation 5 is applied, so an aggregate retention is not the mean of the two directional retentions.

Representation surfaces. A surface is the exact array consumed by a diagnostic, including pooling and normalization. For a global encoder, projected vectors $\tilde{v}_i, \tilde{t}_i$ precede normalization, and the unit-normalized vectors $v_i = \nu(\tilde{v}_i)$, $t_i = \nu(\tilde{t}_i)$, with $\nu(x) = x/\|x\|_2$, are the vectors consumed by the native score. The retrieval studies (Appendix D) compute neighborhood agreement and recall on normalized vectors. The exact controls (Appendix B) evaluate the marginal and relational diagnostics on both surfaces. The CKA study (Appendix C) uses the saved rows of the published protocol without further normalization, and Appendix C.5 repeats it on row-normalized rows. Absolute LID is not compared across surfaces or encoders.

Text units. A text unit is one selected caption in the retrieval and selection studies and one caption group, all captions of an image averaged, in the CKA study (Appendix C.1).

A.2 MARGINAL DIAGNOSTICS

Marginal diagnostics describe one modality at a time on a declared surface. They use I or T alone, never p or S, and they are not combined into a scalar alignment score.

Local intrinsic dimensionality. At each $k \in \{10, 20, 50, 100\}$, we use the maximum-likelihood estimator of Levina & Bickel (2005) as implemented in scikit-dimension (Bac et al., 2021), with its unbiased option and no coordinate standardization. Our implementation follows LayerScope (Arcos-Holzinger et al., 2026a), which builds on GRIDS (Arcos-Holzinger et al., 2026b). Near-duplicate rows are grouped at absolute tolerance 10^{-10}. For a row requiring direct resolution, all members of its near-duplicate group are excluded, the first k positive Euclidean distances $0 < r_1 \le \cdots \le r_k$ are retained, and

$$\widehat{\mathrm{LID}}_i(k) = \frac{k-2}{\sum_{j=1}^{k} \log(r_k/r_j)}. \tag{A.1}$$

The reported cohort value is the harmonic mean of the finite positive row estimates. The raw and unresolved invalid fractions and the number of recomputed rows are stored, and a condition is rejected if more than 5% of rows remain invalid. The estimate describes finite-neighborhood expansion; it is not a claim about a unique manifold dimension.

Neighbor radius. The k-nearest-neighbor radius of row i is the Euclidean distance to its kth neighbor, excluding the row itself. Neighbors are ordered by distance and then by row index, which makes boundary ties deterministic. The retained summary contains every row's radius and its mean, median, standard deviation, minimum, maximum, and 5th, 25th, 75th, and 95th percentiles.

Effective rank of the covariance spectrum. For representations $Z \in \mathbb{R}^{n \times d}$ centered across samples, $Z_c = Z - \mathbf{1}\bar{z}^\top$ with $\bar{z} = n^{-1}\sum_i z_i$, let σ_ℓ be the singular values of Z_c, so that σ_ℓ^2/n are the eigenvalues of the covariance matrix. Following Roy & Vetterli (2007),

$$q_\ell = \frac{\sigma_\ell^2}{\sum_m \sigma_m^2}, \qquad H(q) = -\sum_\ell q_\ell \log q_\ell, \qquad r_{\mathrm{eff}} = \exp H(q). \tag{A.2}$$

Spectral entropy equals $\log r_{\mathrm{eff}}$ and is therefore not treated as a separate diagnostic. The implementation also archives participation rank, which is not a reported result.

What the exact controls fix. A one-sided rotation leaves every marginal diagnostic fixed while native retrieval collapses, and positive radial scaling changes pre-normalization marginal geometry while normalization restores the scored vectors exactly (Appendix B.4). LID, neighbor radius, and effective rank therefore describe the geometry of one modality; none of them establishes relational agreement or retrieval performance. Pair-dependent intrinsic-dimension correlation (Basile et al., 2025) consumes the pairing and is therefore not a marginal diagnostic in this sense, and neural intrinsic-dimension estimates need not recover a single underlying manifold dimension (Schulte & Rügamer, 2026).

A.3 NEIGHBORHOOD AGREEMENT: CHANCE LEVEL AND CORRECTED REPORTING

Neighborhood agreement $A_k(p)$ is defined in Equation 3. The sets $N_k^I(i)$ and $N_k^T(j)$ contain the k nearest Euclidean neighbors of image i and text unit j within their own modality, exclude the item itself, and break distance ties by index. Image j belongs to $p^{-1}(N_k^T(p(i)))$ exactly when $t_{p(j)}$ is a neighbor of $t_{p(i)}$, so

$$A_k(p) = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j \in N_k^I(i)} \mathbf{1}\left[p(j) \in N_k^T(p(i))\right]. \tag{A.3}$$

Lemma A.1 (Chance level)**.** *Fix neighbor sets $N_k^I(i)$ and $N_k^T(j)$ of size k that exclude their own index, and let p be drawn uniformly from all bijections $[n] \to [n]$. Then $\mathbb{E}[A_k(p)] = c_k = k/(n-1)$.*

Proof. Fix i and condition on $p(i) = a$. Every $j \in N_k^I(i)$ differs from i, so $p(j)$ is uniformly distributed on $[n] \setminus \{a\}$. This set contains all k elements of $N_k^T(a)$ because $a \notin N_k^T(a)$. Hence

$\Pr[p(j) \in N_k^T(a) \mid p(i) = a] = k/(n-1)$ for each of the k neighbors j, and the conditional expected inner sum in Equation A.3 is $k^2/(n-1)$. This does not depend on a, so it is also the unconditional expectation. Averaging over i and dividing by k gives $k/(n-1)$. □

The lemma concerns the unrestricted uniform distribution over all bijections, including those with fixed points. The random references used in this paper exclude fixed points in the reassigned subset and exclude replacements with identical normalized text, and their generator is not uniform (Appendix A.5). We therefore use c_k as the analytic reference and report the empirical reference distributions separately. They agree closely: for SigLIP B/16 224 at $k = 50$ and $n = 5{,}000$, $c_{50} = 0.010002$, while the mean raw A_{50} over the 256 random re-pairings of the checkpoint panel is 0.010004.

Chance-corrected agreement. Where a table reports corrected agreement, it is

$$A_k^{\mathrm{corr}}(p) = \frac{A_k(p) - c_k}{A_k(p_0) - c_k}, \tag{A.4}$$

which maps the original pairing to one and analytic chance to zero. Values are not clipped and may exceed one or fall below zero; the denominator is nonzero in every reported cohort.

Lemma A.2 (Retention is invariant to affine rescaling)**.** *Let $M' = aM + b$ with fixed $a \neq 0$ and b, applied to the original, the semantic, and every random assignment of one cohort. If the denominator of Equation 5 is nonzero, then $\rho_{M'}(q) = \rho_M(q)$.*

Proof. The random-reference mean is linear, so $\overline{M'}_{\mathrm{ref}}(q) = a\,\overline{M}_{\mathrm{ref}}(q) + b$. The numerator and the denominator of Equation 5 are each multiplied by a, and b cancels. □

Corollary A.3 (The correction leaves retention unchanged)**.** *Within a cohort, A_k^{corr} is an affine function of A_k with slope $1/(A_k(p_0) - c_k) \neq 0$. Raw and chance-corrected agreement therefore give the same retention, and the correction changes only the reporting scale. The same holds for recall expressed as a fraction or as a percentage.*

A.4 Semantic assignment construction

Feasible assignments. At replacement rate q, the feasible set $\mathcal{P}_q$ of Section 3.2 permutes text partners within a subset $\mathcal{I}_q$ of image identities and keeps all other partners fixed. Within $\mathcal{I}_q$, self-assignments and replacements whose text is identical to the original after normalization are infeasible. Text normalization case-folds and collapses whitespace. For caption groups, the comparison uses the ordered concatenation of the normalized captions (Appendix C.1). The retrieval studies use complete re-pairing, $q = 1$ and $\mathcal{I}_1 = [n]$. The CKA study uses $q \in \{0.2, 0.4, 0.6, 0.8, 1\}$, and $\mathcal{I}_q$ is the prefix of length qn of one seeded ordering of image identities. The changed subsets are therefore nested, and at each q the semantic and random assignments change the same images. Every reported assignment satisfies $F(p) = 1 - q$.

Selector vectors and solver. The external selector is MPNet (Song et al., 2020) through Sentence-BERT (Reimers & Gurevych, 2019): `sentence-transformers/all-mpnet-base-v2` at revision `e8c3b32edf54`, with maximum length 384, attention-mask mean pooling, and ℓ_2 normalization. In the retrieval studies, z_i encodes image i's selected caption. In the CKA study, z_i is the normalized average of the vectors of all captions of image i. Costs $C_{ij} = 1 - z_i^\top z_j$ are float64, and infeasible edges are excluded. SciPy's dense linear-assignment solver minimizes Equation 2 jointly over $\mathcal{I}_q$. Every retained solution is a bijection on $\mathcal{I}_q$ with no fixed point and no normalized-identical replacement, so the multiset of text units is unchanged. No evaluated representation, score, or outcome enters construction.

Two annotation-based selectors. Both annotation selectors use official COCO instance-category sets G_i and the cost $C_{ij}^{\mathrm{A}} = 1 - |G_i \cap G_j|/|G_i \cup G_j|$, with zero similarity for an empty union. They differ in how they resolve the many ties of this coarse cost (Table A.1).

- *Annotation with MPNet tie-break* (retrieval study, Appendix D). Ties are broken deterministically by MPNet distance, so this selector is not free of the text encoder. It supplies the annotation rows of Tables D.4, D.7 and D.8 and the annotation contrasts in Appendix D.6.2.

- *Annotation-only* (CKA study, Appendix C.3). No MPNet tie-break and no cost perturbation are used. Starting from numerically sorted image IDs, the source and target traversal orders are permuted independently with NumPy's `default_rng` seed 20260911, and SciPy 1.16.3's dense linear assignment solves the reordered costs. This yields one deterministic optimum at the recorded versions, not a uniform sample of optima. It is the annotation-only assignment of Section 5.1.

Table A.1: External selectors. Both annotation selectors minimize category-set Jaccard distance but resolve its ties differently, so only the CKA study's selector is free of MPNet.

Selector	Study	Cost and tie rule	Used in
MPNet (canonical)	Both	$1 - z_i^\top z_j$; exact dense optimum	Figure 2, Figure 4, Figure 5; Appendices C and D
Annotation with MPNet tie-break	Retrieval	Jaccard; ties by MPNet distance	Appendices D.3 to D.5 and D.6.2
Annotation-only	CKA	Jaccard; ties by seeded traversal order; no MPNet	Figure 2; Appendices C.3 and C.5

Alternative matchings. Three families of alternatives measure sensitivity to the particular optimum (Table A.2). None is an independent population sample. *Cost-jittered alternatives:* for each objective, the canonical optimum is accompanied by 15 alternatives. For alternative b, every finite edge receives independent noise from $\mathrm{Unif}[-10^{-4}m, 10^{-4}m]$, where m is the median absolute finite cost, and an alternative is eligible only if its unperturbed total cost lies within 1% of the canonical optimum. The implementation also adds a machine-scale target-index term. Its total is constant over complete bijections, so it does not break ties; residual ties follow the deterministic solver and the frozen row order. *Cost-matched alternatives* change at least half of the canonical assignment while matching its distribution of assigned semantic costs. They are constructed separately for the retrieval study (Appendix D.6.1) and for the CKA study (Appendix C.4).

Table A.2: Alternative semantic assignments. Overlap is the fraction of targets shared with the canonical assignment. Every family was constructed from selector costs and assignment overlap alone.

Family	Study	Count	Construction	Details
Cost-jittered	Retrieval	15 + 1	Edge noise; within 1% of optimal cost	Appendix D.6.1
Cost-jittered	CKA	15 + 1	As above; 11 of 15 change only 2 of 5,000 targets	Appendix C.2
Cost-matched	Retrieval	128	Canonical overlap ≤ 0.50; matched cost profile	Appendix D.6.1
Cost-matched	CKA	128	Canonical overlap ≤ 0.50; matched cost profile	Appendix C.4

A.5 Random references

Generator. For each seed, text units with identical normalized text are grouped; groups, and rows within groups, are ordered by seeded hash keys; and the resulting sequence of targets within $\mathcal{I}_q$ is cyclically shifted by the size of the largest group. Each draw is therefore a valid derangement of $\mathcal{I}_q$ that uses every target once and assigns no identical normalized text. The generator is not claimed to be uniform over $\mathcal{P}_q$, so its mean $\overline{M}_{\mathrm{ref}}(q)$ is a reference for this procedure rather than a uniform null (Section 3.4).

Which random assignments each result uses. Table A.3 lists every set of random assignments. Two distinctions matter for reading the tables. First, *a single random draw is not the random reference*. The core study saved one canonical random assignment (seed 20260809), and it appears as one draw in Tables D.4, D.7 and D.8 and in the rank distribution of Table D.6. Retention in Equation 5 instead uses the mean over many draws; Appendix D.3 compares the two for the checkpoint of Figure 5. Second, the 22-checkpoint panel uses one saved set of 256 random assignments (seeds 20260810–20261065) for all 22 checkpoints. These are the first 256 of the 1,000 draws used for

the core study's empirical reference distribution. The extension arms in Appendix F used their own runner's random derangements and are not pooled with the panel.

Table A.3: Random re-pairings. All draws exclude fixed points and identical normalized text within the changed subset. Only means over many draws enter retention (Equation 5).

Set	Draws	Seeds	Used for
CKA study, each q	256	seeded per q	$\overline{M}_{\mathrm{ref}}(q)$ in Figure 2; Appendix C
Core generator, $q = 1$	1,000	20260810–20261809	Empirical reference distributions for SigLIP, CLIP, and SigLIP2 (Appendix D.6)
Checkpoint panel, $q = 1$	256	20260810–20261065	$\overline{M}_{\mathrm{ref}}(1)$ for all 22 checkpoints: Figure 4; Appendix D.2
Extension runner, $q = 1$	256	from 20260810	Late-interaction and audio retention (Appendix F)
Canonical draw, $q = 1$	1	20260809	One-draw rows only; never a reference mean

B Exact controls and proofs

This appendix proves Propositions 1–3 of Section 4 in the main-text order, specifies the controls of Table 1 and their strengths, and reports the measured responses, including the CLIP and SigLIP2 replications.

B.1 Proofs of Propositions 1–3

Write the representations as matrices whose rows are samples, $I \in \mathbb{R}^{n \times d_I}$ and $T \in \mathbb{R}^{n \times d_T}$. An orthogonal transformation acts on rows, $I' = IQ_I$ and $T' = TQ_T$, with $Q_I \in O(d_I)$ and $Q_T \in O(d_T)$. Let Id be the identity matrix and $\mathbf{1}$ the all-ones vector.

Proposition 1 (Orthogonal invariance)**.** *At fixed p, applying orthogonal transformations Q_I and Q_T to I and T, respectively, preserves within-modality Euclidean distances, covariance spectra, and neighbor graphs, and hence the marginal diagnostics used here and $A_k(p)$. The distance-based RBF-CKA of Section 3.3 is also unchanged. For equal-width unit-normalized representations with $s(t, v) = \alpha\langle t, v\rangle + \beta$ and fixed $\alpha > 0$ and β, a shared rotation $Q_I = Q_T$ additionally preserves S and all candidate rankings; a one-sided rotation need not.*

Proof. We treat the image modality; the text modality is identical.

Distances and neighbor graphs. For all rows, $\|(v_i - v_j)Q_I\|_2^2 = (v_i - v_j)Q_IQ_I^\top(v_i - v_j)^\top = \|v_i - v_j\|_2^2$, so the complete matrix of within-modality distances is unchanged. The neighbor sets $N_k^I(i)$ are determined by that matrix and by the index tie rule, so they are unchanged, and so are the neighbor radii and the Levina–Bickel estimates in Equation A.1, which depend on the rows only through these distances.

Covariance spectra and effective rank. Centering commutes with right multiplication: $(IQ_I)_c = IQ_I - \mathbf{1}\bar{v}^\top Q_I = I_cQ_I$. The covariance matrix becomes $Q_I^\top \Sigma_I Q_I$, which is similar to Σ_I and has the same eigenvalues; equivalently, I_cQ_I has the same singular values as I_c. The effective rank in Equation A.2 is a function of these eigenvalues and is unchanged.

Neighborhood agreement. By Equation A.3, $A_k(p)$ depends only on the two neighbor graphs and on p, all of which are fixed.

RBF-CKA. Each RBF kernel entry is $K_{ij} = \exp(-\|v_i - v_j\|_2^2/(2h))$ with bandwidth h equal to the median nonzero squared within-modality distance. Both the entries and the bandwidth are functions of the distance matrix, so each uncentered kernel is unchanged, and so are the centered kernels K_c^I and K_c^T, their Frobenius norms, and CKA(p) in Equation 4. Linear CKA, used as a sensitivity analysis in Appendix C.5, is also unchanged because its Gram matrix $IQ_IQ_I^\top I^\top = II^\top$ is invariant.

Scores. Orthogonal maps preserve Euclidean norms, so unit-normalized rows remain unit-normalized. Let $d_I = d_T$. With $Q_I = Q_T = Q$, $\langle t_jQ, v_iQ\rangle = t_jQQ^\top v_i^\top = \langle t_j, v_i\rangle$, so every entry of S and hence every candidate ranking is unchanged. For a one-sided rotation, $S' = \alpha IQ_IQ_T^\top T^\top + \beta\mathbf{1}\mathbf{1}^\top$,

and $S' = S$ holds exactly when $I(Q_I Q_T^\top - \mathrm{Id})T^\top = 0$. If I and T have full column rank, this forces $Q_I Q_T^\top = \mathrm{Id}$, that is, $Q_I = Q_T$. A one-sided rotation with $Q_T \neq Q_I$ therefore changes S for such representations. Whether it also changes the rankings depends on the data; the measured controls in Appendix B.4 show that the tested rotation changes them for all three checkpoints. □

Proposition 2 (Fixed marginals under re-pairing)**.** *Changing only p to another bijection preserves both representation collections and every marginal diagnostic. The score matrix S and candidate rankings remain fixed; only the relevance judgments encoded by p change.*

Proof. Re-pairing replaces p_0 by another bijection p and leaves the indexed collections I and T untouched. Every marginal diagnostic is a function of I alone or of T alone (Appendix A.2), so it is unchanged. Because p is a bijection, every text unit remains the designated partner of exactly one image; the multiset of designated text units is the multiset of all text units. The score matrix $S_{ij} = s(t_j, v_i)$ depends on I, T, and s but not on p, so every row and column ranking is unchanged. What changes is which candidate is designated as relevant: $t_{p(i)}$ for image query i and $v_{p^{-1}(j)}$ for text query j. Consequently $A_k(p)$, CKA(p), and Recall@K may change. In CKA, re-pairing relabels the rows and columns of K_c^T simultaneously, which preserves its Frobenius norm, so only the numerator of Equation 4 can change. □

Proposition 3 (Positive radial cancellation)**.** *Let $t_i = \nu(\tilde{t}_i)$, where $\tilde{t}_i \neq 0$ and $\nu(x) = x/\|x\|_2$. For any $a_i > 0$, $\nu(a_i \tilde{t}_i) = \nu(\tilde{t}_i)$. Thus, at fixed I, p, and s, positive sample-wise scaling preserves T, the score matrix S, candidate rankings, and Recall@K, while raw geometry can change.*

Proof. For $a_i > 0$, $\|a_i \tilde{t}_i\|_2 = a_i \|\tilde{t}_i\|_2$, so $\nu(a_i \tilde{t}_i) = a_i \tilde{t}_i / (a_i \|\tilde{t}_i\|_2) = \nu(\tilde{t}_i)$. The normalized collection T is therefore unchanged, and so are S, every ranking, and Recall@K at fixed I, p, and s. The pre-normalization collection $\{a_i \tilde{t}_i\}$ generally has different pairwise distances, so diagnostics computed on it can change. The control with $\sigma = 1$ raises pre-normalization text LID by 45.76% (Table B.2). □

Scope. The propositions concern the distance- and spectrum-based diagnostics and the normalized dot-product scorer used here. They do not claim invariance for every statistic or scoring operator. Late-interaction MaxSim (Appendix F.2), for example, sums token-level maxima and is not covered by the shared-rotation statement.

B.2 Control construction and strengths

Table B.1 lists the prespecified strengths. Table 1 reports the maximum of each: $\theta = \pi/2$ for both rotations, $\sigma = 1$ for radial scaling, and a complete derangement ($q = 1$). The intervention seed is 20260806. No strength was chosen from outcomes.

Table B.1: Prespecified control grid. Table 1 reports the largest strength of each control.

Quantity	Values
Neighborhood scale k	$\{10, 20, 50, 100\}$; primary $k = 50$
Rotation angle θ	$\{0, \pi/16, \pi/8, \pi/4, \pi/2\}$
Radial log-scale σ	$\{0, 0.1, 0.25, 0.5, 1.0\}$
Re-paired fraction q	$\{0, 0.1, 0.25, 0.5, 1.0\}$
Retrieval cutoffs	R@$\{1, 5, 10\}$ and full ranks

- *Rotations.* The generator applies a seeded permutation to the embedding coordinates, pairs adjacent coordinates, and rotates every pair by the same Givens angle θ; an odd final coordinate is left unchanged. The shared rotation applies the same Q to images and texts. The one-sided rotation applies Q to the texts only.
- *Radial scaling.* Pre-normalization text vectors are multiplied by $a_i = \exp(\sigma \epsilon_i)$ with independent $\epsilon_i \sim \mathcal{N}(0, 1)$ and no recentering.

- *Random derangement.* At fraction q, exactly qn rows are selected by the seeded order and rearranged by one Sattolo cycle, while all other rows keep their original partners. At $q = 1$, this is a single cycle through all n rows. This control derangement is distinct from the random references of Appendix A.5. The implementation requires qn to be an integer and rejects one-row subsets.

B.3 Implementation

The controls use SigLIP B/16 224 (`google/siglip-base-patch16-224` at `7fd15f0689c7`) on the 5,000-pair COCO cohort of Appendix D.1, with its native scoring rule held fixed. The projected surface is each tower's final projected `pooler_output`; the normalized surface is the model's native `image_embeds` and `text_embeds` (Appendix A.1). Extracted arrays are float32, while metrics are accumulated in float64. Marginal diagnostics are computed on both surfaces at all four k; A_k uses exact Euclidean neighbors with self excluded and index tie-breaking; and retrieval uses the native normalized scores with candidate-index tie-breaking. Every control passed its exact gate, which allows a relative change of at most 2×10^{-4} in any quantity the propositions fix. The observed changes were far smaller: across all strengths, surfaces, and k, fixed marginal diagnostics moved by at most 3×10^{-8} in relative terms (floating-point rounding), and fixed neighborhood agreement and recall values were identical.

B.4 Measured responses and replications

Table B.2 gives the SigLIP responses behind Table 1 at full precision. The radial row is reported on both surfaces: scaling changes the pre-normalization geometry but none of the normalized quantities. Under complete random derangement, both representation collections are unchanged (Proposition 2), while A_{50} falls by 96.24% and bidirectional R@1 by 47.23 points.

Table B.2: SigLIP B/16 224 responses at the maximum strengths, $k = 50$. Marginal and A_{50} entries are relative changes, and R@1 is the change in bidirectional R@1 in percentage points; 0 denotes a relative change below 10^{-7}. Original normalized-surface values: image LID 14.995, text LID 12.383, $A_{50} = 0.2758$, R@1 47.08% (image to text) and 47.38% (text to image). Projected-surface originals: text LID 12.365, text radius 14.52, text effective rank 144.87, $A_{50} = 0.2687$.

Control	Surface	Text LID	Text radius	Text eff. rank	A_{50}	R@1 (pp)
Shared rotation	both	0	0	0	0	0
One-sided rotation	both	0	0	0	0	−47.22
Radial scaling	projected	+45.76%	+62.67%	−74.93%	−48.69%	0
Radial scaling	normalized	0	0	0	0	0
Random derangement	both	0	0	0	−96.24%	−47.23

Image-side marginal diagnostics are unchanged by every control, since none of them alters the images. The one-sided rotation reduces image-to-text R@10 from 81.70% to 0.10% and text-to-image R@10 from 81.30% to 0.06%. Under the random derangement, A_{50} falls from 0.275812 to 0.010364 on the normalized surface, close to $c_{50} = 0.010002$ (Lemma A.1).

CLIP and SigLIP2 replications. Table B.3 repeats the three representation controls for CLIP ViT-B/16 (`openai/clip-vit-base-patch16` at `57c216476eef`; Radford et al., 2021) and fixed-resolution SigLIP2 B/16 224 (`google/siglip2-base-patch16-224` at `75de2d55ec2d`; Tschannen et al., 2025), on the same cohort and at the same maximum strengths. For CLIP, the projected surface applies `visual_projection` and `text_projection` to the pooler outputs. The one-sided rotation leaves corrected agreement at exactly 1 while bidirectional R@1 falls from 34.59% to 0.00% for CLIP and from 46.88% to 0.02% for SigLIP2. Radial scaling and the CLIP shared rotation leave every normalized measurement unchanged. For SigLIP2, the shared rotation was verified at the score level: both Gram matrices and the cross-modal score matrix are reproduced to within 1.3×10^{-7} maximum absolute error, as Proposition 1 requires.

Marginal invariance under the semantic assignments. For every representative assignment, text LID at all four k, the neighbor-radius distribution, effective rank, the row-norm multiset, and the complete text-embedding multiset were unchanged on both surfaces, as Proposition 2 requires. The

Table B.3: Control replications on normalized vectors at $k = 50$. Corrected agreement follows Equation A.4; recalls are percentages. The SigLIP2 shared rotation was checked on Gram and score matrices rather than as a separately scored condition.

Checkpoint	Condition	A_{50}^{corr}	I→T R@1	I→T R@10	T→I R@1	T→I R@10
CLIP B/16 224	Original	1.0000	36.56	72.30	32.62	68.94
	Shared rotation	1.0000	36.56	72.30	32.62	68.94
	One-sided rotation	1.0000	0.00	0.12	0.00	0.08
	Radial scaling	1.0000	36.56	72.30	32.62	68.94
SigLIP2 B/16 224	Original	1.0000	45.98	79.72	47.78	81.50
	Shared rotation	scores reproduced; max. abs. error 1.3×10^{-7}				
	One-sided rotation	1.0000	0.00	0.32	0.04	0.10
	Radial scaling	1.0000	45.98	79.72	47.78	81.50

canonical MPNet and annotation target captions were also re-encoded through the pinned SigLIP processor and text tower; both surfaces matched the cached rows within the declared 5×10^{-4} tolerance. Every reported assignment had 5,000 unique targets, no self-assignment, and no identical normalized caption.

C CKA AND ORIGINAL-PAIR FIDELITY

This appendix supports Section 5.1. It gives the reconstruction protocol, the fidelity curve behind Figure 2, the annotation-only assignment, the 128 alternative matchings, the linear-CKA and row-normalized variants, and the QAP audit behind Figure 3.

C.1 RECONSTRUCTION PROTOCOL

The study reconstructs the DINOv2/All-RoBERTa COCO experiment of Maniparambil et al. (2024) with the pinned public inputs of Table C.1. The accompanying public implementation is resolved at revision `3de1963b533c`. The original release did not pin model revisions, so these revisions are our reconstruction inputs. Reconstruction used Python 3.11.2, NumPy 1.26.4, SciPy 1.16.3, PyTorch 2.6.0, Transformers 4.57.6, and Sentence Transformers 3.4.1.

Table C.1: Public inputs for the CKA study. Revision prefixes identify the full revisions in the retained manifests. Both evaluated representations have width 1,024; MPNet's is 768.

Role	Public checkpoint	Revision
Image representation	`facebook/dinov2-large` (Oquab et al., 2023)	`47b73eefe95e`
Text representation	`sentence-transformers/all-roberta-large-v1`	`ace3f8f363e3`
External selector	`sentence-transformers/all-mpnet-base-v2`	`e8c3b32edf54`

Cohort and pooling. The cohort contains the 5,000 official COCO 2017 validation images (Lin et al., 2014) in the order of the official JSON file, with all captions attached to each image: 25,014 captions in 4,987 groups of five, 12 of six, and one of seven. DINOv2 image rows average the final hidden state over the CLS and patch tokens. All-RoBERTa applies its native attention-mask mean pooling and per-caption ℓ_2 normalization, and the released protocol then averages the caption vectors of each image. The MPNet selector vectors z_i are averaged in the same way and then row-normalized. All-RoBERTa and MPNet use token limits 256 and 384. The result is one image row and one text row for each of the 5,000 images.

Caption-group exclusions. Each image's captions follow annotation order. Two caption groups count as identical when the ordered concatenations of their normalized captions agree. Groups are not sorted, and sharing one caption does not make two groups identical. This is the implemented comparison, not a test of semantic equivalence.

RBF-CKA. For the saved rows z_i of one modality, let $d_{ij} = \max(\|z_i\|^2 + \|z_j\|^2 - 2z_i^\top z_j, 0)$ and let h be the median of the nonzero d_{ij}. Then

$$K_{ij} = \exp\left(-\frac{d_{ij}}{2h}\right), \qquad H = \mathrm{Id} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad K_c = HKH, \tag{C.1}$$

and Equation 4 evaluates $\mathrm{CKA}(p)$ from K_c^I and K_c^T. The median is PyTorch's lower middle element for an even number of nonzero entries. Saved features are float32, while kernels and reductions use float64 and the equivalent double-mean centering. The bandwidths are $h = 789.516856$ for images and $h = 1.263683$ for text. A float32 literal-source check agrees within 5×10^{-5}. No per-feature scaling is applied here; the QAP procedure of Appendix C.6 does apply it.

Partial re-pairing. At replacement rate q, the changed subset $\mathcal{I}_q$ is the prefix of one seeded ordering of image identities (Appendix A.4). The semantic assignment is re-solved inside $\mathcal{I}_q$, and 256 random re-pairings are drawn inside the same subset, so both conditions change the same images and have $F(p) = 1 - q$.

Ordinary-shuffle replication. To compare with the published curve, one ordinary seeded partial shuffle per level reproduces the source procedure. It may retain fixed points, and the original shuffle code and seed were not released, so this series is a protocol replication rather than an exact reconstruction of the published permutations. JSON order is our reconstruction choice; the released extraction script's TorchVision loader sorts image IDs. Consistent relabeling preserves CKA, but seeded subsets depend on row order, so we do not claim the authors' exact sampled assignments.

C.2 The fidelity curve

Table C.2 gives the values plotted in Figure 2. At $F(p) = 0.8$, hard negative re-pairing gives 0.700670, against 0.462279 for the random mean. With no original pairs left, hard negative re-pairing still gives 0.659390, which exceeds the random mean at $F(p) = 0.8$ and also its largest draw, 0.471440. The original pairing scores highest (0.719270).

Table C.2: RBF-CKA against original-pair fidelity (Figure 2). "Ordinary shuffle" is our replication of the published procedure, which may retain fixed points. The hard negative and random conditions change the same images at each fidelity level. MC SE is the Monte Carlo standard error of the 256-draw random mean.

$F(p)$	Published	Ordinary shuffle	Hard negative	Random mean	MC SE
1.0	0.720	0.719270	—	0.719270	—
0.8	0.460	0.468954	0.700670	0.462279	0.000209
0.6	0.270	0.262519	0.689973	0.263955	0.000232
0.4	0.130	0.131983	0.679331	0.123607	0.000178
0.2	0.040	0.041178	0.669915	0.039682	0.000112
0.0	0.010	0.012628	0.659390	0.012945	0.000018

At complete re-pairing, retention (Equation 5) is $(0.659390 - 0.012945)/(0.719270 - 0.012945) = 0.915223$. The canonical assignment and its 15 cost-jittered alternatives span 0.913111–0.915223. Four of these alternatives differ from the canonical assignment at 6.02–6.70% of targets, and the other 11 change only two of 5,000 targets, so they are not treated as allocation-diverse; Appendix C.4 constructs substantially different alternatives. Five cohorts that each omit a different 20% of images, with both kernels rebuilt and the assignment re-solved, give retention 0.909106–0.916212.

C.3 Annotation-only assignment and kernel response

The annotation-only assignment (Appendix A.4) replaces every caption group and was saved before its CKA was evaluated. Its CKA is 0.546883, above the random mean (0.462279) and the largest random draw (0.471440) at $F(p) = 0.8$; its retention is 0.755939. It uses COCO object annotations and no text encoder, so the ordering reversal of Section 5.1 does not depend on MPNet. The check is motivated by the documented relationship between the two text encoders: `all-roberta-large-v1` (Sentence-Transformers contributors, 2021) and `all-mpnet-base-v2` are Sentence-Transformers models

whose model cards describe the same training project and overlapping training data. This follow-up was specified after the original results.

To describe how much each assignment changes the evaluated text geometry, let $L = K_c^T$ be the original centered text kernel and L_p its simultaneous row and column relabeling under p. We report

$$\delta_L(p) = \frac{\|L - L_p\|_F}{\|L\|_F}, \qquad \delta_L(p)^2 = 2\left(1 - \frac{\langle L, L_p\rangle_F}{\|L\|_F^2}\right), \tag{C.2}$$

where the identity uses that permutation preserves the Frobenius norm. By Cauchy–Schwarz, $|\mathrm{CKA}(p) - \mathrm{CKA}(p_0)| \le \delta_L(p)$. This bound describes sensitivity; it does not decompose the observed CKA difference.

Table C.3: Selector sensitivity and text-kernel response. Random rows average the 256 random re-pairings at that fidelity. Substitute cosine is the mean cosine similarity between each image's original and assigned caption-group vectors (normalized All-RoBERTa rows), averaged over all 5,000 rows, including unchanged rows at $F(p) = 0.8$.

Assignment	$F(p)$	CKA	Substitute cosine	δ_L
Original	1.0	0.719270	1.000000	0.000000
Hard negative (MPNet)	0.8	0.700670	0.956324	0.282460
Hard negative (MPNet)	0.0	0.659390	0.830577	0.508814
Annotation-only	0.0	0.546883	0.616015	0.864871
Random mean	0.8	0.462279	0.835758	0.846225
Random mean	0.0	0.012945	0.175648	1.406548

At $F(p) = 0$, the MPNet assignment preserves more of the text geometry than the annotation-only assignment, which preserves more than random re-pairing. Kernel change alone does not order CKA across conditions: the annotation-only assignment changes the text kernel slightly more than random re-pairing at $F(p) = 0.8$, yet it gives the higher CKA. These measurements are consistent with CKA responding to how similar the replacement partners are, but they do not establish a causal mechanism.

C.4 128 ALTERNATIVE MATCHINGS

This check asks whether the reversal depends on the particular MPNet optimum. Before any new assignment was evaluated, we specified a construction that uses only the MPNet caption-group vectors, the same row order, and the same forbidden edges. No DINOv2 or All-RoBERTa value enters construction. The check was specified after the original results.

Exact linear assignments with increasing penalties on canonical edges bound the cost of leaving the canonical assignment. At maximum canonical overlap 0.50, the Lagrangian lower and feasible upper bounds on the minimum semantic-cost ratio are 1.006812 and 1.006839, so a complete re-pairing that changes at least half of the canonical matches exists within the primary 1% cost tolerance. The predeclared 2% and 5% fallback tolerances were not needed.

Each accepted assignment must be a complete derangement with the caption-group exclusions, share at most 50% of targets with the canonical assignment and at most 80% with every other accepted assignment, have mean semantic cost at most 1.01 times the canonical mean, and match the canonical distribution of assigned costs: Wasserstein distance at most 0.02, and largest decile difference at most 0.05, times the canonical mean cost. The search screens 16 proposals per penalty/noise setting and extends a setting to at most 256 proposals once eight are accepted; the overall cap is 1,024. The first three settings accepted 1, 1, and 3 of 16 proposals and stopped. The fourth accepted 128 of 173, for 221 proposals in total. It uses the smallest frontier penalty that attains 0.50 overlap, plus independent uniform edge noise with half-width 0.04 times the canonical mean cost. All assignments and gate diagnostics were saved before any new CKA was computed.

The canonical CKA of 0.659390 lies near the middle of the family: 60 of the 128 alternatives score no higher. Every alternative exceeds 0.471440, the largest of the 256 random re-pairings that keep 80% of original pairs. The reversal therefore holds across substantially different allocations with the same cost profile.

Table C.4: The 128 alternative matchings on the CKA surface. Ranges are over the constructed family and describe sensitivity to the allocation; they are not population intervals. Cost ratios use the canonical mean semantic cost.

Quantity	Value or range
Original pairs retained	0 in every assignment
Canonical matches changed	50.00–52.00%
Largest overlap between two alternatives	74.12%
Mean semantic-cost ratio	1.008986–1.009997
Wasserstein distance / canonical mean cost	0.010527–0.013764
Largest decile difference / canonical mean cost	0.015052–0.037391
RBF-CKA minimum / median / maximum	0.658094 / 0.659450 / 0.661051
Retention minimum / median / maximum	0.913389 / 0.915309 / 0.917575
Alternatives above the random mean at $F(p) = 0.8$	128 / 128
Alternatives above every random draw at $F(p) = 0.8$	128 / 128

C.5 Linear CKA and row-normalized representations

The fidelity-matched comparison. The comparison in Section 5.1 pits re-pairing with no original pairs against the random reference that keeps 80% of them. Table C.5 repeats it under linear CKA, with biased centered Gram matrices, and on row-normalized representations, in which every saved row is ℓ_2-normalized before either kernel is built. All variants evaluate the same saved assignments: the semantic and random assignments at every replacement rate, the annotation-only assignment, and the 128 alternatives. This analysis was run on 2026-09-25, after the original results; it constructs and selects no assignment. Before writing any output, the runner reproduced the original anchors, the 128 alternative values, the annotation-only value, and the six complete-reassignment cells of Table C.6 to within 8×10^{-16}.

Table C.5: The fidelity reversal under linear CKA and row normalization. Every column evaluates the same saved assignments. Under every variant, hard negative re-pairing with no original pairs (bold), the annotation-only assignment, and all 128 cost-matched alternatives exceed both the mean and the largest of the 256 random re-pairings that keep 80% of original pairs. The first column reproduces Figure 2.

	Saved rows		Row-normalized	
Condition	RBF-CKA	Linear CKA	RBF-CKA	Linear CKA
Original pairing, $F(p) = 1$	0.7193	0.7030	0.7184	0.7038
Hard negative, $F(p) = 0.8$	0.7007	0.6847	0.6995	0.6848
Random mean, $F(p) = 0.8$	0.4623	0.4503	0.4623	0.4519
Largest random draw, $F(p) = 0.8$	0.4714	0.4601	0.4716	0.4617
Hard negative, $F(p) = 0$	**0.6594**	**0.6439**	**0.6569**	**0.6415**
Annotation-only, $F(p) = 0$	0.5469	0.5351	0.5440	0.5317
128 alternatives, $F(p) = 0$	0.658–0.661	0.642–0.645	0.656–0.659	0.640–0.643
Random mean, $F(p) = 0$	0.0129	0.0089	0.0128	0.0089

Under every variant, hard negative re-pairing at $F(p) = 0$ exceeds the random mean at $F(p) = 0.8$ by 0.190–0.197, and it also exceeds the largest of those 256 random draws. The annotation-only assignment exceeds that largest draw by 0.070–0.075, and all 128 alternatives exceed it. The curves keep the same shape at the intermediate rates: under each variant, hard negative re-pairing at every fidelity level scores above the random mean at $F(p) = 0.8$, and the random mean decreases monotonically as fidelity decreases. The artifact retains every variant's complete curve and all random draws.

Complete re-pairing with preprocessing matched across diagnostics. Table C.6 reports retention at complete re-pairing when all three relational diagnostics consume identical arrays on each surface, with the same semantic assignment and the same 256 random re-pairings. Neighborhood agreement

A_{50} uses exact Euclidean neighbors on the declared surface, excludes self, and breaks distance ties by index.

Table C.6: Complete re-pairing, $F(p) = 0$, with preprocessing matched within each surface. Retention uses the mean of the 256 random re-pairings. The two surfaces are reported separately.

Surface	Diagnostic	Original	Hard negative	Random mean	Retention
Saved rows	RBF-CKA	0.719270	0.659390	0.012945	0.915223
	Linear CKA	0.702959	0.643851	0.008894	0.914838
	A_{50}	0.370848	0.315332	0.009989	0.846156
Row-normalized	RBF-CKA	0.718416	0.656919	0.012772	0.912849
	Linear CKA	0.703807	0.641547	0.008896	0.910406
	A_{50}	0.400760	0.338592	0.009999	0.840905

High CKA retention persists on both surfaces and under both kernels. Neighborhood agreement retains somewhat less of its contrast, 84.6% on saved rows and 84.1% after normalization.

C.6 QAP AUDIT

Problem setup. Maniparambil et al. (2024) match captions to images by maximizing CKA over one-to-one assignments, a quadratic assignment problem (QAP). We retain their caption-to-image setup. The 320 anchors are the images closest in cosine similarity to the centers of a 320-cluster K-means partition of the DINOv2 image rows (the source defaults, `random_state`=0), and their original pairs are held fixed in all three problems. Each problem must then recover the partners of 500 non-anchor query captions among 500 query images, 820 nodes in total. Query sets are drawn with seeds 0, 1, and 2, and the image order within the query block is shuffled with seed 0. As in the released code, every feature is divided by its standard deviation before the kernels are built. Each modality's kernel A or B is the median-bandwidth RBF kernel of Equation C.1, double-centered and divided by its Frobenius norm. For a permutation matrix P the objective is

$$J(P) = \langle A, PBP^\top \rangle_F, \tag{C.3}$$

which is the CKA of the matching P over all 820 nodes. Because of the feature scaling, J is not numerically identical to the saved-row CKA of Figure 2. The problems use our frozen JSON-ordered rows (Appendix C.1); they do not reconstruct the authors' sampled anchors and queries.

Solver. We use the Fast Approximate QAP (FAQ) algorithm (Vogelstein et al., 2015) as implemented by SciPy's `quadratic_assignment` with `method="faq"`, maximization, the 320 anchors fixed through `partial_match`, randomized doubly stochastic initialization, and tolerance 10^{-5}. Every start receives an explicit seed, and all 500 starts of each problem enter the summaries. An accuracy-blind 32-start feasibility check passed the frozen diversity criteria before the 500-start study.

Outcomes. Fidelity counts exact recoveries among the 500 query partners; the 320 anchors are excluded. Table C.7 gives every problem's summary, and Figure 3 plots all 1,500 outputs. Means and rank correlations weight every start equally, including repeated output permutations. Selecting the highest-CKA output gives fidelity 0.25 and 0.73 percentage points below, and 0.41 points above, the mean over the 500 runs. That mean equals the expected fidelity of choosing a run uniformly at random.

These results do not show that the objective is worse than random run selection, nor that it carries no information about exact recovery. They show that the highest-CKA output and the highest-fidelity output differ in every problem: another output from the same runs recovers six, six, and seven more partners. The highest-fidelity output is identified with ground-truth labels for this audit only and is not proposed as a matching rule.

C.7 VALIDATION

The pipeline reconstructs all assignments, both kernels, the random means, and the reported ratios. A feature-space identity recovers the complete-re-pairing linear CKA to 4.44×10^{-16}, and direct

Table C.7: All 500 randomized FAQ runs per query set. Fidelity is exact recovery over 500 queries. "Top-CKA" is the fidelity of the highest-objective output, "Best" the highest fidelity among outputs, and "Gap" their difference in percentage points. Correlations are between J and fidelity over all 500 runs.

Seed	Unique	J range	Mean fid.	Top-CKA	Best	Gap	Spearman	Kendall
0	70	2.279×10^{-4}	56.85%	56.6%	57.8%	1.2	−0.128	−0.120
1	26	2.617×10^{-4}	65.13%	64.4%	65.6%	1.2	0.138	0.102
2	167	6.246×10^{-4}	60.59%	61.0%	62.4%	1.4	0.446	0.321

neighbor-set intersection recovers A_{50} exactly. A small-array literal-centering implementation checks RBF-CKA on both surfaces. The construction records reproduce all ensemble costs and the exclusion and overlap gates. All 1,500 QAP outputs satisfy the anchor constraints, their objectives reconstruct directly from Equation C.3, and an independent inverse-permutation calculation reproduces every reported fidelity. Artifacts A12, A13, A16, A17, A19, and A22 (Table G.2) hold the machine-readable records.

D Relational agreement and assigned-pair retrieval

This appendix supports Section 5.2. It gives the protocol for the 22 checkpoints, their complete results, the representative checkpoints and the random reference behind Figure 5, assigned ranks conditional on original top-1 success, qualitative examples, and the sensitivity analyses, with the checkpoints each analysis covers.

D.1 Protocol

Cohort and caption selection. The cohort contains 5,000 unique COCO validation images (Lin et al., 2014) from `lmms-lab/COCO-Caption2017` at revision `3bdd5827e243`. Image identities are ranked by a seeded stable hash (seed 20260804), and one caption per selected image is then chosen by a separate seeded hash (seed 20260805). Neither step observes a representation, an annotation, or an outcome. The study therefore uses one outcome-independent caption realization, and its magnitudes are conditional on it.

One designated partner per query. Each image query has exactly one relevant caption and each caption query exactly one relevant image, in galleries of 5,000 images and 5,000 captions. Standard COCO evaluation instead counts any of an image's five captions as relevant, so absolute recalls here are not comparable with standard benchmark numbers.

Checkpoints and preprocessing. Table D.1 lists the 22 checkpoints. The 16 CLIP, SigLIP, and SigLIP2 checkpoints (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) are loaded with their `AutoProcessor` at the pinned revision. Images are converted to RGB and passed through the processor's resize, crop, and normalization configuration; text uses max-length padding and truncation under the versioned tokenizer. The scored vectors are the native normalized `image_embeds` and `text_embeds`. SigLIP So400m/14 224 has a published 16-token text limit and truncates 762 of the 5,000 captions (15.24%); no other checkpoint in the panel truncates a caption.

The six additional encoders use their native processors, with no common manual resize: 512-pixel inputs and a 1,024-token processing length for Jina-CLIP-v2 (Koukounas et al., 2024), under which no caption is truncated; 224-pixel center crops and a 248-token window for both LongCLIP variants (Zhang et al., 2024); any-resolution 336-pixel tiling for E5-V (Jiang et al., 2024); and a dynamic budget of 4,096–1,310,720 pixels with the library's query path for Qwen3-VL-Embedding-2B and -8B (Li et al., 2026). E5-V follows its published templates: the caption, or `<image>`, is followed by a newline and `Summary above sentence in one word:` or `Summary above image in one word:`, respectively. Every output is explicitly ℓ_2-normalized; the Jina and Qwen outputs are renormalized in float64 because their loader outputs did not meet a strict unit-norm tolerance.

Table D.1: The 22 global image–text checkpoints of Section 5.2 at their evaluated revisions. d is the width of the scored vectors. For the 16 CLIP, SigLIP, and SigLIP2 checkpoints, the image column gives the native input resolution and the text column the native token limit; the six additional encoders use the native processors described in Appendix D.1. Full revisions are retained with each record.

Checkpoint	Identifier	Revision	d	Image	Text
CLIP B/32 224	`openai/clip-vit-base-patch32`	`3d74acf9a28c`	512	224	77
CLIP B/16 224	`openai/clip-vit-base-patch16`	`57c216476eef`	512	224	77
CLIP L/14 224	`openai/clip-vit-large-patch14`	`32bd64288804`	768	224	77
CLIP L/14 336	`openai/clip-vit-large-patch14-336`	`ce19dc912ca5`	768	336	77
SigLIP B/16 224	`google/siglip-base-patch16-224`	`7fd15f0689c7`	768	224	64
SigLIP B/16 256	`google/siglip-base-patch16-256`	`b078df89e446`	768	256	64
SigLIP B/16 384	`google/siglip-base-patch16-384`	`41aec1c83b32`	768	384	64
SigLIP L/16 256	`google/siglip-large-patch16-256`	`d0da9f876e7d`	1024	256	64
SigLIP L/16 384	`google/siglip-large-patch16-384`	`ce005573a409`	1024	384	64
SigLIP So400m/14 224	`google/siglip-so400m-patch14-224`	`d04cf29fca7b`	1152	224	16
SigLIP2 B/16 224	`google/siglip2-base-patch16-224`	`75de2d55ec2d`	768	224	64
SigLIP2 B/16 256	`google/siglip2-base-patch16-256`	`3f9f96cb90da`	768	256	64
SigLIP2 B/16 384	`google/siglip2-base-patch16-384`	`f775b65a7976`	768	384	64
SigLIP2 L/16 256	`google/siglip2-large-patch16-256`	`787800c8990e`	1024	256	64
SigLIP2 L/16 384	`google/siglip2-large-patch16-384`	`1b426889ea62`	1024	384	64
SigLIP2 So400m/14 224	`google/siglip2-so400m-patch14-224`	`78e403963a4f`	1152	224	64
E5-V	`royokong/e5-v`	`684c4c91ebab`	4096	native	native
Jina-CLIP-v2	`jinaai/jina-clip-v2`	`e10d47f5691d`	1024	native	native
LongCLIP-GmP ViT-L/14	`zer0int/LongCLIP-GmP-ViT-L-14`	`b476fa98520c`	768	native	native
LongCLIP-L (Diffusers)	`zer0int/LongCLIP-L-Diffusers`	`89090388cbab`	768	native	native
Qwen3-VL-Embedding-2B	`Qwen/Qwen3-VL-Embedding-2B`	`9f2f7e710d6d`	2048	native	native
Qwen3-VL-Embedding-8B	`Qwen/Qwen3-VL-Embedding-8B`	`2c4565515e0f`	4096	native	native

Extraction of the six additional encoders. All six were extracted afresh on 2026-09-24 at the revisions of Table D.1, with PyTorch 2.6.0, Sentence Transformers 5.4.0, and Transformers 4.57.6 (Jina, both LongCLIPs) or 5.5.0 (E5-V, both Qwens). Before the full cohort, each model was checked on the first 16 rows: outputs had to be finite and normalized, reach bidirectional R@1 of at least 0.25 on that small gallery, and reproduce under a repeated extraction with the same batches to within 5×10^{-4}; Qwen's cosine scorer also had to agree with the normalized dot products to within 5×10^{-5}. These are setup checks, not results. The fresh original-pair recalls differ from the earlier extension records by at most 0.16 percentage points.

Scores, neighbors, and retention. Scores are float32 dot products of the normalized vectors, and ties are broken by candidate index. Neighbor sets use exact Euclidean distances on the normalized vectors, exclude the item itself, and break ties by index. The primary scale is $k = 50$; $k \in \{10, 20, 100\}$ are secondary. All 22 checkpoints use one saved array of assignments: the identity, the canonical MPNet assignment, its 15 alternatives, and 256 random re-pairings (Appendix A.5). Retention follows Equation 5 with $\overline{M}_{\mathrm{ref}}(1)$ the mean over those 256 re-pairings; for bidirectional recall, the two directions are averaged within each condition first.

Protocol status. The 16-checkpoint panel's roster and analysis rules were fixed before its outcomes were inspected; three of its rows were already known from the representative studies. The six-encoder analysis of neighborhood agreement was specified after their retrieval outcomes were known, and its cohort, assignments, graph scales, and reporting rule were fixed before their agreement outcomes were inspected (Appendix G.1).

D.2 Per-checkpoint results for all 22 checkpoints

Table D.2 gives every checkpoint's original values and its three retentions, the values plotted in Figure 4. All 22 satisfy $\rho_{A_{50}} > \rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$. Neighborhood agreement keeps 71.9–81.6% of its original-minus-random gap, bidirectional R@10 keeps 46.0–53.1%, and bidirectional R@1 keeps 7.3–14.9%. Every retention uses the same 256 random re-pairings, so the rows are directly comparable.

Table D.2: Complete results for the 22 checkpoints under complete hard negative re-pairing ($q = 1$), the values plotted in Figure 4. Original A_{50} and bidirectional recalls (%) are original-pairing values; each ρ is Equation 5 with the mean of the same 256 random re-pairings for every checkpoint. All 22 rows satisfy $\rho_{A_{50}} > \rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$. Ranges: $\rho_{A_{50}}$ 0.7193–0.8160, $\rho_{\mathrm{R@10}}$ 0.4601–0.5313, $\rho_{\mathrm{R@1}}$ 0.0732–0.1485.

	Neighborhood agreement		Bidirectional R@10		Bidirectional R@1	
Checkpoint	$A_{50}(p_0)$	$\rho_{A_{50}}$	p_0 (%)	$\rho_{\mathrm{R@10}}$	p_0 (%)	$\rho_{\mathrm{R@1}}$
CLIP B/32 224	0.2742	0.8081	69.12	0.4704	32.04	0.1485
CLIP B/16 224	0.2779	0.8091	70.62	0.4876	34.59	0.1306
CLIP L/14 224	0.2835	0.8116	72.19	0.4838	37.22	0.1262
CLIP L/14 336	0.2902	0.8160	73.11	0.4834	38.30	0.1232
SigLIP B/16 224	0.2758	0.7787	81.50	0.4709	47.23	0.0959
SigLIP B/16 256	0.2794	0.7825	81.70	0.4710	47.30	0.0877
SigLIP B/16 384	0.2764	0.7804	82.63	0.4765	49.17	0.0938
SigLIP L/16 256	0.2564	0.7699	83.75	0.4744	51.10	0.0855
SigLIP L/16 384	0.2495	0.7668	84.19	0.4802	52.87	0.0842
SigLIP So400m/14 224	0.2392	0.7795	83.05	0.4754	50.81	0.0852
SigLIP2 B/16 224	0.2868	0.7833	80.61	0.4762	46.88	0.0962
SigLIP2 B/16 256	0.2878	0.7864	80.90	0.4770	47.85	0.0980
SigLIP2 B/16 384	0.2874	0.7859	82.59	0.4747	49.57	0.0950
SigLIP2 L/16 256	0.2719	0.7666	82.47	0.4662	50.16	0.0869
SigLIP2 L/16 384	0.2704	0.7692	82.95	0.4601	50.68	0.0856
SigLIP2 So400m/14 224	0.2671	0.7656	82.35	0.4619	49.63	0.0848
E5-V	0.4868	0.7422	82.99	0.5311	48.52	0.1099
Jina-CLIP-v2	0.3876	0.7629	78.44	0.5207	42.40	0.1231
LongCLIP-GmP ViT-L/14	0.4822	0.7449	83.02	0.5313	48.72	0.1047
LongCLIP-L (Diffusers)	0.3458	0.7971	76.83	0.5062	42.74	0.1144
Qwen3-VL-Embedding-2B	0.4135	0.7344	84.23	0.5017	50.93	0.0831
Qwen3-VL-Embedding-8B	0.4017	0.7193	86.74	0.4905	54.83	0.0732

The six additional encoders before and after the follow-up. The six additional encoders were first evaluated for retrieval only, in an extension run that drew its random derangements with its own generator (rejection sampling, seeded from 20260810) and resolved score ties optimistically. The follow-up recomputed all three diagnostics under the panel's conventions, including the stable tie rule (Table D.3). Retentions change by at most 0.004 for R@10 and 0.003 for R@1. In particular, the largest R@10 retention among the six is 0.5324 (E5-V) in the earlier run and 0.5313 (LongCLIP-GmP) in the follow-up; Figure 4 uses the follow-up values, so its range ends at 0.531. Only the follow-up values share the panel's random reference and are reported alongside the 16 other checkpoints.

Table D.3: The six additional encoders before and after the follow-up. The earlier extension run measured retrieval only, with its own random derangements and optimistic handling of score ties. The follow-up re-extracted all six from the recorded revisions and recomputed all three diagnostics with the panel's saved assignments and stable tie rule. Recalls are bidirectional percentages; the follow-up values are those of Table D.2 and Figure 4.

	Earlier extension run			Follow-up (panel reference)			
Encoder	R@10	$\rho_{\mathrm{R@10}}$	$\rho_{\mathrm{R@1}}$	R@10	$\rho_{\mathrm{R@10}}$	$\rho_{\mathrm{R@1}}$	$\rho_{A_{50}}$
E5-V	82.97	0.5324	0.1126	82.99	0.5311	0.1099	0.7422
Jina-CLIP-v2	78.60	0.5167	0.1230	78.44	0.5207	0.1231	0.7629
LongCLIP-GmP ViT-L/14	83.08	0.5287	0.1034	83.02	0.5313	0.1047	0.7449
LongCLIP-L (Diffusers)	76.86	0.5050	0.1157	76.83	0.5062	0.1144	0.7971
Qwen3-VL-Embedding-2B	84.17	0.5008	0.0837	84.23	0.5017	0.0831	0.7344
Qwen3-VL-Embedding-8B	86.73	0.4917	0.0739	86.74	0.4905	0.0732	0.7193

D.3 Representative checkpoints and the random reference

Table D.4 reports image-to-text values for the three representative checkpoints. It separates two quantities that must not be conflated: one random re-pairing (the canonical draw, seed 20260809) and the random reference $\overline{M}_{\mathrm{ref}}$, the mean over the 256 panel draws that enters Equation 5. For SigLIP B/16 224, image-to-text mean reciprocal rank is 0.1536 under the MPNet assignment, 0.1013 under the annotation assignment, and 0.0015 under the one-draw random assignment. For CLIP, it is 0.4855, 0.1454, and 0.0015 for the original, MPNet, and one-draw conditions, and text-to-image R@10 is 68.94%, 33.84%, and 0.08%. For SigLIP2, text-to-image R@1 is 47.78%, 5.00%, and 0.00%, and text-to-image R@10 is 81.50%, 39.70%, and 0.10%. Recall and reciprocal rank always evaluate the designated partner, not the original one. The CLIP one-draw image-to-text R@10 of 0.14% lies below the prespecified screening reference of 0.52%, analytic chance $10/5000$ plus five binomial standard errors; because globally constrained assignments do not give independent query outcomes, this is a heuristic screen rather than a calibrated test.

Single draw and random reference for Figure 5. Figure 5 shows SigLIP B/16 224 image-to-text values, and its brackets are the retentions of Equation 5. Table D.5 places the canonical single draw next to $\overline{M}_{\mathrm{ref}}$. For R@10, the single draw (0.06%) lies at the 1st percentile of the core generator's 1,000 draws, whose mean is 0.18%, while the 256-draw panel mean is 0.19%. The random reference is 0.0100 for A_{50}, 0.19% for R@10, and 0.01% for R@1. Retention computed against either the single draw or $\overline{M}_{\mathrm{ref}}$ rounds to the displayed 0.78, 0.47, and 0.09, because every random value is small relative to its original-minus-random gap.

D.4 Assigned ranks conditional on original top-1 success

With scores fixed, the original partner stays in the gallery. Under a complete derangement with deterministic tie handling, a query's top-ranked candidate can be its original partner or its designated partner, but not both, so $\mathrm{R@1}(p_0) + \mathrm{R@1}(p) \leq 1$. For SigLIP B/16 224 image-to-text retrieval, original R@1 is 47.08%, so this bound allows assigned R@1 up to 52.92%; it does not force the observed 4.10%. To remove the constraint, we condition on whether the original partner ranks first, using the saved single-caption representations and assignments (Table D.7). The analysis was specified after the original results and before these summaries were computed.

Table D.4: Representative checkpoints, image-to-text, $k = 50$, normalized vectors. The one-draw rows show the single canonical random re-pairing (seed 20260809). The reference rows are the mean over the 256 panel draws, the $\overline{M}_{\mathrm{ref}}$ of Equation 5. The annotation row uses the MPNet tie-break (Table A.1). Recalls are percentages.

Checkpoint	Condition	A_{50}	A_{50}^{corr}	R@10	R@1
SigLIP B/16 224	Original pairing	0.275812	1.0000	81.70	47.08
	Hard negative (MPNet)	**0.216984**	**0.7787**	**38.44**	**4.10**
	Hard negative (annotation)	0.182965	0.6507	24.32	2.74
	Random, one draw	0.010508	0.0019	0.06	0.00
	Random reference $\overline{M}_{\mathrm{ref}}$	0.010004	0.0000	0.19	0.01
SigLIP2 B/16 224	Original pairing	0.286752	1.0000	79.72	45.98
	Hard negative (MPNet)	**0.226776**	**0.7833**	**37.26**	**4.04**
	Random, one draw	0.010220	0.0008	0.20	0.00
	Random reference $\overline{M}_{\mathrm{ref}}$	0.010012	0.0000	0.18	0.01
CLIP B/16 224	Original pairing	0.277868	1.0000	72.30	36.56
	Hard negative (MPNet)	**0.226720**	**0.8091**	**35.22**	**4.26**
	Random, one draw	0.010268	0.0010	0.14	0.00
	Random reference $\overline{M}_{\mathrm{ref}}$	0.009995	0.0000	0.19	0.01

Table D.5: Single random draw and random reference for SigLIP B/16 224, image-to-text, the checkpoint and direction of Figure 5. The random reference of Equation 5 is the mean over the 256 panel draws. Retention against either rounds to the values displayed in Figure 5: 0.78, 0.47, and 0.09.

			Random		Retention against	
Diagnostic	Original	Hard negative	One draw	$\overline{M}_{\mathrm{ref}}$ (256)	One draw	$\overline{M}_{\mathrm{ref}}$
A_{50}	0.2758	0.2170	0.0105	0.0100	0.778	0.779
R@10 (%)	81.70	38.44	0.06	0.19	0.470	0.469
R@1 (%)	47.08	4.10	0.00	0.01	0.087	0.087

Table D.6: Assigned-pair Recall@K (%) for SigLIP B/16 224, image-to-text, over the full 5,000-caption gallery. The score matrix and every candidate ranking are identical in all four rows; only the designated partner changes. Both hard negative assignments lie between the original pairing and one random re-pairing at every cutoff.

Condition	$K = 1$	5	10	20	50	100	250	500	1,000
Original pairing	47.08	72.14	81.70	88.74	96.26	99.00	99.76	99.92	99.96
Hard negative (MPNet)	4.10	25.82	38.44	52.18	71.82	83.32	91.60	95.52	97.94
Hard negative (annotation)	2.74	15.68	24.32	36.04	54.70	67.78	79.56	86.76	91.86
Random, one draw	0.00	0.04	0.06	0.42	0.96	2.06	4.98	9.88	20.10

Table D.7: SigLIP B/16 224 assigned ranks conditional on original top-1 success. Recalls are percentages; median ranks are out of 5,000. "Random" is the canonical single draw, not the reference mean. Each direction conditions on its own original ranking.

Direction	Assignment	Original top-1	Queries	R@1	R@10	Median rank
Image → text	Hard negative (MPNet)	yes	2,354	0.00	43.76	14
	Hard negative (MPNet)	no	2,646	**7.75**	33.71	23
	Hard negative (annotation)	yes	2,354	0.00	26.30	40
	Hard negative (annotation)	no	2,646	5.18	22.56	41
	Random, one draw	yes	2,354	0.00	0.00	2,547.5
	Random, one draw	no	2,646	0.00	0.11	2,501.5
Text → image	Hard negative (MPNet)	yes	2,369	0.00	38.75	17
	Hard negative (MPNet)	no	2,631	**9.46**	38.31	18
	Hard negative (annotation)	yes	2,369	0.00	25.75	39
	Hard negative (annotation)	no	2,631	5.47	24.17	40
	Random, one draw	yes	2,369	0.00	0.17	2,521
	Random, one draw	no	2,631	0.04	0.08	2,590

Among queries whose original partner is not ranked first, assigned R@1 under the MPNet assignment is 7.75% image-to-text and 9.46% text-to-image, with R@10 of 33.71% and 38.31%. Low assigned-pair R@1 therefore persists outside the stratum that the constraint forces to fail. This is not a test of whether the replacement captions are acceptable descriptions. Every saved rank is recovered exactly from the cached score matrix, and the weighted conditional summaries recover the pooled values. For text queries, the saved rank at source index i belongs to query $p(i)$ and is inverse-permuted before conditioning.

D.5 Qualitative examples and object overlap

Figure D.1 shows the first three rows of the frozen cohort order. The display rule uses no MPNet distance, rank, agreement, or other outcome, so the examples are not selected for visual success. They illustrate how a hard negative caption can preserve an object or scene while belonging to a different image.

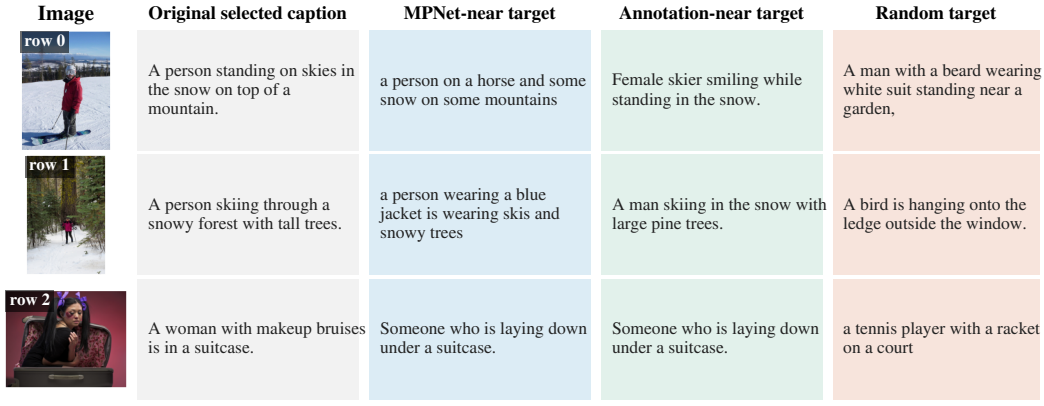


Figure D.1: The first three rows of the frozen cohort order. From left to right: the image, its original caption, the captions assigned to it by the MPNet and the annotation hard negative assignments, and the caption assigned by one random re-pairing. Hard negative captions can preserve a scene, object, or action while describing a different image; the displayed random captions are unrelated. Every assignment is a complete bijection, and every displayed caption belongs to a different image. The figure illustrates the change of designated partner, not human judgments of caption validity.

We also compare official COCO instance-category sets between each source image and the image whose caption it is assigned (Table D.8). This check uses the three single-caption assignments of the retrieval study, including the annotation assignment with its MPNet tie-break; it does not use the

CKA study's annotation-only assignment. An empty union has Jaccard zero, and 48 of the 5,000 source images have an empty category set.

Table D.8: Object-category overlap under complete re-pairing. The random row is the canonical single draw. Sharing object categories, or even an identical category set, does not make a caption an acceptable description.

Assignment	Mean Jaccard	Any shared	Identical nonempty	Empty unions
Hard negative (MPNet)	0.496039	91.84%	20.86%	20
Hard negative (annotation)	0.859223	99.04%	63.10%	48
Random, one draw	0.075186	34.00%	0.26%	0

The MPNet assignment favors images that share annotated objects, relative to random re-pairing. The check measures neither the attributes nor the relations in the captions, and no human evaluation was conducted.

D.6 Sensitivity analyses and their coverage

Table D.9 lists each analysis and the checkpoints it covers. The neighborhood scales and the 16 MPNet assignments cover the full 22-checkpoint panel. The allocation, cohort, query, and reference-precision analyses were run on the representative checkpoints, and the ranking analysis on the 16-checkpoint panel for which it was specified.

Table D.9: Coverage of the retrieval-study sensitivity analyses. "Representative" denotes SigLIP B/16 224, CLIP B/16 224, and SigLIP2 B/16 224.

Analysis	Checkpoints	What varies	Location
Scales $k = 10, 20, 100$	All 22	Graph scale	Table D.10
16 MPNet assignments	All 22	Semantic assignment	Table D.10
128 cost-matched alternatives	Representative	Allocation at matched cost	Appendix D.6.1
Annotation selector (MPNet tie-break)	SigLIP	Selector	Tables D.4 and D.12
25 rebuilt cohorts	Representative	Cohort, graphs, assignments, references	Appendix D.6.2
Source-row resampling	SigLIP	Queries	Appendix D.6.3
1,000-draw reference distributions	Representative	Random draws	Appendix D.6.3
Retention-ranking analysis	16-checkpoint panel	Query groups, assignments	Appendix D.6.4

D.6.1 Assignment sensitivity

16 MPNet assignments. Across the canonical assignment and its 15 cost-jittered alternatives, the ordering $\rho_{A_{50}} > \rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$ holds for every checkpoint and every assignment (Table D.10). For SigLIP B/16 224, all 16 assignments exceed the random-reference mean for corrected agreement, R@10, and mean reciprocal rank; the prespecified gate required at least 14 for agreement and R@10. The annotation assignment has the same direction for all three outcomes.

128 cost-matched alternatives. Random re-pairing changes both the semantic cost of each assigned pair and the global allocation. This control asks whether the canonical allocation remains exceptional among allocations with approximately the same cost profile. It was frozen separately and constructed from selector costs alone. Define canonical overlap $o(p, p^\star) = n^{-1} \sum_i \mathbf{1}[p(i) = p^\star(i)]$ and the constrained frontier $\Phi(\gamma) = \min_{p \in \mathcal{P}_1 : o(p,p^\star) \le \gamma} \sum_i C_{i,p(i)}$. Penalized assignments $L(\lambda) = \min_{p \in \mathcal{P}_1} \sum_i [C_{i,p(i)} + \lambda \mathbf{1}[p(i) = p^\star(i)]]$ give the dual lower bound $L(\lambda) - \lambda\gamma n \le \Phi(\gamma)$. At the primary target $\gamma = 0.50$, the feasible cost ratio is 1.006704 and the dual lower bound is 1.006633; the target $\gamma = 0.25$ requires ratio 1.026166, outside the 1% budget. The generator solves $p_b = \arg\min_{p \in \mathcal{P}_1} \sum_i [C_{i,p(i)} + 0.014\, \mathbf{1}[p(i) = p^\star(i)] + \epsilon^{(b)}_{i,p(i)}]$ with independent $\epsilon^{(b)}_{ij} \sim \mathrm{Unif}[-h, h]$.

Table D.10: Neighborhood-scale and assignment sensitivity for all 22 checkpoints. Left: ρ_{A_k} at each graph scale under the canonical assignment; at every scale it exceeds $\rho_{\mathrm{R@10}}$ in Table D.2 (22/22 checkpoints). Right: ranges over the canonical MPNet assignment and its 15 cost-jittered alternatives (Table A.2). The ordering $\rho_{A_{50}} > \rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$ holds in 352 of 352 checkpoint–assignment combinations.

	ρ_{A_k}, canonical assignment				Range over 16 MPNet assignments		
Checkpoint	$k = 10$	$k = 20$	$k = 50$	$k = 100$	$\rho_{A_{50}}$	$\rho_{\mathrm{R@10}}$	$\rho_{\mathrm{R@1}}$
CLIP B/32 224	0.7277	0.7474	0.8081	0.8442	0.806–0.808	0.469–0.473	0.148–0.155
CLIP B/16 224	0.7196	0.7608	0.8091	0.8476	0.806–0.809	0.486–0.491	0.130–0.138
CLIP L/14 224	0.7110	0.7567	0.8116	0.8446	0.808–0.812	0.483–0.487	0.126–0.129
CLIP L/14 336	0.7149	0.7607	0.8160	0.8474	0.812–0.816	0.482–0.487	0.123–0.126
SigLIP B/16 224	0.6938	0.7281	0.7787	0.8145	0.777–0.781	0.470–0.474	0.096–0.099
SigLIP B/16 256	0.6861	0.7286	0.7825	0.8174	0.781–0.785	0.471–0.474	0.088–0.094
SigLIP B/16 384	0.6874	0.7224	0.7804	0.8158	0.779–0.782	0.476–0.478	0.094–0.098
SigLIP L/16 256	0.6610	0.7170	0.7699	0.8024	0.767–0.771	0.472–0.475	0.085–0.088
SigLIP L/16 384	0.6706	0.7144	0.7668	0.8015	0.764–0.767	0.478–0.481	0.084–0.085
SigLIP So400m/14 224	0.6770	0.7259	0.7795	0.8108	0.776–0.779	0.473–0.476	0.085–0.088
SigLIP2 B/16 224	0.6849	0.7265	0.7833	0.8126	0.781–0.784	0.475–0.477	0.096–0.097
SigLIP2 B/16 256	0.6887	0.7295	0.7864	0.8137	0.784–0.787	0.476–0.477	0.098–0.099
SigLIP2 B/16 384	0.6857	0.7209	0.7859	0.8151	0.784–0.786	0.473–0.475	0.095–0.097
SigLIP2 L/16 256	0.6764	0.7133	0.7666	0.8060	0.763–0.767	0.465–0.469	0.086–0.088
SigLIP2 L/16 384	0.6756	0.7141	0.7692	0.8062	0.767–0.770	0.460–0.463	0.085–0.087
SigLIP2 So400m/14 224	0.6752	0.7163	0.7656	0.8022	0.764–0.767	0.461–0.464	0.085–0.089
E5-V	0.6119	0.6642	0.7422	0.7899	0.741–0.742	0.529–0.532	0.110–0.113
Jina-CLIP-v2	0.6420	0.6904	0.7629	0.8073	0.762–0.764	0.519–0.521	0.123–0.127
LongCLIP-GmP ViT-L/14	0.6046	0.6608	0.7449	0.7886	0.744–0.745	0.528–0.531	0.105–0.108
LongCLIP-L (Diffusers)	0.6907	0.7396	0.7971	0.8318	0.794–0.797	0.506–0.507	0.113–0.117
Qwen3-VL-Embedding-2B	0.6135	0.6624	0.7344	0.7781	0.733–0.735	0.500–0.503	0.083–0.084
Qwen3-VL-Embedding-8B	0.5958	0.6427	0.7193	0.7650	0.718–0.720	0.489–0.491	0.073–0.074

The predeclared schedule was $h \in \{0.005, 0.008, 0.010, 0.012\}$; 0.008 was the first passing scale, and the first 128 qualifying assignments were retained. Eligibility required mean cost ratio at most 1.01, normalized sorted-cost Wasserstein error at most 0.02, largest normalized interior-decile error at most 0.05, canonical overlap at most 0.50, and pairwise overlap at most 0.80. The retained family has mean cost ratios 1.00750–1.00854, Wasserstein errors 0.00810–0.00944, decile errors 0.01193–0.02127, and canonical overlap 0.4810–0.4998, and every two alternatives differ on at least 1,006 targets. All are complete derangements with the original exclusions. The family matches the distribution of assigned costs, not each image's cost.

Table D.11: Canonical MPNet assignment versus the 128 cost-matched alternatives. Intervals are the 2.5–97.5% quantiles of the family, and "Percentile" is the fraction of the 128 values no greater than canonical. The generator is not a uniform conditional sampler, so these are not randomization p-values.

Checkpoint	Metric	Canonical	Mean	Central 95%	Percentile
SigLIP B/16 224	A_{50}	0.216984	0.217240	[0.215705, 0.218649]	0.398
SigLIP B/16 224	A_{50}^{corr}	0.778684	0.779647	[0.773874, 0.784948]	0.398
SigLIP B/16 224	I→T R@10	0.3844	0.3809	[0.3758, 0.3854]	0.891
SigLIP B/16 224	T→I R@10	0.3852	0.3868	[0.3820, 0.3918]	0.258
CLIP B/16 224	A_{50}	0.226720	0.226658	[0.225074, 0.228201]	0.508
CLIP B/16 224	A_{50}^{corr}	0.809054	0.808822	[0.802910, 0.814583]	0.508
CLIP B/16 224	I→T R@10	0.3522	0.3565	[0.3514, 0.3608]	0.055
CLIP B/16 224	T→I R@10	0.3384	0.3377	[0.3327, 0.3420]	0.617

Every canonical value lies inside its central 95% range (Table D.11). Monte Carlo standard errors are at most 0.000281 for corrected agreement and 0.000233 for R@10. The same neighborhood-agreement and R@10 regime therefore appears without the exact canonical allocation. For SigLIP2 B/16 224, the same 128 alternatives give corrected agreement 0.7771–0.7984, image-to-text R@10 35.94–37.22%, and R@1 3.70–4.24%, against random-reference maxima of 0.0028, 0.42%, and 0.08%. Its canonical value lies inside the central 95% range for agreement, image-to-text R@1 and mean reciprocal rank, and text-to-image R@10.

D.6.2 COHORT AND GRAPH SENSITIVITY

The prespecified reconstruction partitions the 5,000 identities deterministically into 25 groups of 200. Each reconstruction removes one group, keeps the same 4,800 image and caption identities in every condition, rebuilds the neighbor graphs, re-solves the MPNet and annotation assignments inside the retained cohort, and draws 256 new random re-pairings. Full-cohort values use the 1,000-draw references. The contrast is the hard negative value minus the within-cohort random-reference mean, in chance-corrected agreement at $k = 50$. For the $G = 25$ reconstruction estimates $\hat{\theta}_{(-g)}$, we report the jackknife-scaled interval, the full-cohort estimate plus or minus $1.96\,\mathrm{SE}_{\mathrm{JK}}$, with

$$\mathrm{SE}_{\mathrm{JK}} = \sqrt{\frac{G-1}{G}\sum_{g=1}^{G}(\hat{\theta}_{(-g)} - \bar{\theta})^2}, \qquad \bar{\theta} = \frac{1}{G}\sum_{g}\hat{\theta}_{(-g)}. \tag{D.1}$$

It summarizes sensitivity to this deterministic pipeline perturbation, not population sampling uncertainty.

Table D.12: Contrasts at $k = 50$ across the 25 rebuilt cohorts. Every contrast is positive in all 25 reconstructions, and every jackknife-scaled interval excludes zero; this is not a population significance test. The SigLIP annotation contrast uses the MPNet tie-break.

Contrast	Full	Mean	Median	Range	Jackknife SE	JK-scaled interval
SigLIP MPNet	0.778680	0.779902	0.779307	[0.772023, 0.787279]	0.018629	[0.742167, 0.815192]
SigLIP annotation	0.650649	0.652019	0.652536	[0.642217, 0.658561]	0.021803	[0.607916, 0.693381]
CLIP MPNet	0.809074	0.810852	0.810292	[0.801860, 0.821293]	0.022201	[0.765561, 0.852586]

Table D.13: Every reconstruction estimate at $k = 50$. The group index identifies the removed group of 200 identities; M and A denote the MPNet and annotation selectors.

Group	SigLIP M	SigLIP A	CLIP M	Group	SigLIP M	SigLIP A	CLIP M
0	0.780692	0.653811	0.801860	13	0.778899	0.642391	0.812611
1	0.778360	0.651058	0.806480	14	0.776389	0.648981	0.807505
2	0.785368	0.652319	0.810292	15	0.784254	0.649576	0.814846
3	0.772023	0.655846	0.807835	16	0.780897	0.656111	0.817357
4	0.783694	0.653232	0.812439	17	0.780897	0.655069	0.817539
5	0.778497	0.653076	0.811210	18	0.783314	0.656556	0.813766
6	0.782534	0.658561	0.807740	19	0.783191	0.657612	0.817824
7	0.778063	0.646461	0.806445	20	0.781546	0.658101	0.807825
8	0.773015	0.652536	0.805187	21	0.787279	0.649994	0.814045
9	0.779307	0.657843	0.811582	22	0.784294	0.652639	0.821293
10	0.778351	0.646542	0.808393	23	0.773585	0.650013	0.809536
11	0.776658	0.650107	0.812713	24	0.777737	0.642217	0.806259
12	0.778702	0.649833	0.808719	–	–	–	–

Table D.14: Secondary neighborhood scales in the rebuilt cohorts. Every estimate is positive in all 25 reconstructions. These scales are descriptive sensitivity analyses.

Contrast	k	Full cohort	Reconstruction mean	Reconstruction range
SigLIP MPNet	10	0.693821	0.695438	[0.684960, 0.710334]
	20	0.728141	0.727194	[0.718173, 0.738984]
	100	0.814513	0.814160	[0.808764, 0.820831]
SigLIP annotation	10	0.499492	0.503693	[0.483831, 0.513656]
	20	0.558116	0.560517	[0.542462, 0.570559]
	100	0.705217	0.705341	[0.694120, 0.710384]
CLIP MPNet	10	0.719648	0.720228	[0.709879, 0.736996]
	20	0.760792	0.762811	[0.751077, 0.775702]
	100	0.847623	0.846807	[0.839961, 0.853999]

Re-solving inside a retained cohort changes 14–18% of eligible targets in every reconstruction (MPNet: 14.35–17.54%, overlap with the full-cohort assignment 0.8246–0.8565; annotation: 14.36–

17.96%, overlap 0.8204–0.8564), so the reconstructions also vary the assignment. The full-cohort reconstruction audit reproduced all 12 stored agreement values across the four k with maximum absolute error 1.11×10^{-16}.

For SigLIP2 B/16 224, the same 25 cohorts use the frozen within-cohort MPNet assignment, rebuilt graphs, and 256 new random re-pairings each (Table D.15). Every reconstruction is positive at every scale.

Table D.15: SigLIP2 B/16 224 contrasts in chance-corrected agreement. "Full" subtracts the 1,000-draw full-cohort reference mean; each reconstruction subtracts its own 256-draw mean.

k	Full	Reconstruction range	Jackknife-scaled interval	Positive
10	0.68487	[0.67268, 0.70587]	[0.60276, 0.76698]	25/25
20	0.72654	[0.72358, 0.74396]	[0.67524, 0.77783]	25/25
50	0.78330	[0.77983, 0.79354]	[0.74690, 0.81971]	25/25
100	0.81264	[0.80716, 0.81929]	[0.78662, 0.83867]	25/25

D.6.3 Query resampling and reference precision

Source-row resampling. Holding galleries and assignments fixed, each of 2,000 draws resamples the 5,000 saved source rows with replacement, using the same indices for the MPNet assignment and the one-draw random assignment (Table D.16). Image-to-text rows pair the same image query across conditions. Text-to-image rows are indexed by source image i, with text query $p(i)$, so paired rows can represent different text queries. The two directions use separate seeded draws, and the endpoints are the 2.5th and 97.5th percentiles. These intervals describe conditional source-row sensitivity, not uncertainty over assignments or populations.

Table D.16: SigLIP B/16 224: 2,000 paired fixed-gallery source-row resamples per direction. The contrast is the MPNet assignment minus the one-draw random assignment.

Direction	Contrast	Estimate	95% interval
Image → text	R@10 (percentage points)	+38.38	[37.04, 39.76]
Text → image	R@10 (percentage points)	+38.40	[37.14, 39.80]
Image → text	Mean reciprocal rank	+0.1520	[0.1458, 0.1585]
Text → image	Mean reciprocal rank	+0.1594	[0.1529, 0.1667]

Reference distributions. For SigLIP B/16 224, none of the 1,000 core draws reached the hard negative value for corrected agreement, image-to-text R@10, or mean reciprocal rank; the add-one upper-tail frequency is $1/1001$. This is a descriptive comparison with the generator, not a randomization p-value. At $k = 50$, the SigLIP reference mean of corrected agreement is 4.09×10^{-6} (SD 8.64×10^{-4}, Monte Carlo SE 2.73×10^{-5}), and the CLIP mean is -1.98×10^{-5} (SD 8.96×10^{-4}, Monte Carlo SE 2.83×10^{-5}); across rebuilt cohorts, the largest Monte Carlo SE is 6.23×10^{-5}. Table D.17 gives the SigLIP2 reference; every SigLIP2 hard negative outcome lies beyond the range of all 1,000 draws.

Table D.17: SigLIP2 B/16 224 reference distribution over 1,000 random re-pairings.

Outcome	Mean	Central 95%	Maximum	MC SE of mean
A_{50}^{corr}	−0.000017	[−0.001714, 0.001699]	0.002811	0.000028
Image-to-text R@1	0.0108%	[0.0000, 0.0400]%	0.0800%	0.00046 pp
Image-to-text R@10	0.1828%	[0.0800, 0.3000]%	0.4200%	0.00187 pp
Image-to-text MRR	0.001696	[0.001349, 0.002146]	0.002633	0.0000068
Text-to-image R@10	0.1809%	[0.0800, 0.3200]%	0.3600%	0.00187 pp

D.6.4 Retention rankings across the 16-checkpoint panel

The prespecified ranking analysis asks whether the three retentions order the 16 CLIP, SigLIP, and SigLIP2 checkpoints differently. It was not extended to the six additional encoders. We use *broad lineage* for CLIP versus the SigLIP lineage and keep the three prespecified family labels for the inversion analysis; SigLIP2 is a newer generation within the SigLIP lineage. A robust inversion requires both paired model-difference intervals, from the fixed-gallery delete-group analysis over 25 query groups of 200, to exclude zero.

Table D.18: Prespecified retention-ranking analysis over the 16-checkpoint panel. "Cross-family" uses the three labels CLIP, SigLIP, and SigLIP2.

Retention pair	τ_b	ρ_s	Inversions	Cross-family	Robust	Robust cross-family
A_{50} vs. R@10	0.4833	0.6529	31	16	2	0
A_{50} vs. R@1	0.6667	0.8765	20	7	1	0
R@10 vs. R@1	0.3500	0.4147	39	23	4	0

No cross-family inversion among the three retentions meets the robustness criterion (Table D.18; ρ_s is Spearman's coefficient). The rank associations are nevertheless imperfect, and seven robust within-family inversions remain: for A_{50} versus R@10, SigLIP B/16 224 and B/16 256 each reverse order with SigLIP L/16 384; for A_{50} versus R@1, CLIP B/32 224 reverses with CLIP L/14 336; and for R@10 versus R@1, CLIP B/32 224 reverses with CLIP B/16 224, L/14 224, and L/14 336, and SigLIP B/16 224 with SigLIP L/16 384. These are seven metric-pair occurrences rather than seven distinct checkpoint pairs. The four CLIP checkpoints form the Pareto front over the three retentions, which describes the profile rather than defining a leaderboard.

Across the 25 query groups, the Kendall τ_b replicate ranges are [0.3333, 0.5167], [0.6000, 0.7333], and [0.2333, 0.4500] for the three pairs, with jackknife standard errors 0.2105, 0.2049, and 0.2820. These quantify sensitivity to the fixed query cohort; the 16 correlated checkpoints are not treated as an independent sample of models. Repeating the analysis across the 16 MPNet assignments gives τ_b ranges [0.3833, 0.4833], [0.6000, 0.7000], and [0.3000, 0.4000], with 31–37, 18–24, and 36–42 inversions, so the pattern is not specific to the canonical assignment. Original agreement and retention are distinct checkpoint descriptors: original A_{50} has Kendall τ_b 0.5833, 0.3333, and 0.5500 with the three retentions, with 11, 6, and 9 robust cross-family inversions.

E Checkpoint selection

This appendix supports Section 5.3. The Flickr30k study, which provides the held-out evidence, comes first: its protocol, the complete validation and test panel, the paired bootstrap interval, and the five-caption evaluation. The retrospective COCO analyses and the post hoc follow-ups come after it.

E.1 Flickr30k protocol

Protocol fixed before outcomes. The study uses Flickr30k (Young et al., 2014) with the standard Karpathy split: all 1,014 validation images and 1,000 test images, excluding the training split. The 16-checkpoint roster (Table D.1), revisions, extraction settings, caption rule, metrics, selection rules, and analysis were fixed on 2026-09-09 at 00:09:23 UTC, before any representation was extracted or any outcome examined. Validation metrics and selected checkpoints were saved at 00:44:03 UTC, before any test outcome was computed, and the completed result was recorded at 00:46:17 UTC. This is a prospective protocol within the project, not an external preregistration. The study evaluates checkpoint selection on original pairs; it involves no re-pairing and no CKA.

Data order and captions. Images are ordered by numeric file identity within validation and then within test, and each image's five captions are ordered by sentence ID. The primary analysis selects one caption per image with the outcome-independent rule and seed of the COCO cohort (seed 20260805). The caption graph uses the same selected captions as retrieval.

Selection rules. On validation images only, the agreement rule selects the checkpoint with the highest A_{50}, and the recall rule selects the checkpoint with the highest bidirectional one-caption R@1. Both modality graphs use exact self-excluding neighbors with $k = 50$. Exact checkpoint ties would go to the first roster entry and score ties to the first gallery index; every reported maximum is unique. Test images never select a checkpoint, and no reverse orientation is evaluated.

Five-caption evaluation. The prespecified secondary metric lets each image query rank all five captions of every image in its split and counts success when its top caption belongs to that image; each caption query ranks all images and succeeds when its own image ranks first. The two directions receive equal weight. A second recall rule maximizes this validation metric, while the agreement rule keeps the one-caption A_{50}. The primary selected checkpoints are also evaluated on this five-caption test metric.

Extraction controls. Native forward reconstruction and deterministic re-encoding are checked for every checkpoint. The normalized float32 caches share ordered image identities, captions, and selected caption indices across the roster. Text limits are the native ones, including 16 tokens for SigLIP So400m/14 224 and 64 for SigLIP2 So400m/14 224.

E.2 Complete Flickr30k panel

Table E.1 extends Table 2 with the test A_{50} and both five-caption recalls.

Table E.1: Complete Flickr30k validation and test panel. A_{50} is a fraction; R@1 columns are bidirectional percentages. "One" uses the selected caption, and "Five" counts all five captions of an image as positives. Bold validation values are the selected maxima; test values do not select checkpoints. Validation has 1,014 images and test 1,000, and each five-caption text gallery is five times as large. Absolute values are not comparable with the larger COCO galleries.

	Validation			Test		
Checkpoint	A_{50}	One R@1	Five R@1	A_{50}	One R@1	Five R@1
CLIP B/32 224	0.202288	62.72	69.20	0.201140	59.50	69.12
CLIP B/16 224	0.207160	66.37	72.99	0.207600	61.40	72.32
CLIP L/14 224	0.207968	69.77	77.25	0.211140	65.40	74.99
CLIP L/14 336	0.214162	71.94	79.77	0.211780	67.95	77.27
SigLIP B/16 224	0.231085	75.49	82.33	0.223760	74.35	81.88
SigLIP B/16 256	**0.232978**	76.13	83.47	0.228000	75.20	82.78
SigLIP B/16 384	0.230888	78.90	85.12	0.227240	78.25	85.57
SigLIP L/16 256	0.224260	80.87	86.79	0.223020	79.65	85.38
SigLIP L/16 384	0.222032	**82.64**	**88.66**	0.215520	80.95	87.56
SigLIP So400m/14 224	0.206114	76.38	84.48	0.207420	75.65	83.12
SigLIP2 B/16 224	0.222150	73.62	81.94	0.219320	73.15	81.72
SigLIP2 B/16 256	0.221558	75.74	82.62	0.220260	74.70	82.34
SigLIP2 B/16 384	0.222801	77.81	84.91	0.219200	77.15	83.94
SigLIP2 L/16 256	0.221065	78.30	85.89	0.218560	77.35	83.92
SigLIP2 L/16 384	0.221696	79.98	86.84	0.215380	78.35	84.96
SigLIP2 So400m/14 224	0.225819	78.94	85.52	0.223100	77.85	85.03

E.3 Held-out consequences and paired bootstrap interval

The agreement rule selects SigLIP B/16 256, and both recall rules select SigLIP L/16 384. Their one-caption test recalls are 75.20% and 80.95%, a 5.75-point gap (recall-selected minus agreement-selected), with 95% paired bootstrap interval [3.80, 7.80]. On the five-caption test metric the recalls are 82.78% and 87.56%, a 4.78-point gap (Table E.2). The five-caption result is the same whether the recall rule uses one-caption or five-caption validation recall, because both choose the same checkpoint.

Table E.2: Test consequences of validation selection on Flickr30k. The agreement-selected checkpoint is SigLIP B/16 256 and the recall-selected checkpoint is SigLIP L/16 384 in both rows. Recalls are percentages; gaps and intervals are percentage points. Intervals describe paired image-level variability conditional on the selected checkpoints and the test galleries.

Test evaluation	Agreement-selected	Recall-selected	Gap	95% interval
One caption (primary)	75.20	80.95	$+5.75$	$[+3.80, +7.80]$
Five captions (secondary)	82.78	87.56	$+4.78$	$[+3.62, +5.97]$

Bootstrap. Each interval uses 10,000 paired bootstrap draws over image clusters, generated with NumPy's PCG64 and seed 20260909, and takes the 2.5th and 97.5th percentiles. A cluster contains an image query and its caption queries; for the five-caption metric, caption-query outcomes are averaged within each image before the two directions are averaged. Checkpoints, galleries, ranks, and validation choices stay fixed. The intervals therefore exclude model-selection uncertainty, gallery variation, and variation across datasets, and they are not population-level inference about future datasets or checkpoints.

Subgroups and correlations. Restricting either rule to the SigLIP family, or to the combined SigLIP/SigLIP2 lineage, selects the same two checkpoints as the full panel, with the same gaps. Within CLIP, both rules select CLIP L/14 336. Within SigLIP2, the agreement rule selects SigLIP2 So400m/14 224 and both recall rules select SigLIP2 L/16 384: the one-caption gap is $+0.50$ points ($[-1.15, +2.15]$) and the five-caption gap is -0.07 points ($[-1.21, +1.08]$). These prespecified subgroup comparisons are descriptive and do not show an advantage within every family. Across the 16 checkpoints, the agreement–recall Kendall τ_b is $+0.3667$ for one-caption and $+0.3167$ for five-caption recall on validation, and $+0.3167$ and $+0.3833$ on test. These associations are positive, yet agreement still selects a worse retriever, which is the point of Section 5.3.

E.4 Retrospective COCO selection

These analyses use the 16-checkpoint COCO panel of Appendix D with original pairs. They were carried out after the full-cohort outcomes were known and are retrospective.

Full cohort. On the complete 5,000-pair cohort, agreement selection picks CLIP L/14 336 and recall selection picks SigLIP L/16 384, whose bidirectional R@1 values differ by 14.57 points (Table E.3). The panel-wide agreement–recall Kendall τ_b is -0.4667. This description uses the same images for selection and evaluation.

Within the combined SigLIP/SigLIP2 lineage, agreement selects SigLIP2 B/16 256 (47.85%) and recall selects SigLIP L/16 384 (52.87%), a 5.02-point gap. The prespecified family gaps are 0 for CLIP, 5.57 for SigLIP, and 2.83 for SigLIP2.

Disjoint halves. One outcome-independent permutation of the 5,000 identities, generated with `numpy.random.Generator(numpy.random.PCG64(20260905))`, splits the cohort into halves A (first 2,500) and B; neither half is re-sorted, and both modalities share the order. The partition was specified after the full-cohort results. For each half and checkpoint, both 50-neighbor graphs are rebuilt within the half, and scores are ranked against that half's 2,500-item gallery. One half selects and the other evaluates (Table E.4). Raw A_{50} is higher in a half because $c_{50} = 50/(n-1)$ grows as n shrinks; the graph algorithm is unchanged.

Secondary split gaps are $+5.22$ and $+4.50$ points within the combined SigLIP lineage, $+5.62$ and $+5.42$ within SigLIP, and $+2.16$ and $+2.92$ within SigLIP2. Within CLIP they are zero, because both rules select CLIP L/14 336 in both orientations.

These selection results compare original-pair measurements. They are distinct from the retention-ranking analysis of Appendix D.6.4, which concerns responses to re-pairing normalized by each checkpoint's own original and random values.

Table E.3: Original-pair COCO metrics for the 16-checkpoint panel. Bold entries are the maxima selected by each rule. Kendall $\tau_b = -0.4667$.

Checkpoint	$A_{50}(p_0)$	Bidirectional R@1
CLIP B/32 224	0.274164	32.04%
CLIP B/16 224	0.277868	34.59%
CLIP L/14 224	0.283524	37.22%
CLIP L/14 336	**0.290168**	38.30%
SigLIP B/16 224	0.275812	47.23%
SigLIP B/16 256	0.279436	47.30%
SigLIP B/16 384	0.276444	49.17%
SigLIP L/16 256	0.256448	51.10%
SigLIP L/16 384	0.249548	**52.87%**
SigLIP So400m/14 224	0.239244	50.81%
SigLIP2 B/16 224	0.286752	46.88%
SigLIP2 B/16 256	0.287848	47.85%
SigLIP2 B/16 384	0.287448	49.57%
SigLIP2 L/16 256	0.271912	50.16%
SigLIP2 L/16 384	0.270448	50.68%
SigLIP2 So400m/14 224	0.267132	49.63%

Table E.4: COCO disjoint-half selection. The gap is the recall-selected minus the agreement-selected evaluation R@1, in percentage points. The two orientations are a split-sensitivity analysis of one cohort.

Orientation	Checkpoint selected by	Selection A_{50}	Selection R@1	Evaluation R@1	Gap
A $\to$ B	Agreement: CLIP L/14 336	0.341536	46.90%	45.00%	—
	Recall: SigLIP L/16 384	0.281768	62.20%	59.84%	+14.84
B $\to$ A	Agreement: CLIP L/14 336	0.338864	45.00%	46.90%	—
	Recall: SigLIP L/16 384	0.283496	59.84%	62.20%	+15.30

E.5 POST HOC FOLLOW-UPS

The analyses below were specified after the Flickr30k and COCO outcomes were known. They add no new held-out evaluation.

Transfer across datasets. Selecting on one dataset and reading the other's recall (Table E.5) reuses the complete COCO panel and the Flickr30k results, with lexicographic tie-breaking by checkpoint key. Both transfer gaps were known before this follow-up was specified, so these are not prospective replications.

Table E.5: Transfer of the two selection rules across datasets. Bidirectional one-caption R@1 in percent; the gap is recall-selected minus agreement-selected. COCO uses 5,000 images and Flickr30k test 1,000. No interval is reported.

Source → target	Agreement-selected	Recall-selected	Agreement R@1	Recall R@1	Gap
COCO → Flickr30k test	CLIP L/14 336	SigLIP L/16 384	67.95	80.95	+13.00
Flickr30k validation → COCO	SigLIP B/16 256	SigLIP L/16 384	47.30	52.87	+5.57

The sign reversal between COCO and Flickr30k. The agreement–recall correlation is negative on COCO ($\tau_b = -0.467$) and positive on Flickr30k validation ($+0.367$). This follow-up reuses the 16 COCO and Flickr30k validation caches; the subset construction and the complete scale grid were specified before any new outcome was computed. Original caption choices and pair identities stay fixed, and there is no new extraction, re-pairing, or test evaluation.

Most of the reversal occurs between families. At $k = 50$, CLIP's within-family τ_b is $+1.000$ on both datasets and SigLIP's is -0.467 on both (Table E.6). Across families, agreement and recall order 18 of 84 checkpoint pairs concordantly on COCO and 64 of 84 on Flickr30k; within families, the counts are 14 and 18 of 36, with no ties. Net concordance (concordant minus discordant pairs) rises from -56 to $+44$, which decomposes into $+92$ between families and $+8$ within families. This is a pair-count identity, not a causal attribution. Across the 120 checkpoint pairs, the agreement ordering changes in 59 and the recall ordering in nine: CLIP has relatively high agreement but low recall on COCO, and relatively low values of both on Flickr30k.

Table E.6: Original-cohort agreement–recall associations at $k = 50$. COCO uses 5,000 items and Flickr30k validation 1,014. Subgroup coefficients describe fixed groups of four or six checkpoints.

Checkpoints	COCO τ_b	Flickr30k validation τ_b
All 16	-0.467	$+0.367$
CLIP (4)	$+1.000$	$+1.000$
SigLIP (6)	-0.467	-0.467
SigLIP2 (6)	-0.467	$+0.067$

For each dataset, 25 seeded permutations define nested 500- and 1,000-image subsets, shared across checkpoints, metrics, and scales; the two datasets use separate random streams. Retained images keep their order and selected caption, and both neighbor sets and retrieval galleries are rebuilt inside each subset. At matched $n = 1{,}000$ and $k = 50$, the median τ_b is -0.583 on COCO (range -0.762 to -0.400) and $+0.400$ on Flickr30k ($+0.333$ to $+0.500$), and every subset keeps its dataset's sign at all four scales (Table E.7). At $n = 500$ and $k = 50$, the signs again differ in every subset, with medians -0.393 and $+0.544$. Gallery-size mismatch and the single choice $k = 50$ therefore do not explain the reversal at matched size. The grid also shows scale dependence: original COCO has $\tau_b = +0.100$ at $k = 10$, and some 500-image conditions have mixed signs (Table E.7). At $k = 50$, the agreement and recall maxima differ in all 100 subset conditions. On Flickr30k at $n = 1{,}000$, every draw selects SigLIP B/16 256 by agreement and SigLIP L/16 384 by recall, matching the original validation choices; COCO at that size selects CLIP L/14 336 by agreement in 18 of 25

Table E.7: Complete size/scale sensitivity grid. Subset rows show the median Kendall τ_b, full minimum–maximum range, and positive/zero/negative counts across 25 draws. Rows with $n = 5{,}000$ or $1{,}014$ are single original-cohort references. Ranges are descriptive, not confidence intervals.

Dataset	n	k	k/n	τ_b	Min, max	+ / 0 / –
COCO	500	5	1.0%	−0.312	−0.644, +0.143	1 / 0 / 24
COCO	500	10	2.0%	−0.450	−0.740, −0.194	0 / 0 / 25
COCO	500	20	4.0%	−0.467	−0.790, −0.167	0 / 0 / 25
COCO	500	50	10.0%	−0.393	−0.756, −0.075	0 / 0 / 25
COCO	500	100	20.0%	−0.100	−0.346, +0.295	7 / 1 / 17
COCO	1,000	10	1.0%	−0.383	−0.633, −0.183	0 / 0 / 25
COCO	1,000	20	2.0%	−0.533	−0.795, −0.350	0 / 0 / 25
COCO	1,000	50	5.0%	−0.583	−0.762, −0.400	0 / 0 / 25
COCO	1,000	100	10.0%	−0.510	−0.711, −0.200	0 / 0 / 25
COCO	5,000	10	0.2%	+0.100	—	—
COCO	5,000	20	0.4%	−0.117	—	—
COCO	5,000	50	1.0%	−0.467	—	—
COCO	5,000	100	2.0%	−0.600	—	—
Flickr30k val.	500	5	1.0%	+0.159	−0.110, +0.510	19 / 1 / 5
Flickr30k val.	500	10	2.0%	+0.319	+0.000, +0.567	24 / 1 / 0
Flickr30k val.	500	20	4.0%	+0.427	+0.176, +0.605	25 / 0 / 0
Flickr30k val.	500	50	10.0%	+0.544	+0.303, +0.628	25 / 0 / 0
Flickr30k val.	500	100	20.0%	+0.633	+0.483, +0.733	25 / 0 / 0
Flickr30k val.	1,000	10	1.0%	+0.243	+0.092, +0.317	25 / 0 / 0
Flickr30k val.	1,000	20	2.0%	+0.343	+0.300, +0.417	25 / 0 / 0
Flickr30k val.	1,000	50	5.0%	+0.400	+0.333, +0.500	25 / 0 / 0
Flickr30k val.	1,000	100	10.0%	+0.527	+0.483, +0.550	25 / 0 / 0
Flickr30k val.	1,014	10	1.0%	+0.226	—	—
Flickr30k val.	1,014	20	2.0%	+0.333	—	—
Flickr30k val.	1,014	50	4.9%	+0.367	—	—
Flickr30k val.	1,014	100	9.9%	+0.533	—	—

draws and SigLIP L/16 384 by recall in 25 of 25. These are subset-selection disagreements, not new held-out costs.

All 32 original checkpoint–dataset records reproduce their saved agreement and recall values, and direct distance and set-intersection reconstructions agree exactly in 64 preselected subset cases covering every checkpoint and both datasets. The overlapping subsets measure sensitivity within the retained datasets; their ranges are not population intervals, and the remaining dataset-dependent ranking difference does not isolate image content, caption properties, architecture, or training data as its cause.

Oracle regret, rank weighting, and model-pool sensitivity. These diagnostics use only the retained scalar tables of the Flickr30k study and the COCO split analysis. For a checkpoint pool $\mathcal{M}$, let r_m be checkpoint m's held-out bidirectional R@1 as a fraction and $r^* = \max_{m \in \mathcal{M}} r_m$. For the agreement (G) or recall (R) rule g,

$$O_g(\mathcal{M}) = 100\,(r^* - r_{g(\mathcal{M})}), \qquad \mathrm{Rel@1}_g(\mathcal{M}) = r_{g(\mathcal{M})}/r^*, \tag{E.1}$$

an observed single-pick analogue of model-selection regret (Renggli et al., 2022), with Rel@1 following Agostinelli et al. (2022). The paired gap $100(r_R - r_G) = O_G - O_R$ equals the agreement rule's regret only when the recall rule attains the oracle. Correlations use tie-corrected Kendall τ_b and the additive hyperbolic weighted Kendall coefficient of `scipy.stats.weightedtau`: with zero-based importance ranks u_i, each pair has weight $1/(u_i + 1) + 1/(u_j + 1)$. Our τ_w averages the coefficients computed with descending lexicographic ranks of (agreement, recall) and (recall, agreement), matching `scipy.stats.weightedtau` with `rank=True`; τ_w^R uses recall-first ranks only. No significance test is inferred. In the leave-one-model-out analysis, each checkpoint is omitted in turn, both rules reselect on the remaining 15, and the oracle is recomputed among them (Tables E.8 to E.10). These overlapping pools measure sensitivity to the finite roster; they are not independent replications or population intervals.

Table E.8: Full-panel selection diagnostics. Oracle regrets are in percentage points; Rel@1 is a fraction. Correlations use selection-split agreement and one-caption recall. G and R identify the agreement and recall rules. The Flickr30k rows share their selection pairs. τ_w^R uses recall-first importance ranks; τ_w averages both importance orders.

Study	Oracle G	Oracle R	Rel@1 G	Rel@1 R	τ_b	τ_w	τ_w^R
Flickr30k, one caption	5.75	0.00	0.9290	1.0000	+0.3667	+0.1914	+0.2364
Flickr30k, five captions	4.78	0.00	0.9454	1.0000	+0.3667	+0.1914	+0.2364
COCO A → B	14.84	0.00	0.7520	1.0000	−0.5272	−0.5889	−0.6614
COCO B → A	15.30	0.00	0.7540	1.0000	−0.5667	−0.6011	−0.7250

Table E.9: Leave-one-model-out sensitivity. Each row summarizes all 16 omissions, with both choices and the oracle recomputed among the remaining 15 checkpoints. Ranges are extrema across omissions, not confidence intervals. The paired gap is R minus G; it equals oracle regret G minus oracle regret R.

Study	Oracle regret G	Oracle regret R	Paired gap	R better
Flickr30k, one caption	[4.45, 6.60]	[0.00, 0.00]	[4.45, 6.60]	16/16
Flickr30k, five captions	[2.79, 5.68]	[0.00, 0.19]	[2.60, 5.68]	16/16
COCO A → B	[13.36, 16.16]	[0.00, 1.58]	[11.78, 16.16]	16/16
COCO B → A	[4.50, 15.30]	[0.00, 0.48]	[4.50, 15.30]	16/16

Table E.10: Oracle regret for every checkpoint omission (percentage points). Each G/R pair gives the agreement/recall rule's regret against the best held-out checkpoint in that reduced pool. No interval is inferred from these overlapping subsets.

	Flickr30k one		Flickr30k five		COCO A→B		COCO B→A	
Omitted checkpoint	G	R	G	R	G	R	G	R
CLIP B/32 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
CLIP B/16 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
CLIP L/14 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
CLIP L/14 336	5.75	0.00	4.78	0.00	16.16	0.00	4.50	0.00
SigLIP B/16 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP B/16 256	6.60	0.00	5.68	0.00	14.84	0.00	15.30	0.00
SigLIP B/16 384	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP L/16 256	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP L/16 384	4.45	0.00	2.79	0.19	13.36	1.58	13.72	0.48
SigLIP So400m/14 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP2 B/16 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP2 B/16 256	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP2 B/16 384	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP2 L/16 256	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP2 L/16 384	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00
SigLIP2 So400m/14 224	5.75	0.00	4.78	0.00	14.84	0.00	15.30	0.00

F Additional operators and modalities

This appendix supports the late-interaction and audio results at the end of Section 5.2 and reports two further extensions: within-video segment assignment and dense captions. The six additional global image–text encoders are part of the 22-checkpoint panel and are reported in Appendix D.

F.1 Protocols and provenance

Status. Each runner serializes its roster, revisions, assignments, capability gate, and scoring route with its output, and these were fixed before the final result-bearing run of each arm. The arms were combined into the present cross-setting argument after their outcomes were inspected, so we treat the patterns across rosters as descriptive breadth, not as population-level inference about models.

Checkpoints. Every system has public weights and a public model or paper record (Table F.1). The LanguageBind loader exposes no revision argument; the listed revision records the intended checkpoint, the run stores the revision the loader resolved, and the two match. All other loaders enforce the listed revision.

Table F.1: Extension systems at their evaluated revisions.

Role	Public checkpoint	Revision
Late interaction	`vidore/colqwen2.5-v0.2` (Faysse et al., 2025; Vidore, 2025a)	`dcbe8d9cede5`
Late interaction, video, audio	`vidore/colqwen-omni-v0.1` (Vidore, 2025b)	`c0f50696f483`
Late interaction	`tencent/EVIE-Preview-4.5B` (Tencent, 2026)	`1c43543b000b`
Video	`LanguageBind/LanguageBind_Video` (Zhu et al., 2024)	`84bed6cfe757`
Video, audio	`Tevatron/OmniEmbed-v0.1` (Ma et al., 2025)	`fa8be315e74d`
Video	`Qwen/Qwen3-VL-Embedding-2B` and `-8B` (Li et al., 2026)	Table D.1
Audio	`laion/clap-htsat-unfused` (Wu et al., 2023)	`8fa0f1c6d043`
Audio	`laion/clap-htsat-fused` (Wu et al., 2023)	`365dea6ef167`
Audio	`laion/larger_clap_general` (Wu et al., 2023)	`ada0c23a36c4`

Preprocessing. Late-interaction processors are dynamic: 3,136–602,112 pixels for ColQwen2.5, up to 802,816 pixels for ColQwen-Omni, and 65,536–786,432 pixels for EVIE. Captions are passed verbatim to `encode_query` with no template, and images to `encode_document`. Audio is mono float32, resampled to 48 kHz for CLAP and 16 kHz for the Whisper-fronted ColQwen-Omni and OmniEmbed paths. OmniEmbed is evaluated with its declared query prompt and with bare text; these are separate configurations and are not averaged.

Assignments, reference, and gate. The late-interaction cells use the COCO cohort and reuse the canonical MPNet assignment of the retrieval study. The audio arm selects one of the five captions of each AudioCaps test clip with a deterministic hash (seed 20260805), which observes no embedding or model, and solves the MPNet assignment of Appendix A.4 over those 975 captions. Both arms use as random reference the mean of 256 random derangements from the extension runner, seeded from 20260810, each a complete bijection with no fixed point and no identical normalized caption (Table A.3). Retention follows Equation 5 without clipping, and bidirectional values average the raw directional recalls first. Before any retention is reported, a capability gate requires an original-minus-reference bidirectional R@10 of at least 0.05 and an original R@10 above five times the uniform-chance R@10. A configuration that fails is left uninterpreted rather than counted as low retention. In the audio arm, a negative directional retention remains visible but precludes interpreting that configuration's bidirectional value.

F.2 Native late interaction

The three systems use their deployed MaxSim operator, provided by `sentence_transformers.MultiVectorEncoder` and not reimplemented. For cap-

tion token vectors $a_1, \ldots, a_m$ and image vectors $b_1, \ldots, b_\ell$,

$$s(a, b) = \sum_{u=1}^{m} \max_{1 \le w \le \ell} \langle a_u, b_w \rangle. \tag{F.1}$$

Token vectors and any normalization are those returned by each checkpoint. A single pooled representation, and hence a neighbor graph, is not defined, so only retrieval is measured. One text-query/image-document matrix is fixed for every assignment. Image-to-text and text-to-image denote row-wise and column-wise ranking of that same asymmetric matrix; no reverse image-query operator is computed.

Table F.2: Native late interaction on the COCO cohort. Both directions rank the same text-query/image-document MaxSim matrix. The bidirectional rows average raw directional recalls before retention (Appendix A.1). The ordering $\rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$ holds in five of six directions and in all three bidirectional aggregates; ColQwen2.5 image-to-text is the exception.

Model	Direction	$\rho_{\mathrm{R@10}}$	$\rho_{\mathrm{R@1}}$	$\rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$
ColQwen2.5	Image to text	0.3627	0.3725	**no**
	Text to image	0.4673	0.0994	yes
	Bidirectional	0.4468	0.1060	yes
ColQwen-Omni	Image to text	0.4276	0.4192	yes
	Text to image	0.4620	0.0889	yes
	Bidirectional	0.4578	0.0931	yes
EVIE	Image to text	0.3059	0.1850	yes
	Text to image	0.4502	0.0835	yes
	Bidirectional	0.4036	0.0912	yes

Table F.2 gives both directions and the bidirectional aggregate. R@10 retention exceeds R@1 retention in five of six directions and in all three aggregates. The exception is ColQwen2.5 image-to-text, where each image ranks candidate captions by text-query MaxSim, which sums over caption tokens without length normalization. Caption length is a candidate explanation that we did not test.

F.3 Audio–text retrieval

On AudioCaps (Kim et al., 2019), all six configurations pass the capability gate. Five interpretable configurations keep $\rho_{\mathrm{R@10}} > \rho_{\mathrm{R@1}}$ in both directions and in the aggregate: three CLAP checkpoints and OmniEmbed with and without its query prompt (Tables F.3 and F.4). ColQwen-Omni's bidirectional value is not interpreted, because its audio-to-text R@1 retention is negative; its directional values remain in the tables. The cohort has 975 clips, so absolute recalls are not compared with COCO.

F.4 Within-video segment assignment

This arm changes which annotated segment of a video each caption designates, with every score fixed; it has no relational diagnostic. ActivityNet Captions (Krishna et al., 2017) supplies 4,917 decoded validation videos, of which 3,899 have at least three annotated segments. In 39 of these videos, 54 of the 17,505 annotated spans (0.31%) extend past the decoded file and are trimmed to the available footage. DiDeMo (Hendricks et al., 2017) supplies 1,019 decoded test videos. Each description's annotator spans are reduced to the modal five-second span, with the earliest span breaking ties, and videos with fewer than three segments or only one distinct span are excluded, leaving 851. Segment clips use eight uniformly spaced interior frames within the annotated bounds; LanguageBind applies its own matching eight-frame sampler.

For a video with m eligible segments, the identity assigns caption j to segment j, reversal maps j to $m - 1 - j$, the cyclic shift moves every target by one, and the seeded derangement has no fixed point and differs from both reversal and shift. Segment R@1 is computed within the video's fixed $m \times m$ score matrix and averaged over videos; chance is the cohort mean of $1/m$. LanguageBind's whole-file decoder cannot process the long DiDeMo files within the shared resource contract, so that

Table F.3: AudioCaps Recall@1 for all six configurations. Original, hard negative, and reference recalls are percentages; ρ is dimensionless. A and T denote audio and text. Mean first averages the directional recalls. All six configurations pass the capability gate; the final column records retention interpretation. The ColQwen-Omni combined profile is not interpreted because of its negative audio-to-text R@1 retention; its directional results remain visible.

Configuration	Direction	Original	Hard neg.	Reference	ρ	Status
CLAP HTSAT unfused	A→T	31.18	5.13	0.0701	0.1626	Pass
	T→A	29.64	6.36	0.0769	0.2125	Pass
	Mean	30.41	5.74	0.0735	0.1869	Pass
CLAP HTSAT fused	A→T	32.00	5.54	0.0677	0.1713	Pass
	T→A	27.38	6.15	0.0721	0.2227	Pass
	Mean	29.69	5.85	0.0699	0.1950	Pass
Larger CLAP general	A→T	26.46	6.26	0.0753	0.2343	Pass
	T→A	17.85	5.54	0.0901	0.3068	Pass
	Mean	22.15	5.90	0.0827	0.2635	Pass
OmniEmbed, instructed	A→T	11.69	1.85	0.1006	0.1506	Pass
	T→A	36.62	5.54	0.0645	0.1498	Pass
	Mean	24.15	3.69	0.0825	0.1500	Pass
OmniEmbed, bare	A→T	21.23	3.38	0.0777	0.1563	Pass
	T→A	33.95	5.23	0.0761	0.1522	Pass
	Mean	27.59	4.31	0.0769	0.1538	Pass
ColQwen-Omni	A→T	1.23	0.10	0.1082	−0.0050	Negative
	T→A	20.92	5.64	0.0889	0.2665	Pass
	Mean	11.08	2.87	0.0986	—	Not interpreted

Table F.4: AudioCaps Recall@10 for all six configurations. Original, hard negative, and reference recalls are percentages; ρ is dimensionless. A and T denote audio and text. Mean first averages the directional recalls. All six configurations pass the capability gate; the final column records retention interpretation. The ColQwen-Omni combined profile is not interpreted because of its negative audio-to-text R@1 retention; its directional results remain visible.

Configuration	Direction	Original	Hard neg.	Reference	ρ	Status
CLAP HTSAT unfused	A→T	81.13	44.00	0.9207	0.5371	Pass
	T→A	72.72	42.46	0.9427	0.5785	Pass
	Mean	76.92	43.23	0.9317	0.5566	Pass
CLAP HTSAT fused	A→T	78.56	40.51	0.9367	0.5098	Pass
	T→A	68.41	39.08	0.9411	0.5652	Pass
	Mean	73.49	39.79	0.9389	0.5356	Pass
Larger CLAP general	A→T	76.00	43.28	0.9808	0.5639	Pass
	T→A	66.26	38.97	0.9483	0.5823	Pass
	Mean	71.13	41.13	0.9645	0.5724	Pass
OmniEmbed, instructed	A→T	31.90	14.15	0.9948	0.4258	Pass
	T→A	81.54	35.69	0.9351	0.4312	Pass
	Mean	56.72	24.92	0.9649	0.4297	Pass
OmniEmbed, bare	A→T	58.97	26.36	0.9559	0.4378	Pass
	T→A	76.62	34.36	0.9311	0.4417	Pass
	Mean	67.79	30.36	0.9435	0.4400	Pass
ColQwen-Omni	A→T	6.56	3.59	1.0553	0.4601	Pass
	T→A	61.54	37.74	0.9559	0.6072	Pass
	Mean	34.05	20.67	1.0056	—	Not interpreted

cell is absent rather than evaluated on a shortened cohort. This yields five ActivityNet cells and four DiDeMo cells (Table F.5).

Table F.5: Within-video segment R@1. The identity assignment exceeds chance in all nine cells, and every scrambled assignment is at or below chance. Reversal can keep a middle fixed point when m is odd, so it is not a complete derangement.

Model	Cohort	Identity	Reverse	Shift	Derange	Chance
ColQwen-Omni	ActivityNet	**0.4137**	0.2575	0.2296	0.2309	0.2783
LanguageBind	ActivityNet	**0.4138**	0.2614	0.2329	0.2251	0.2783
OmniEmbed	ActivityNet	**0.4518**	0.2499	0.2213	0.2120	0.2783
Qwen3-VL-Embedding-2B	ActivityNet	**0.4430**	0.2530	0.2204	0.2171	0.2783
Qwen3-VL-Embedding-8B	ActivityNet	**0.4560**	0.2504	0.2224	0.2087	0.2783
ColQwen-Omni	DiDeMo	**0.3331**	0.2476	0.2234	0.2228	0.2528
OmniEmbed	DiDeMo	**0.3606**	0.2415	0.2239	0.2099	0.2528
Qwen3-VL-Embedding-2B	DiDeMo	**0.3583**	0.2331	0.2273	0.2052	0.2528
Qwen3-VL-Embedding-8B	DiDeMo	**0.3765**	0.2361	0.2213	0.1985	0.2528

These results show that the deployed scorers support the annotated segment assignments; they do not isolate sequence reasoning.

F.5 Dense captions

This arm asks whether longer, denser captions change R@10 retention. DOCCI (Onoe et al., 2024) uses its official 5,000-image test split, each model's native image processor, and the model's own text window; cells with more than 5% truncation are flagged rather than removed. The prespecified outcome is `no_effect`: mean checkpoint-level changes in R@10 retention from COCO to DOCCI are -0.0292 across three long-context checkpoints and -0.0226 across the 16 short-context checkpoints, short of the frozen -0.05 threshold (Table F.6). The long-context group truncates 0–3.76% of DOCCI captions and the short-context group 94.40–100%. The group values are means of per-checkpoint changes, not retentions recomputed from averaged recalls. The COCO side of each change comes from the extension run's own snapshot of the panel (Appendix G.3). E5-V was also evaluated but belongs to neither group. Because the cohorts differ in images as well as captions, and many dense captions are truncated, this result does not identify a causal effect of caption density.

Table F.6: The two primary caption-granularity groups and their members. Each $\Delta\rho_{\mathrm{R@10}}$ is DOCCI minus COCO, with bidirectional recalls averaged before reference adjustment. Group values are arithmetic means of checkpoint-level changes. Truncation is the percentage of DOCCI captions truncated by that checkpoint's evaluated text pipeline.

Group	Checkpoint	$\Delta\rho_{\mathrm{R@10}}$	Truncated (%)
Long context	Jina-CLIP-v2	−0.0357	0.00
	LongCLIP-GmP ViT-L/14	−0.0293	3.76
	LongCLIP-L (Diffusers)	−0.0226	3.76
Group mean	3 checkpoints	−0.0292	—
Short context	CLIP B/16 224	−0.0304	94.40
	CLIP B/32 224	−0.0377	94.40
	CLIP L/14 224	−0.0172	94.40
	CLIP L/14 336	−0.0139	94.40
	SigLIP2 B/16 224	−0.0396	98.76
	SigLIP2 B/16 256	−0.0324	98.76
	SigLIP2 B/16 384	−0.0313	98.76
	SigLIP2 L/16 256	−0.0211	98.76
	SigLIP2 L/16 384	−0.0111	98.76
	SigLIP2 So400m/14 224	−0.0216	98.76
	SigLIP B/16 224	−0.0196	98.52
	SigLIP B/16 256	−0.0173	98.52
	SigLIP B/16 384	−0.0208	98.52
	SigLIP L/16 256	−0.0127	98.52
	SigLIP L/16 384	−0.0208	98.52
	SigLIP So400m/14 224	−0.0142	100.00
Group mean	16 checkpoints	−0.0226	—

G Reproduction and protocol records

This appendix records when each analysis was specified relative to its outcomes, the software environments, and where the code and results needed to check every reported quantity are found.

G.1 Study status

Table G.1 states, for each analysis, what was fixed before its outcomes were seen. These are protocol records kept within the project, not external preregistrations, and later analyses do not make earlier results prospective.

G.2 Environments

The research profiles target Python 3.11. The repository's `requirements.txt` pins the review and CPU-check dependencies, and `requirements-experiments.txt` pins the primary extraction versions for SigLIP, CLIP, SigLIP2, and the CKA study; `docs/ENVIRONMENT.md` documents separate environments for systems that need other library versions. Every result family also stores the software environment of its run. The CKA study used Python 3.11.2, NumPy 1.26.4, SciPy 1.16.3, PyTorch 2.6.0, Transformers 4.57.6, and Sentence Transformers 3.4.1. The six additional encoders used PyTorch 2.6.0, Sentence Transformers 5.4.0, and Transformers 4.57.6 (Jina-CLIP-v2 and both LongCLIP variants) or 5.5.0 (E5-V and both Qwen3-VL-Embedding models). The pins are not complete transitive lockfiles, so reconstruction is validated against stated tolerances and structural identities rather than byte equality.

G.3 Code, data, and artifacts

The code-and-results supplement contains the implementation, dependencies, protocols, the compact records behind every table and figure, and a file manifest. It omits full embedding caches, model weights, and source datasets, which are public inputs identified by revision in the protocols. Table G.2

Table G.1: What was fixed before each analysis's outcomes, and how each analysis relates to earlier observations.

Analysis	Fixed before its outcomes	Relation to earlier observations	Section
Exact controls and semantic assignment	Cohort and caption rules, strengths, metrics, external objectives, random generator	Original study	Appendices B and D.3
CLIP and SigLIP2 replications	Checkpoints, shared assignments, graph scales, validation rules	Follow-ups to the SigLIP result	Appendices B.4 and D.3
Rebuilt cohorts; cost-matched alternatives	Reconstruction and allocation rules, before each run	Sensitivity analyses of existing results	Appendix D.6
16-checkpoint panel	Roster, ranking and inversion rules	Three rows already known	Appendices D.2 and D.6.4
Six-encoder agreement follow-up	Cohort, saved assignments, graph scales, reporting rule, before agreement outcomes	Post hoc; retrieval outcomes already known	Appendix D.2
CKA reconstruction	Source audit and reconstruction choices	Published shuffle values known; semantic outcomes unseen	Appendix C.2
CKA alternatives, matched preprocessing, conditional ranks	Construction gates, preprocessing, summary rules	After the original CKA and rank results	Appendices C.4, C.5 and D.4
Annotation-only CKA and kernel response	MPNet-free tie rule; assignment saved before evaluation	After the original results	Appendix C.3
Linear and row-normalized fidelity comparison	Reuse of saved assignments; no construction or selection	Post hoc (2026-09-25)	Appendix C.5
QAP audit	Problems, starts, feasibility and evaluation rules	Follow-up of the published procedure	Appendix C.6
Flickr30k selection	Entire protocol before extraction; validation choices saved before test	Prospective within the project, after the COCO results	Appendices E.1 and E.3
COCO selection	Partition before half-specific metrics	Retrospective	Appendix E.4
Transfer, sign reversal, oracle diagnostics	Subsets, scale grid, and validation cases before new outcomes	Post hoc	Appendix E.5
Operator and modality extensions	Rosters, gates, and analysis rules before the final runs	Cross-setting synthesis is descriptive	Appendix F

maps each result family to its compact records and to the additional inputs that a full recomputation requires.

Table G.2: Result families. Compact records reproduce every reported number; full recomputation additionally needs the inputs in the last column.

ID	Result family	Compact records included	Additional full inputs
A1	Exact controls	Protocol, implementation, result report	Embedding caches, control arrays
A2	Semantic assignment (SigLIP)	Assignment and metric code, protocol, report	Assignments, graphs, ranks
A3	CLIP replication	Protocol, code, directional and control report	CLIP caches, full arrays
A4	Rebuilt cohorts	Reconstruction and validation code, all-25 report	Retained-cohort graphs, assignments
A5	Cost-matched alternatives	Frontier and ensemble code, protocols, reports	Costs, 128 assignments, rank arrays
A6	SigLIP2 replication	Protocol, code, reference and cohort reports	SigLIP2 caches, full arrays
A7	16-checkpoint panel	Canonical panel record, profile and ranking report	Embeddings, graphs, query-level arrays
A8	Late interaction	Final records, roster and scoring code	Model caches, score matrices
A9	Video and temporal	Final records, cohort and sampling code	Videos, segment representations
A10	Audio	All six configuration records, gates, code	Audio files, representations
A11	Dense captions	Per-model records, group analysis, truncation	DOCCI images, embedding caches
A12	CKA reconstruction	Source audit, protocol, all curve points, validation	Image and caption-group representations
A13	Diagnostic comparison	Diagnostic records and implementation	Saved representations, assignments
A14	COCO full-cohort selection	Original metrics, selections, report	Embeddings

ID	Result family	Compact records included	Additional full inputs
A15	COCO split selection	Split indices, half-specific metrics, choices, gaps	Checkpoint embedding caches
A16	QAP audit	Input and feasibility manifests, outcomes, code	Matching inputs, 1,500 permutations
A17	CKA alternatives and matched preprocessing	Construction records, all 128 CKA values, six preprocessing cells, conditional ranks	Assignment, feature, and rank arrays
A18	Flickr30k selection	Protocol, split and caption manifest, validation choices, test metrics, query-level successes, intervals, code	Images, embedding caches
A19	Annotation-only CKA and transfer	Specification, annotation-only assignment and categories, kernel responses, object overlap, transfer	A12 inputs, A1/A2 retrieval inputs
A20	Sign-reversal follow-up	Protocol, subset indices, metric and correlation tables, decomposition, validation	A7 and A18 embedding caches
A21	Six-encoder agreement and 22-checkpoint summary	Cohort, saved assignments, per-model records and arrays, extraction records, combined summary	Six embedding caches; A7 arrays
A22	Linear and row-normalized CKA	All variant values, random draws, anchor checks	A12 and A17 arrays

Locations and commands. Reports for A1–A7 are under `experiments/siglip_coco_fingerprint/`; the compact A7 record is `artifacts/checkpoint_panel/`. A8–A11 and A14–A15 are under `artifacts/extensions/`, whose index identifies superseded records. A12, A13, A16, A17, and A22 are under `artifacts/published_cka_usefulness/`, A18 under `artifacts/flickr30k_selection/`, A19 under `artifacts/published_cka_usefulness/mechanism_transfer/`, A20 under `artifacts/selection_sign_flip/`, and A21 under `artifacts/global_graph_retention/`. The oracle, Rel@1, weighted-rank, and model-omission diagnostics are derived from A15 and A18 and stored in `artifacts/selection_diagnostics/`. From the repository root, the combined appendix tables and relabeled figures are regenerated from included records by

```
python paper/iclr2027-appendix/make_tables.py
python paper/iclr2027-appendix/make_figures.py
```

and the combined 22-checkpoint summary and the linear/normalized CKA comparison are recomputed, given their external arrays, by `experiments/global_graph_retention/panel_summary.py` and `experiments/published_cka_usefulness/fidelity_variants.py`. Each experiment directory's README gives its commands.

Validation. The recorded checks cover ordered cohorts, finite representations, native-score reconstruction, intervention identities, bijection and exclusion constraints, tie rules, and agreement between full arrays and reported outcomes; extension checks cover capability gates, directions, and temporal counts. Where only compact records are included, they can be checked against the reported numbers, but the omitted embeddings cannot be regenerated from them. Correctness is assessed through structural identities and scientific outcomes rather than file digests.