# SPARSE-LANDMARK NON-RIGID ALIGNMENT OF SPECTRAL POINT SETS FOR CROSS-MODAL RETRIEVAL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Cross-modal retrieval typically learns shared representations from abundant paired data. We study a complementary regime in which unimodal representations are independently constructed and kept fixed, while reliable cross-modal correspondences are scarce. The resulting challenge is therefore not representation learning, but sparse registration between two latent spaces. Our key observation is that independently learned spectral spaces can retain non-random semantic neighborhood agreement despite occupying different coordinate frames. This motivates modeling their mismatch as a dominant global coordinate discrepancy followed by residual, position-dependent non-rigid deformation. We accordingly formulate sparse cross-modal alignment as a global-to-local geometric registration problem and propose Sparse-Landmark Non-Rigid Alignment (SLNA). Using verified pairs as sparse landmarks, SLNA first recovers a global distance-preserving coordinate relation and then allocates nonlinear capacity only to the remaining residuals through regularized corrections in complementary coordinate and angular geometries. Experiments on Flickr30k and MSCOCO show that SLNA consistently improves over rigid Procrustes alignment and also outperforms a direct RBF mapping under matched model-selection settings. Structural diagnostics reveal statistically significant cross-modal neighborhood agreement, while controlled-deformation experiments show that residual modeling recovers local distortions beyond a single global transformation. These results suggest that sparse cross-modal alignment benefits from treating global coordinate recovery and local deformation modeling as distinct estimation problems, rather than learning a single flexible mapping from sparse pairs.

## 1 INTRODUCTION

Image–text retrieval is a fundamental task in cross-modal retrieval and multimodal data management (Kalyan et al., 2024; Duan et al., 2025). Most existing methods learn a shared representation space from large-scale paired image–text data (Radford et al., 2021; Jia et al., 2021; Yamaguchi et al., 2025). Such approaches are effective when abundant and reliable cross-modal pairs are available. However, in many specialized applications, collecting images or texts is easier than obtaining reliable cross-modal pairs. In these cases, the main challenge is not learning new representations. It is aligning existing representations with only a small number of trusted correspondences. This naturally raises a new question: how can we align two independently learned representation spaces without updating the original encoders?

A key observation is that independently learned representations may still preserve similar semantic structures, even though their coordinate systems are different. For example, spectral representations capture local neighborhood structures within each modality (Shaham et al., 2018; Yacobi et al., 2025). However, since the two spaces are learned separately, they do not share the same coordinate system. As illustrated in Figure 1(a), corresponding samples may have similar local structures but different geometric positions. Therefore, sparse cross-modal alignment is not only a problem of finding semantic correspondences. It is also a problem of recovering the geometric relationship between two latent spaces. This geometric perspective suggests that different types of mismatch need to be handled at different spatial scales.



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.

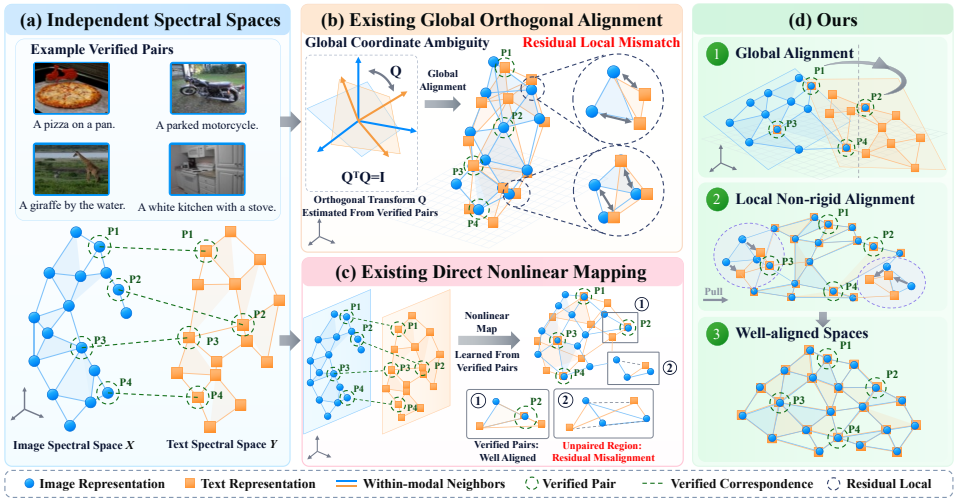



Figure 1: **Motivation for multi-scale geometric alignment.** (a) Independently learned image and text spectral spaces preserve related semantic structure but occupy different coordinate frames. (b) Existing global orthogonal alignment resolves the dominant coordinate mismatch but leaves local residual deviations. (c) Existing direct nonlinear mapping jointly models global and local mismatch from sparse verified pairs. (d) SLNA explicitly separates global alignment from local non-rigid correction.

However, existing alignment strategies do not explicitly model this geometric mismatch. Orthogonal Procrustes alignment estimates a global transformation from sparse correspondences (Schönemann, 1966). It is stable with limited supervision, but it assumes that one transformation can explain all samples. As shown in Figure 1(b), this assumption fails when different semantic regions contain different deviations. At the other extreme, nonlinear mappings provide more flexibility (Yacobi et al., 2025; Xu et al., 2022; Du et al., 2020). As illustrated in Figure 1(c), such methods try to learn the global transformation and local deformation at the same time. However, under sparse supervision, these two factors are difficult to separate. As a result, the learned mapping can easily overfit the limited correspondences.

Based on this observation, we argue that sparse cross-modal alignment should be viewed as a multi-scale geometric registration problem. Specifically, the mismatch between independently learned latent spaces contains two distinct components: (1) a global mismatch that determines the relative orientation and coordinate frame of the two spaces, and (2) local deformation that describes residual, position-dependent deviations after global alignment. Importantly, these components require different modeling strategies. The global mismatch can be estimated reliably from sparse correspondences, whereas local deformation is more complex and is better modeled after the global mismatch has been corrected. In other words, sparse correspondences are sufficient to estimate a stable global structure, while local flexibility should be reserved for modeling the remaining residual deformation. Following this, we propose **Sparse-Landmark Non-Rigid Alignment (SLNA)**, as illustrated in Figure 1(d). SLNA first estimates a global transformation using sparse verified correspondences as landmarks. It then learns a regularized residual model to correct local deformation after global alignment. This decomposition allows sparse landmarks to provide stable global information while the residual model focuses only on remaining local errors.

Our contributions are summarized as follows:

- We identify sparse cross-modal alignment as a multi-scale geometric mismatch problem, where independently learned representations require both global coordinate recovery and local deformation correction.
- We propose Sparse-Landmark Non-Rigid Alignment (SLNA), which decomposes sparse alignment into landmark-based global alignment and regularized local residual modeling without updating the original encoders.

- We evaluate SLNA on Flickr30k and MSCOCO through retrieval experiments, structural analyses, and controlled deformation studies, showing that independently learned spaces retain alignable structure and that global–local geometric decomposition improves sparse cross-modal alignment.

## 2 Related Work

**Cross-Modal Retrieval.** Existing cross-modal retrieval methods typically learn a shared representation space to establish semantic associations across modalities. Image–text contrastive models such as CLIP rely on large-scale paired data to learn cross-modal representations (Radford et al., 2021; Feng et al., 2025). CCA and Deep CCA align different modalities by modeling cross-view correlations (Andrew et al., 2013), while CSA further improves cross-modal matching through weighted canonical similarities (Li et al., 2025). Other approaches reduce modality discrepancy using MMD, adversarial learning, or optimal transport (Song & Soleymani, 2019; Li et al., 2018; Han et al., 2024). Despite their effectiveness, these methods generally rely on sufficient paired supervision. When only a small number of verified image–text pairs are available, the limited supervision makes it difficult to learn stable shared representations or complex cross-modal mappings, leaving the sparse-pair setting insufficiently addressed.

**Sparse-Pair Alignment.** Under sparse paired supervision, existing methods align independently learned modality-specific representations using a limited number of verified pairs. Orthogonal Procrustes provides a simple global alignment (Schönemann, 1966), while nonlinear mappings offer greater flexibility for modeling cross-modal discrepancies (Yacobi et al., 2025; Xu et al., 2022; Du et al., 2020). SUE further reduces the dependence on large-scale paired data by aligning independently learned spectral representations with sparse supervision (Yacobi et al., 2025). However, these methods do not explicitly model the geometric relationship between the two latent spaces, making it difficult to jointly account for global coordinate mismatch and local non-rigid variation. We therefore formulate sparse-pair alignment as a global-to-local geometric registration problem, recovering the global coordinate relation before correcting local non-rigid residuals with coordinate and angular information.

## 3 Method

In this section, we introduce **Sparse-Landmark Non-Rigid Alignment (SLNA)**, as illustrated in Figure 2. We formulate the problem and spectral representations in Section 3.1, introduce landmark-based global alignment in Section 3.2, present local non-rigid alignment and the optional posterior correction in Section 3.3, and finally describe bidirectional retrieval in Section 3.4.

### 3.1 Problem Formulation and Spectral Representation

Given an image set $\mathcal{I} = \{I_i\}_{i=1}^{N_I}$ and a text set $\mathcal{T} = \{T_j\}_{j=1}^{N_T}$, we extract their features using fixed encoders $E_I$ and $E_T$. This gives $\mathbf{u}_i^I = E_I(I_i) \in \mathbb{R}^{d_I}$ for images and $\mathbf{u}_j^T = E_T(T_j) \in \mathbb{R}^{d_T}$ for texts. Two independent SpectralNet encoders (Shaham et al., 2018) then map the features into $d$-dimensional spectral spaces: $\mathbf{x}_i = h_I(\mathbf{u}_i^I) \in \mathbb{R}^d$ and $\mathbf{y}_j = h_T(\mathbf{u}_j^T) \in \mathbb{R}^d$. Each spectral encoder approximates low-frequency functions of its own unimodal neighborhood graph. We stack all spectral coordinates into $\mathbf{X} \in \mathbb{R}^{N_I \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N_T \times d}$. These representations form semantic point sets rather than physical shapes. We use $d = 30$ throughout.

The available supervision contains only $m$ verified image–text pairs, $\mathcal{A} = (a_r^I, a_r^T)_{r=1}^m$. Each pair identifies the indices of a matched image and text, and is treated as a landmark. All other cross-modal correspondences in the alignment pool are hidden. Our goal is to learn cross-modal mappings from these sparse landmarks while keeping the image, text, and spectral encoders fixed. We use row vectors throughout and write $\mathrm{n}(\mathbf{z}) = \mathbf{z}/\|\mathbf{z}\|_2$ for unit normalization. The source-to-target construction is applied independently with the modalities exchanged for the reverse direction.

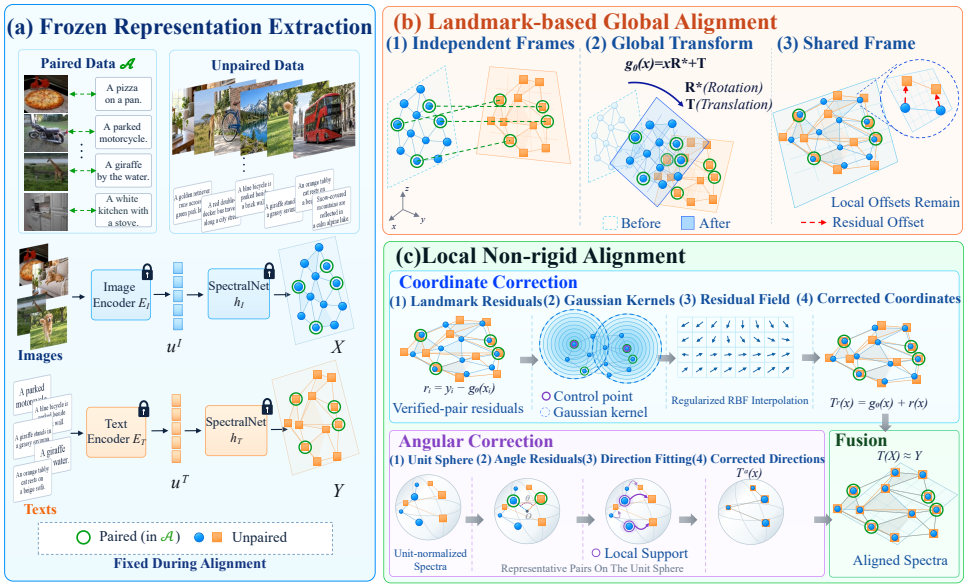



Figure 2: **Overview of the proposed SLNA framework.** (a) Spectral point-set construction and sparse verified landmarks. (b) Landmark-based global alignment estimates the dominant coordinate relation between the two spectral spaces. (c) Local non-rigid alignment models the remaining mismatch through coordinate-space and angular residual corrections and fuses them into SLNA; the optional SLNA-soft further refines the frozen mapping using posterior information from unverified candidates. The alignment is fitted independently in both directions for bidirectional retrieval.

### 3.2 Landmark-Based Global Alignment

To remove the dominant global coordinate mismatch, we first estimate a distance-preserving global map using the verified image–text pairs as sparse landmarks. Let $\mathbf{X}_A \in \mathbb{R}^{m\times d}$ and $\mathbf{Y}_A \in \mathbb{R}^{m\times d}$ denote the spectral coordinates of the $m$ verified image and text landmarks, respectively. We first compute the landmark means as $\boldsymbol{\mu}_x = \frac{1}{m}\sum_{r=1}^{m} \mathbf{x}_{a_r^I}$ and $\boldsymbol{\mu}_y = \frac{1}{m}\sum_{r=1}^{m} \mathbf{y}_{a_r^T}$, where $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y \in \mathbb{R}^{1\times d}$ denote the centers of the image and text landmarks. We then center the two landmark sets as $\mathbf{X}_c = \mathbf{X}_A - \mathbf{1}\boldsymbol{\mu}_x$ and $\mathbf{Y}_c = \mathbf{Y}_A - \mathbf{1}\boldsymbol{\mu}_y$, where $\mathbf{1} \in \mathbb{R}^{m\times 1}$ is the all-ones vector. Based on the centered landmark coordinates, we estimate the global rotation between the two spectral spaces using the Orthogonal Procrustes objective (Schönemann, 1966):

$$\mathbf{R}^\star = \underset{\mathbf{R}\in \mathrm{SO}(d)}{\arg\min} \|\mathbf{X}_c\mathbf{R} - \mathbf{Y}_c\|_F^2, \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{d\times d}$ is the rotation matrix to be estimated, $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathrm{SO}(d) = \{\mathbf{R} \mid \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \det(\mathbf{R}) = 1\}$ denotes the $d$-dimensional special orthogonal group. The optimal matrix $\mathbf{R}^\star$ minimizes the overall alignment error between the two centered landmark sets. Given $\mathbf{R}^\star$, the globally aligned representation of an arbitrary source-modality spectral point $\mathbf{z} \in \mathbb{R}^d$ is defined as

$$g_0(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_x)\mathbf{R}^\star + \boldsymbol{\mu}_y. \tag{2}$$

The map first translates the source representation to the source landmark center, applies the global rotation $\mathbf{R}^\star$, and then translates it to the target landmark center. Since $\mathbf{R}^\star$ is orthogonal, this transformation preserves Euclidean distances within the source modality while placing it in the target coordinate frame. Appendix A.1 gives the closed-form solution for $\mathbf{R}^\star$.

### 3.3 Local Non-Rigid Alignment

**Coordinate-Space Residual Modeling.** A single global alignment cannot correct local offsets that vary across semantic regions. We therefore further model the residuals after global alignment using Gaussian radial basis functions (RBFs). Specifically, we select $M \le m$ landmarks from the globally

aligned source landmarks by farthest-point sampling and use them as control points $\{\mathbf{c}_\ell\}_{\ell=1}^{M}$. For a spectral-space location $\mathbf{s} \in \mathbb{R}^d$, its Gaussian feature with respect to the $\ell$-th control point is defined as $\phi_\ell(\mathbf{s}) = \exp[-\|\mathbf{s}-\mathbf{c}_\ell\|_2^2/(2\sigma^2)]$, where $\sigma$ is scaled according to the median inter-control distance and controls the local range of influence.

For the $r$-th verified image–text landmark pair, the residual after global alignment is $\mathbf{y}_{a_r^T} - g_0(\mathbf{x}_{a_r^I})$. We collect the RBF responses of all landmarks into $\mathbf{\Phi} \in \mathbb{R}^{m\times M}$, where $\Phi_{r\ell} = \phi_\ell(g_0(\mathbf{x}_{a_r^I}))$, and stack the corresponding residuals row-wise into $\mathbf{\Delta} \in \mathbb{R}^{m\times d}$. The residual-field coefficients are learned by the regularized objective

$$\mathbf{W}^\star = \underset{\mathbf{W}\in\mathbb{R}^{M\times d}}{\arg\min} \|\mathbf{\Phi}\mathbf{W} - \mathbf{\Delta}\|_F^2 + \alpha\|\mathbf{W}\|_F^2, \tag{3}$$

where $\mathbf{W}$ contains the residual-field coefficients. The first term fits the landmark residuals, while $\alpha > 0$ regularizes the field. Since nearby locations produce similar Gaussian responses, the resulting corrections vary smoothly across local regions. The closed-form solution is given in Appendix A.1.

The learned residual field is then added to the global map $g_0$. For an arbitrary source-modality spectral representation $\mathbf{z} \in \mathbb{R}^d$, the coordinate-branch prediction is

$$T^r(\mathbf{z}) = g_0(\mathbf{z}) + \rho\,\boldsymbol{\phi}(g_0(\mathbf{z}))\mathbf{W}^\star, \tag{4}$$

where $\boldsymbol{\phi}(\mathbf{s}) = [\phi_1(\mathbf{s}), \ldots, \phi_M(\mathbf{s})]$ denotes the Gaussian responses of $\mathbf{s}$ to all control points, and $\rho$ controls the magnitude of the local correction. Setting $\rho = 0$ recovers global alignment.

**Angular Residual Modeling.** The raw-coordinate branch models local residuals directly in the spectral coordinate space, so its regression neighborhoods depend on both vector magnitude and direction. Since the norms of cross-modal spectral representations do not carry equally reliable semantic information, we further introduce a complementary angular branch that models local offsets only in directional geometry.

Specifically, the globally aligned source representation is normalized as $\mathbf{a} = \mathrm{n}(g_0(\mathbf{z}))$, while the corresponding target representation is normalized as $\mathbf{b} = \mathrm{n}(\mathbf{y})$. The angular branch constructs RBF neighborhoods from the normalized representations and uses either the chord displacement $\mathbf{b} - \mathbf{a}$ or the spherical tangent vector $\mathrm{Log}_{\mathbf{a}}(\mathbf{b})$ as the directional residual. These targets are fitted with the same ridge-regression objective (Hoerl & Kennard, 2000). For chord displacements, the predicted residual is added to $\mathbf{a}$ and then normalized; for tangent vectors, the prediction is first projected onto the tangent space and then mapped back to the unit sphere through the spherical exponential map. We denote the resulting unit-direction prediction by $T^a(\mathbf{z})$.

**Coordinate–Angular Fusion.** To jointly exploit coordinate and directional information, we fuse the raw-coordinate branch $T^r(\mathbf{z})$ and the angular branch $T^a(\mathbf{z})$ to obtain the final SLNA representation:

$$T_A(\mathbf{z}) = \mathrm{n}[\beta_{\mathrm{raw}}\,\mathrm{n}(T^r(\mathbf{z})) + (1-\beta_{\mathrm{raw}})T^a(\mathbf{z})]\,, \tag{5}$$

where $\beta_{\mathrm{raw}} \in [0,1]$ controls the fusion between the coordinate-based and direction-based predictions. The raw-coordinate branch preserves both magnitude and directional information during residual fitting, whereas the angular branch focuses only on normalized directional relations. Appendix A.2 provides the spherical mappings and numerical details.

**Optional posterior residual correction.** SLNA relies only on verified landmarks, so its local correction can be limited in regions with sparse landmark coverage. As an optional extension, we introduce **SLNA-soft** to exploit additional spatial information from unverified candidates while keeping the fitted SLNA model fixed. For a mapped source candidate $\mathbf{a}_i = T_A(\mathbf{x}_i)$ and normalized target candidates $\mathbf{b}_j = \mathrm{n}(\mathbf{y}_j)$, we construct a posterior barycentric target $\widetilde{\mathbf{b}}_i = \sum_{j=1}^{n_T} \pi_{ij}\mathbf{b}_j$, where $\pi_{ij} = P_{ij}/\sum_k P_{ik}$ denotes the normalized correspondence weight. The correspondence matrix $\mathbf{P}$ is obtained from temperature-scaled cross-modal similarities using row normalization or balanced Sinkhorn (Cuturi, 2013). Before regression, the target is rescaled as $\widehat{\mathbf{b}}_i = \|\mathbf{a}_i\|_2\,\mathrm{n}(\widetilde{\mathbf{b}}_i)$ to preserve the scale of the SLNA output. Using a posterior barycenter avoids forcing an uncertain candidate to a single target and provides a smoother estimate of the remaining local displacement.

Rather than refitting $T_A$, we fit a second low-weight RBF residual field in the frozen SLNA space. Using $M_s$ control points, its coefficients are obtained by

$$\mathbf{V}^\star = \underset{\mathbf{V}\in\mathbb{R}^{M_s\times d}}{\arg\min} \left\| \mathbf{D}^{1/2}(\mathbf{\Psi}\mathbf{V} - \mathbf{E}) \right\|_F^2 + \alpha_s \|\mathbf{V}\|_F^2, \tag{6}$$

where $\mathbf{\Psi}$ contains the RBF features of verified landmarks and candidate samples. The target matrix $\mathbf{E}$ assigns zero additional correction to verified landmarks and uses $\widehat{\mathbf{b}}_i - \mathbf{a}_i$ for candidate samples. The diagonal matrix $\mathbf{D}$ controls their relative contributions, with $\eta$ determining the influence of unverified candidates relative to reliable landmarks.

The resulting optional mapping is

$$T_S(\mathbf{z}) = T_A(\mathbf{z}) + \rho_s\, r_s(T_A(\mathbf{z})), \qquad r_s(\mathbf{a}) = \boldsymbol{\psi}(\mathbf{a})\mathbf{V}^\star, \tag{7}$$

where $\rho_s$ controls the magnitude of the additional correction. The candidate posterior is used only when fitting this residual field offline and is not recomputed for individual retrieval queries. Further details of the posterior construction and weighted regression are provided in Appendix A.3.

### 3.4 Bidirectional Retrieval and Computational Cost

The image-to-text $T_{I\to T}$ and text-to-image $T_{T\to I}$ mappings use different modalities as their source spectral spaces. Their local structures and residual distributions can therefore differ. We fit the two directions independently and combine their cosine similarities for retrieval:

$$S_{ij} = \omega \left\langle \mathrm{n}(T_{I\to T}(\mathbf{x}_i)), \mathrm{n}(\mathbf{y}_j) \right\rangle + (1-\omega) \left\langle \mathrm{n}(\mathbf{x}_i), \mathrm{n}(T_{T\to I}(\mathbf{y}_j)) \right\rangle, \tag{8}$$

where $\omega \in [0,1]$ balances the two directions. The similarity matrix $\mathbf{S}$ is used to obtain the image-to-text and text-to-image rankings. This fusion is applied only during retrieval and does not affect the spectral-space alignment. An equivalent weighted-concatenation form is given in Appendix A.2.

For fixed spectral dimension $d$, fitting one RBF branch costs $O(mM^2 + M^3)$, and mapping one sample costs $O(Md)$. The two branches and two mapping directions introduce only constant-factor overhead. Gallery representations can be precomputed after fitting, leaving query mapping and cosine similarity as the main online costs. Runtime results are reported in Section E.

## 4 Experiments

In this section, we evaluate retrieval performance and the contribution of each alignment component, then examine the structural assumptions, landmark coverage, geometric distortion, and computational cost.

### 4.1 Experimental Setup

**Data and representations.** We evaluate on Flickr30k (Young et al., 2014) and MSCOCO (Lin et al., 2014). Unless stated otherwise, image and text features are extracted using DINOv2 ViT-B/14 (Oquab et al., 2024) and all-MiniLM-L6-v2 (Reimers & Gurevych, 2019), respectively, and are further mapped to 30-dimensional spectral representations using two independently trained SpectralNet encoders (Shaham et al., 2018). Both the feature encoders and spectral encoders remain fixed throughout the alignment process.

**Evaluation protocols.** Unless otherwise stated, the main experiments use **300 verified pairs** as landmarks for fitting the cross-modal alignment. For each dataset, 400 samples are reserved as an independent evaluation set and are excluded from landmark sampling and model selection. Following SUE (Yacobi et al., 2025), we use **100 verified pairs for MSCOCO** and **500 verified pairs for Flickr30k** in Section 4.2. Experiments that explicitly study landmark-budget sensitivity use the pair counts specified in their corresponding sections.

**Selection and metrics.** For the main experiments, model configurations are selected by three-fold cross-validation over the 300 landmarks using $J = 0.50\mathrm{R@10} + 0.25\mathrm{R@1} + 0.25\mathrm{mAP}$, after which the selected configuration is fixed. Direct RBF and SLNA share the same search space and search budget, while the capacity-matched comparison additionally shares the selected model capacities

and fusion parameters. We use five shared landmark seeds (344, 9736, 7719, 4466, and 6030) and report the bidirectional mean of R@1, R@5, R@10, and mAP. The main baselines include bidirectional Procrustes and Direct RBF. Additional protocol and evaluation details are provided in Appendix A.4 and Appendix B.

### 4.2 Comparison with Existing Sparse-Pair Methods

Table 1: **Comparison with existing sparse-pair methods on bidirectional image–text retrieval using random seed 4466. Bold** values indicate the best performance for each metric, while underlined values indicate the second-best performance within each dataset and pair-budget setting.

| Dataset | Pairs | Method | Image → Text | | | | Text → Image | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | R@1 ↑ | R@5 ↑ | R@10 ↑ | mAP ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | mAP ↑ |
| MSCOCO | 100 | Contrastive | 1.50 | 8.50 | 13.00 | 5.00 | 0.80 | 5.80 | 12.20 | 4.90 |
| | | CSA | 0.00 | 1.25 | 3.00 | 1.48 | 0.00 | 1.00 | 2.25 | 1.38 |
| | | SUE | 5.75 | 21.50 | 34.25 | 9.72 | 5.25 | 18.25 | 33.25 | 9.80 |
| | | SLNA | **11.75** | **44.50** | 65.00 | **27.35** | **11.50** | **45.00** | **68.25** | 27.25 |
| | | SLNA-soft | 10.86 | 43.75 | **65.75** | 26.77 | **11.50** | **45.00** | 68.00 | **27.40** |
| | 300 | Contrastive | 2.50 | 11.50 | 19.75 | 8.54 | 2.50 | 12.50 | 18.50 | 8.78 |
| | | CSA | 0.00 | 0.50 | 2.50 | 1.24 | 0.00 | 0.50 | 1.00 | 1.19 |
| | | SUE | 8.25 | 33.25 | 52.00 | 20.60 | 6.75 | 33.75 | 51.75 | 19.60 |
| | | SLNA | **13.50** | **50.50** | **74.75** | **30.65** | 14.00 | **54.00** | **76.75** | 31.89 |
| | | SLNA-soft | 13.25 | 49.75 | 74.50 | 30.53 | **14.25** | **54.00** | 76.50 | **32.22** |
| Flickr30k | 300 | Contrastive | 1.75 | 7.50 | 11.00 | 5.60 | 2.00 | 6.75 | 12.75 | 5.71 |
| | | CSA | 0.50 | 1.50 | 2.75 | 1.88 | 0.75 | 2.50 | 3.50 | 2.21 |
| | | SUE | 4.00 | 19.25 | 28.25 | 12.60 | 4.00 | 19.25 | 31.25 | 12.63 |
| | | SLNA | 8.25 | 26.75 | 41.75 | 18.86 | **9.75** | **26.75** | 38.75 | **19.33** |
| | | SLNA-soft | **9.00** | **29.75** | **42.25** | **19.73** | 9.25 | 26.25 | **39.50** | 18.85 |
| | 500 | Contrastive | 3.00 | 9.50 | 16.20 | 5.80 | 2.50 | 9.80 | 15.00 | 5.80 |
| | | CSA | 0.25 | 1.25 | 2.50 | 1.67 | 0.25 | 0.75 | 2.75 | 1.67 |
| | | SUE | 4.25 | 19.75 | 32.00 | 8.91 | 5.75 | 22.00 | 32.75 | 8.07 |
| | | SLNA | 9.25 | 29.50 | **43.50** | 20.43 | **10.00** | **30.50** | **43.50** | **21.39** |
| | | SLNA-soft | **9.75** | **29.75** | 43.00 | **20.51** | 9.50 | 30.25 | 43.00 | 21.00 |

Table 1 compares SLNA with existing sparse-pair methods. Following SUE, we use 100 verified pairs on MSCOCO and 500 on Flickr30k. Under these pair budgets, SLNA consistently outperforms the published SUE results across all directional retrieval metrics, improving mAP by **181.4%/178.1%** on MSCOCO and **129.3%/165.1%** on Flickr30k for image-to-text/text-to-image retrieval, respectively. The optional SLNA-soft correction further improves image-to-text R@10 and text-to-image mAP on MSCOCO, and image-to-text R@1, R@5, and mAP on Flickr30k.

We additionally use a common budget of 300 verified pairs on both datasets. Under this setting, SLNA-soft improves text-to-image R@1 and mAP on MSCOCO. On Flickr30k, it improves all image-to-text metrics and further increases text-to-image R@10. The slight declines observed in some metrics may be attributed to uncertainty in the candidate correspondences used by the optional SLNA-soft correction, which can introduce small ranking perturbations across retrieval directions.

### 4.3 Component Ablation

Table 2 examines the contributions of global alignment, local residual modeling, and the optional posterior correction. Compared with Procrustes, SLNA improves R@10 by **6.48** points on Flickr30k and **10.15** points on MSCOCO, showing that global rigid alignment alone is insufficient to model the remaining local mismatch. Compared with Direct RBF, SLNA further improves R@10 by **0.65** and **0.75** points, respectively, supporting the explicit decomposition of global alignment and local residual correction. Adding the optional Soft correction further increases R@10 by **0.72** points on Flickr30k and **0.02** points on MSCOCO.

### 4.4 Structural Premise and Registration Mechanism

We conduct two diagnostic experiments to examine the structural premise and registration mechanism underlying SLNA. First, we test whether independently learned image and text spectral spaces

Table 2: **Component ablation on Flickr30k and MSCOCO.** Ablation of global alignment, local residual modeling, and the optional posterior correction. **Bold** values indicate the best performance within each block.

| Method | Global | RBF fit | Soft | R@1 ↑ | R@5 ↑ | R@10 ↑ | mAP ↑ |
|---|---|---|---|---|---|---|---|
| **Flickr30k** | | | | | | | |
| Procrustes | ✓ | – | – | $6.33 \pm 0.75$ | $22.15 \pm 1.30$ | $32.65 \pm 0.96$ | $14.98 \pm 0.76$ |
| Direct RBF | – | direct | – | $8.48 \pm 1.02$ | $26.48 \pm 1.28$ | $38.48 \pm 1.38$ | $18.25 \pm 1.00$ |
| SLNA | ✓ | residual | – | $8.40 \pm 1.07$ | $26.90 \pm 1.15$ | $39.13 \pm 1.68$ | $18.47 \pm 1.19$ |
| SLNA-soft | ✓ | residual | ✓ | $\mathbf{8.60 \pm 0.78}$ | $\mathbf{27.23 \pm 1.82}$ | $\mathbf{39.85 \pm 1.07}$ | $\mathbf{18.63 \pm 1.13}$ |
| **MSCOCO** | | | | | | | |
| Procrustes | ✓ | – | – | $9.55 \pm 0.55$ | $40.73 \pm 0.99$ | $64.88 \pm 1.31$ | $24.90 \pm 0.53$ |
| Direct RBF | – | direct | – | $12.98 \pm 0.86$ | $51.15 \pm 0.74$ | $74.28 \pm 0.85$ | $30.12 \pm 0.48$ |
| SLNA | ✓ | residual | – | $13.58 \pm 1.52$ | $\mathbf{51.98 \pm 0.85}$ | $75.03 \pm 1.38$ | $31.01 \pm 0.93$ |
| SLNA-soft | ✓ | residual | ✓ | $\mathbf{13.90 \pm 1.43}$ | $51.85 \pm 0.67$ | $\mathbf{75.05 \pm 1.26}$ | $\mathbf{31.19 \pm 0.91}$ |

retain sufficiently consistent local structures to support landmark-based alignment. Second, we construct a synthetic transformation with known ground truth to test whether local residual modeling can recover non-rigid deformation that cannot be captured by a single global rigid transformation.

Table 3: **Structural analysis and synthetic deformation recovery.** A: evaluates cross-modal structural consistency using normalized Frobenius distance (NFD), permutation testing, and intrinsic dimensionality. B: evaluates recovery of a known non-rigid deformation under different landmark budgets using RMSE and neighborhood retention.

**A. Cross-modal structure**

| Dataset | Alignment NFD ↓ | Shuffled NFD | $p$-value | TwoNN I/T |
|---|---|---|---|---|
| MSCOCO | 0.4698 | $0.9686 \pm 0.0035$ | 0.001 | 3.34/3.46 |
| Flickr30k | 0.7718 | $0.9685 \pm 0.0031$ | 0.001 | 5.21/3.23 |

**B. Synthetic non-rigid registration**

| Landmarks | Procrustes | | +RBF | |
|---|---|---|---|---|
| | RMSE ↓ | $k$NN ↑ | RMSE ↓ | $k$NN ↑ |
| 10 | $0.2947 \pm 0.0203$ | $0.8645 \pm 0.0145$ | $0.2075 \pm 0.0289$ | $0.8827 \pm 0.0107$ |
| 50 | $0.2648 \pm 0.0087$ | $0.8645 \pm 0.0142$ | $0.0906 \pm 0.0075$ | $0.9165 \pm 0.0039$ |
| 300 | $0.2612 \pm 0.0079$ | $0.8640 \pm 0.0143$ | $0.0322 \pm 0.0022$ | $0.9454 \pm 0.0020$ |

**Cross-modal structural consistency.** Following the structural diagnostic in SUE, we construct Gaussian-weighted $k = 15$ neighborhood graphs for the two modalities and quantify their structural discrepancy using the normalized Frobenius distance (NFD), $\mathrm{NFD}(\mathbf{P}_I, \mathbf{P}_T) = \|\mathbf{P}_I - \mathbf{P}_T\|_F / \|\mathbf{P}_I + \mathbf{P}_T\|_F$, where a smaller value indicates stronger local structural agreement. As shown in Table 3A, Alignment NFD measures the structural discrepancy under the verified image–text pair alignment, whereas Shuffled NFD provides a null baseline from 1,000 random permutations. The Alignment NFD is substantially lower than the Shuffled NFD on both MSCOCO (0.4698 vs. $0.9686 \pm 0.0035$) and Flickr30k (0.7718 vs. $0.9685 \pm 0.0031$), with permutation $p$-values of 0.001. TwoNN I/T estimates the intrinsic dimensions of the image and text spectral spaces, which remain well below the ambient dimension of 30. These results support non-random and locally low-dimensional cross-modal structural agreement. Additional details are provided in Appendix D.

**Recovery of local deformation.** Table 3B evaluates recovery of a known non-rigid deformation by comparing global Procrustes alignment with residual modeling. RMSE measures deformation-recovery error, while $k$NN retention measures preservation of the target local neighborhood. At 300 landmarks, the residual model reduces RMSE from 0.2612 to 0.0322 (87.7%) and increases neighborhood retention from 0.8640 to 0.9454. This shows that global rigid alignment alone cannot recover the local deformation, whereas the residual field can capture the remaining non-rigid mismatch when sufficient landmark coverage is available.

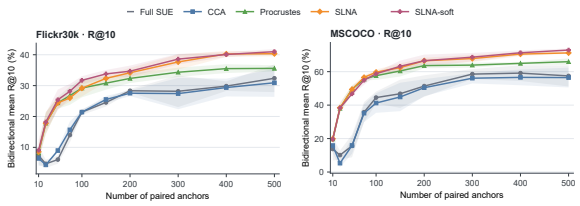



Figure 3: **Landmark efficiency.** Fixed-configuration R@10 under different landmark budgets. Results are three-seed means ± sample standard deviations. Full R@1/R@10 curves are provided in Appendix B.

| | Flickr30k | | MSCOCO | |
|---|---|---|---|---|
| Method | R@10 ↑ | Ret. ↑ | R@10 ↑ | Ret. ↑ |
| CCA | 30.92 | 0.384 | 56.46 | 0.628 |
| Procrustes | 35.58 | 1.000 | 65.96 | 1.000 |
| SLNA | 40.33 | 0.839 | 71.21 | 0.938 |
| SLNA-soft | 41.00 | 0.853 | 72.92 | 0.920 |

Table 4: **Retrieval performance and neighborhood preservation.** Ret. denotes the fraction of original 10-nearest neighbors (10-NN) retained after alignment.

### 4.5 Pair Efficiency and Geometry Preservation

**Landmark efficiency.** Figure 3 shows that R@10 improves rapidly as the number of verified landmarks increases, with the gain gradually saturating after approximately 100–200 landmarks. The advantage of non-rigid alignment also becomes clearer as landmark coverage increases, indicating that local residual modeling benefits from sufficient spatial supervision. We therefore use **300 verified landmarks** in the main experiments as a balance between alignment performance and annotation cost. Complete R@1/R@10 curves are in Appendix B.

**Geometry preservation.** Table 4 shows that Procrustes fully preserves the original neighborhood structure (Ret.=1.000), but achieves only 35.58/65.96 R@10 on Flickr30k/MSCOCO. SLNA improves R@10 to 40.33/71.21 while retaining 0.839/0.938 of the original 10-NN structure, and SLNA-soft further reaches 41.00/72.92 with retention of 0.853/0.920. These results indicate that SLNA improves retrieval through moderate local geometric adjustment while preserving most of the original neighborhood structure. Additional analyses are provided in Appendix G.

## 5 Limitations

SLNA relies on a small set of verified cross-modal correspondences, so its performance depends on landmark quality and coverage. The capacity-matched experiments support the benefit of the global–local decomposition, while gains from the optional SLNA-soft correction vary across datasets and retrieval directions. Our current evaluation uses cached representations and 400-sample independent evaluation sets, and therefore does not assess the full representation-learning pipeline on unseen data. In addition, the Flickr30k features lack an item-level manifest, the five-caption experiment is not an identity-matched Karpathy split (Karpathy & Fei-Fei, 2015), and several diagnostic studies use fixed configurations. Larger-scale and end-to-end evaluation is therefore left for future work.

## 6 Conclusion

In this paper, we study cross-modal retrieval under sparse verified correspondences from a geometric perspective. We view independently learned spectral spaces as structurally related point sets and use sparse reliable landmarks to guide global registration and local non-rigid correction. Based on this formulation, we propose Sparse-Landmark Non-Rigid Alignment (SLNA), with SLNA-soft as an optional refinement using unverified candidates. Results on Flickr30k and MSCOCO demonstrate improved retrieval over existing sparse-pair methods and direct nonlinear alignment.

## Reproducibility Statement

The released artifacts record split manifests, final anchor seeds, development folds, frozen configurations, run hashes, and method-level metrics. A release script verifies the artifacts and exports LaTeX-ready tables. Evaluation correspondences are used only for reporting retrieval metrics and specified diagnostics, not for fitting the reported alignment mappings.

## REFERENCES


Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255, 2013.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Youtian Du, Xue Wang, Yunbo Cui, Hang Wang, and Chang Su. Kernel-based mixture mapping for image and text association. *IEEE Transactions on Multimedia*, 22(2):365–379, 2020. doi: 10.1109/TMM.2019.2930336.

Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. Fuzzy multimodal learning for trusted cross-modal retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20747–20756, 2025. doi: 10.1109/CVPR52734.2025.01932.

Yuheng Feng, Changsong Wen, Zelin Peng, Li Jiaye, and Siyu Zhu. Retaining knowledge and enhancing long-text representations in clip through dual-teacher distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24895–24904, 2025. doi: 10.1109/CVPR52734.2025.02318.

Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26669–26678, 2024. doi: 10.1109/CVPR52733.2024.02519.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, February 2000. ISSN 0040-1706. doi: 10.2307/1271436. URL `https://doi.org/10.2307/1271436`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916, 2021.

T Pavan Kalyan, Piyush Singh Pasi, Sahil Nilesh Dharod, Azeem Azaz Motiwala, Preethi Jyothi, Aditi Chaudhary, and Krishna Srinivasan. WikiDO: A new benchmark evaluating cross-modal retrieval for vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 140812–140827. Curran Associates, Inc., 2024. doi: 10.52202/079017-4469.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015. doi: 10.1109/CVPR.2015.7298932.

Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4242–4251, 2018. doi: 10.1109/CVPR.2018.00446.

Po-han Li, Sandeep P. Chinchali, and Ufuk Topcu. CSA: Data-efficient mapping of unimodal features to multimodal features. In *International Conference on Learning Representations*, 2025.

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 *Computer Vision–ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin 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. *Transactions on Machine Learning Research*, 2024.

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*, volume 139 of *Proceedings of Machine Learning Research*, 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. doi: 10.18653/v1/D19-1410.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. SpectralNet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018.

Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, 2019.

Xing Xu, Kaiyi Lin, Yang Yang, Alan Hanjalic, and Heng Tao Shen. Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3030–3047, 2022. doi: 10.1109/TPAMI.2020.3045530.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1552.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4256–4266, June 2025.

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. doi: 10.1162/tacl_a_00166.

# Appendix

## Table of Contents


**A Additional Method Details** . . . . . **13**

A.1 Closed-Form Global and Residual Estimators . . . . . 13

A.2 Spherical Residuals and Retrieval Representations . . . . . 13

A.3 Posterior Construction and Weighted Regression . . . . . 13

A.4 Frozen Configurations and Fixed-Configuration Protocol . . . . . 14

**B Complete Retrieval Results** . . . . . **15**

**C Representation and Registration Decomposition** . . . . . **17**

**D Structural and Synthetic Diagnostics** . . . . . **19**

**E Robustness and Efficiency** . . . . . **20**

**F Correspondence Diagnostics** . . . . . **22**

**G Geometry and Qualitative Visualization** . . . . . **23**

**H CSA Diagnostic** . . . . . **30**

**I AI Use Statement** . . . . . **30**

# A Additional Method Details

## A.1 Closed-Form Global and Residual Estimators

The global objective in Equation 1 is solved by the proper-rotation Procrustes estimator,

$$\mathbf{R}^\star = \mathbf{U}\,\mathrm{diag}\big(1,\dots,1,\det(\mathbf{U}\mathbf{V}^\top)\big)\mathbf{V}^\top, \tag{9}$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{X}_c^\top\mathbf{Y}_c$. The determinant correction enforces $\mathbf{R}^\star \in \mathrm{SO}(d)$. The landmark means then restore the translation in $g_0$.

With the global map fixed, the strictly regularized objective in Equation 3 has the unique solution

$$\mathbf{W}^\star = (\mathbf{\Phi}^\top\mathbf{\Phi} + \alpha\mathbf{I}_M)^{-1}\mathbf{\Phi}^\top\mathbf{\Delta}, \tag{10}$$

where $\mathbf{\Phi} \in \mathbb{R}^{m\times M}$ is the landmark kernel matrix, $\mathbf{\Delta} \in \mathbb{R}^{m\times d}$ contains the verified residuals, and $\mathbf{I}_M$ is the identity matrix. In implementation, we solve the corresponding positive-definite linear system rather than forming the inverse.

## A.2 Spherical Residuals and Retrieval Representations

The angular branch represents displacements between unit vectors. For non-antipodal row vectors $\mathbf{a}$ and $\mathbf{b}$, the tangent target is defined by the spherical logarithm,

$$\mathrm{Log}_{\mathbf{a}}(\mathbf{b}) = \frac{\theta}{\sin\theta}\big(\mathbf{b} - (\mathbf{a}\mathbf{b}^\top)\mathbf{a}\big), \tag{11}$$

where $\theta = \arccos(\mathbf{a}\mathbf{b}^\top)$ is the angle between the two vectors. This target lies in the tangent space at $\mathbf{a}$. The chord alternative uses $\mathbf{b} - \mathbf{a}$ without converting it to a geodesic displacement.

Let $\boldsymbol{\xi}(\mathbf{a})$ denote the scaled RBF prediction. The tangent variant projects this prediction and reconstructs a unit vector through

$$\mathrm{Exp}_{\mathbf{a}}(\boldsymbol{\xi}_\perp) = \cos(\|\boldsymbol{\xi}_\perp\|_2)\mathbf{a} + \frac{\sin(\|\boldsymbol{\xi}_\perp\|_2)}{\|\boldsymbol{\xi}_\perp\|_2}\boldsymbol{\xi}_\perp, \tag{12}$$

where $\boldsymbol{\xi}_\perp = \boldsymbol{\xi}(\mathbf{a})(\mathbf{I}_d - \mathbf{a}^\top\mathbf{a})$ is the tangential component. The chord variant instead returns $\mathrm{n}(\mathbf{a} + \boldsymbol{\xi}(\mathbf{a}))$. Continuous limits give $\mathrm{Log}_{\mathbf{a}}(\mathbf{a}) = 0$ and $\mathrm{Exp}_{\mathbf{a}}(0) = \mathbf{a}$. For numerical stability, cosines are clipped to $[-1+10^{-7}, 1-10^{-7}]$ and tangent directions use their measured norm with a $10^{-10}$ threshold. Reconstructed outputs are normalized.

Both branch outputs are combined by the same normalized fusion operator,

$$\mathcal{F}_\beta(\mathbf{v}^r, \mathbf{v}^a) = \mathrm{n}[\beta\ \mathrm{n}(\mathbf{v}^r) + (1-\beta)\,\mathrm{n}(\mathbf{v}^a)]\,, \tag{13}$$

where $\beta = \beta_{\mathrm{raw}}$ and the superscripts indicate raw-coordinate and angular outputs. We apply this operator separately to the source and target of a fitted direction. For Anchor, both target branches reduce to the same normalized target. The matched Direct RBF uses the same selected kernel capacities and fusion weights but regresses target coordinates directly; normalizing its angular-kernel inputs does not change its raw-coordinate regression labels.

To implement Equation 8 as one cosine search, we concatenate the directional outputs with square-root weights:

$$\mathbf{z}^q = \big[\sqrt{\omega}\,\mathrm{n}(\mathbf{v}^{q,\rightarrow}),\ \sqrt{1-\omega}\,\mathrm{n}(\mathbf{v}^{q,\leftarrow})\big]\,, \tag{14}$$

where $q \in \{I, T\}$ denotes the modality and the arrows identify independently fitted directions. For the forward fit, the image is mapped and the text remains in its target frame; the reverse fit exchanges these roles. Each $\mathbf{v}$ is the corresponding fused Anchor or Soft output. The concatenated vectors have unit norm, so their inner product $\mathbf{z}_i^I(\mathbf{z}_j^T)^\top$ equals the score in Equation 8.

## A.3 Posterior Construction and Weighted Regression

The posterior balances feature agreement against optional landmark-relative structure. Its cost is

$$\mathbf{C} = \mathcal{N}\big(\mathbf{C}^{\mathrm{feat}} + \lambda_s\mathbf{C}^{\mathrm{struct}}\big)\,, \tag{15}$$

where $\lambda_s \geq 0$ controls the structural term and $\mathcal{N}$ divides a nonnegative matrix by the median of entries larger than $10^{-12}$. The divisor is bounded below by $10^{-12}$ and defaults to one when no entry

qualifies. The feature term is the separately normalized matrix of squared distances between unit source and target candidates. It is equivalent to median-normalized cosine distance because their common factor of two cancels.

To obtain landmark-relative structure, we select $L$ corresponding landmarks by farthest-point sampling and compute a response vector for each candidate. Its $\ell$th component is

$$q_\ell(\mathbf{v}; \mathbf{L}) = \exp\left(-\frac{\|\mathbf{v} - \mathbf{L}_{\ell,:}\|_2^2}{2\tau^2}\right), \tag{16}$$

where $\mathbf{L}$ contains the selected coordinates in the candidate's modality and $\tau$ is the median positive candidate-to-landmark distance pooled across both modalities. The source and target landmark matrices use the same verified identities. We normalize each response vector and define $\mathbf{C}^{\text{struct}}$ as the median-normalized squared-distance matrix between source and target responses. This construction uses known landmarks, not hidden candidate correspondences.

We convert the cost into affinities with $K_{ij} = \exp(-C_{ij}/\varepsilon)$, where $\varepsilon > 0$ controls posterior concentration. The row-normalized variant directly normalizes each row of $\mathbf{K}$. The balanced variant uses Sinkhorn scaling,

$$\mathbf{P} = \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v}), \tag{17}$$

where the scaling vectors enforce uniform row and column marginals. Rows of $\mathbf{P}$ are subsequently normalized to obtain the conditional correspondence weights $\pi_{ij}$ used to construct the posterior barycentric targets. The selected models restore each barycenter to its source norm before regression; near-zero source or barycenter norms fall back to the source vector. A fit is rejected if the required rescaling factor is non-finite or exceeds 10.

The incremental field in Equation 6 is solved by weighted ridge regression:

$$\mathbf{V}^\star = (\mathbf{\Psi}^\top \mathbf{D} \mathbf{\Psi} + \alpha_s \mathbf{I}_{M_s})^{-1} \mathbf{\Psi}^\top \mathbf{D} \mathbf{E}, \tag{18}$$

where $\mathbf{\Psi} \in \mathbb{R}^{(m+n)\times M_s}$ contains the second field's kernel features and $\mathbf{E} \in \mathbb{R}^{(m+n)\times d}$ stacks zero anchor targets followed by candidate residuals. The first $m$ diagonal entries of $\mathbf{D}$ equal $1/(1+\eta)$; the remaining $n$ entries equal $\eta m/[n(1+\eta)]$. This penalizes anchor drift without requiring it to vanish. The selected configurations do not use confidence pruning, residual clipping, or query support gating.

### A.4 Frozen Configurations and Fixed-Configuration Protocol

This section provides the configurations used in the main evaluation and diagnostic experiments. For the main experiments, model configurations are selected before final evaluation and then kept fixed. The component ablation in Table 2 uses the same model capacities and major configuration settings for Direct RBF and SLNA while preserving their different mapping formulations. The coordinate and angular branches follow the formulation in Section 3.3.

Table 5: **Frozen configurations for the main experiments.** Each tuple gives the ridge coefficient, bandwidth scale, and number of controls, $(\alpha, s, M)$; all denotes all 300 landmarks. The angular-target column gives the SLNA parameterization. Direct RBF uses the same numerical capacities and fusion weights but directly regresses target coordinates. All configurations use $\rho = 1$.

| Dataset | Raw $(\alpha, s, M)$ | Angular target | Angular $(\alpha, s, M)$ | $\beta_{\text{raw}}$ | $\omega$ |
|---|---|---|---|---|---|
| Flickr30k | $(.03, .75, 150)$ | chord | $(.03, 1.0, 100)$ | .75 | .50 |
| MSCOCO | $(.003, .84375, \text{all})$ | chord | $(.03, .50, 100)$ | .25 | .50 |

Direct RBF and SLNA use the same numerical configuration but retain different mapping formulations. Direct RBF directly predicts target coordinates, whereas SLNA first recovers the global coordinate relation and then models the remaining residual deformation.

**SLNA-soft configurations.** SLNA-soft is applied after SLNA has been fitted and frozen, following Equation 6. We use 750 unverified candidates, 100 incremental controls, $\alpha_s = 0.03$, and $\rho_s = 1$. The barycentric targets restore the source norm, and no confidence pruning or query-support gating is used.

For Flickr30k, we use balanced Sinkhorn with $\varepsilon = 0.05$. The posterior also uses 64 anchor-distance signatures with weight 0.25, $\eta = 0.5$, and a bandwidth scale of 0.75. For MSCOCO, we use balanced Sinkhorn with $\varepsilon = 0.1$ and a $k = 20$ local-density signature with weight 0.5. We set $\eta = 0.05$ and the bandwidth scale to 0.75, with incremental controls selected from verified landmarks. These configurations are fixed before final evaluation.

**Fixed-configuration protocol.** The remaining diagnostic experiments use a common fixed configuration. We set $\alpha = 0.1$, $\rho = 1$, use the automatic median-distance bandwidth, and use at most 500 controls. These values are kept unchanged when varying the landmark budget or other diagnostic factors.

For Flickr30k, the fixed evaluation set contains the first 400 rows of the released feature files. For MSCOCO, we use the traceable ID-sorted `val2014` subset. SUE does not release the exact identities of its sampled 400-item evaluation set. Therefore, Table 1 follows the same pair budgets but cannot provide an identity-matched test-set comparison.

The pair-budget and sensitivity experiments are used to analyze trends under a common configuration. Their configurations are not re-selected for each experimental condition.

The hard-correspondence controls use the same candidate pool, posterior construction, and residual-field hyperparameters as SLNA-soft. Row-wise argmax selects the highest-probability target independently for each source and may produce repeated targets. Hungarian matching instead enforces a one-to-one assignment between the two candidate pools.

## B Complete Retrieval Results

This section provides additional retrieval results under the fixed-configuration protocol. All experiments follow the settings described in Appendix A.4.

Figures 4 and 5 show that the advantage of SLNA becomes clearer as more verified landmarks are available. SLNA consistently improves over rigid Procrustes at moderate and larger landmark budgets, with a more pronounced gain in R@10. SLNA-soft generally maintains or further improves the retrieval performance of SLNA. These results support the use of local residual modeling once the landmarks provide sufficient coverage of the spectral space.

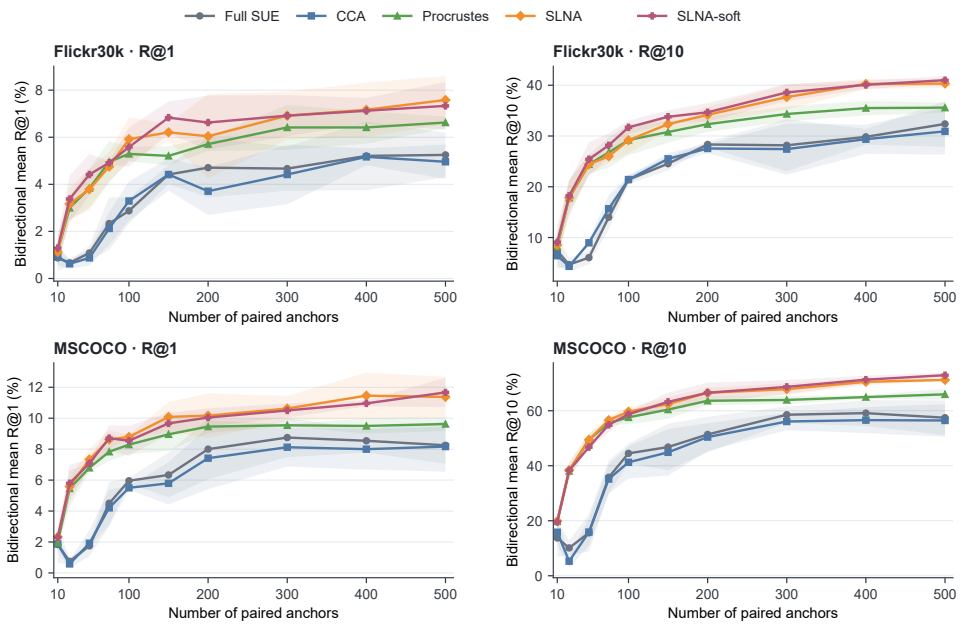



Figure 4: **Complete landmark-budget curves under the fixed-configuration protocol.** Bidirectional R@1 and R@10 are reported over 10–500 verified landmarks using three-seed means and sample standard deviations. The R@10 panels are also shown in Figure 3.

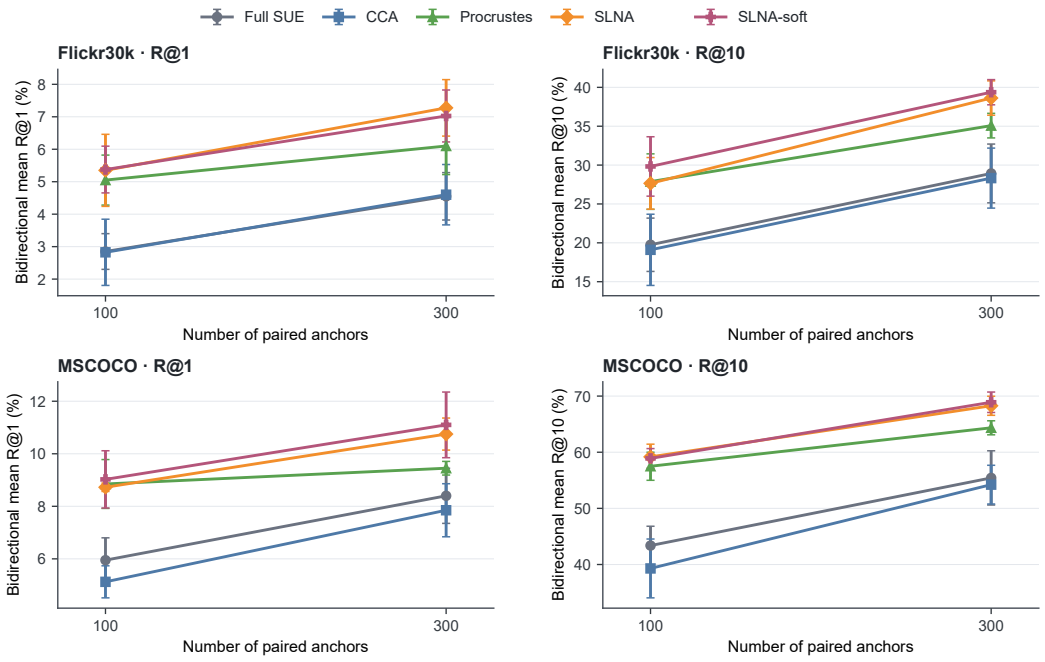



Figure 5: **Fixed-configuration retrieval results at 100 and 300 landmarks.** Bidirectional R@1 and R@10 are reported as five-seed means with sample standard deviations.

## C Representation and Registration Decomposition

This section further separates the contribution of the spectral representation from the alignment stage. We examine the gain introduced by the global–local decomposition, reproduce the main components of SUE, and evaluate SLNA with an alternative encoder pair.

Figure 6 directly measures the retrieval gain introduced by SLNA relative to rigid Procrustes and Direct RBF. Compared with Procrustes, SLNA provides clear positive gains across the main landmark budgets, showing the benefit of modeling the residual deformation after global alignment. Compared with Direct RBF, the gains are smaller but become more consistent when sufficient landmarks are available.

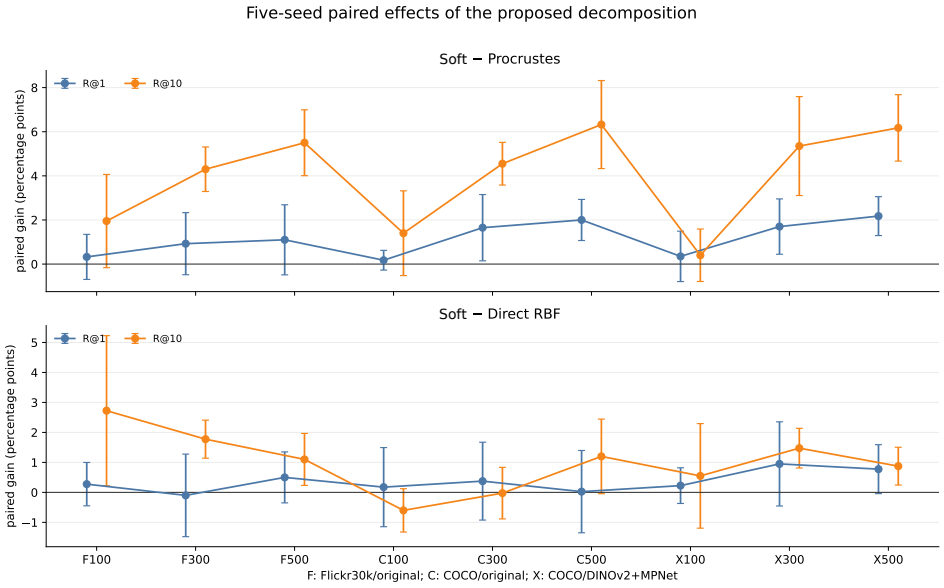



Figure 6: **Paired effects of the global–local decomposition under the fixed-configuration protocol.** Error bars show paired 95% bootstrap intervals over five shared seeds. F/C/X denote Flickr30k, primary MSCOCO, and alternative-backbone MSCOCO, respectively; suffixes indicate the landmark budget.

Figure 7 shows that the spectral representation provides a strong basis for sparse alignment: combining spectral embeddings with CCA increases R@10 substantially over the raw-feature variants, while the additional MMD refinement provides a smaller gain. Our contribution therefore focuses on improving the alignment stage after the spectral spaces have been constructed. Figure 8 further shows that the advantage of SLNA is retained with an independently re-encoded image–text feature pair, especially at 300 and 500 landmarks. SLNA and SLNA-soft remain stronger than the rigid alignment baseline in these settings, indicating that the benefit of global–local alignment is not limited to the primary encoder pair.

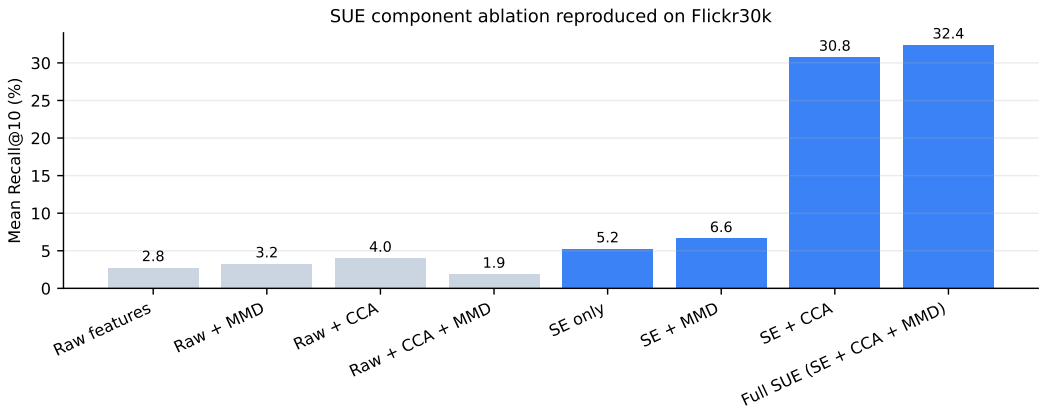



Figure 7: **Component study of the locally reproduced SUE pipeline on Flickr30k.** Bidirectional mean R@10 is shown for combinations of raw features, spectral embeddings (SE), CCA, and MMD.

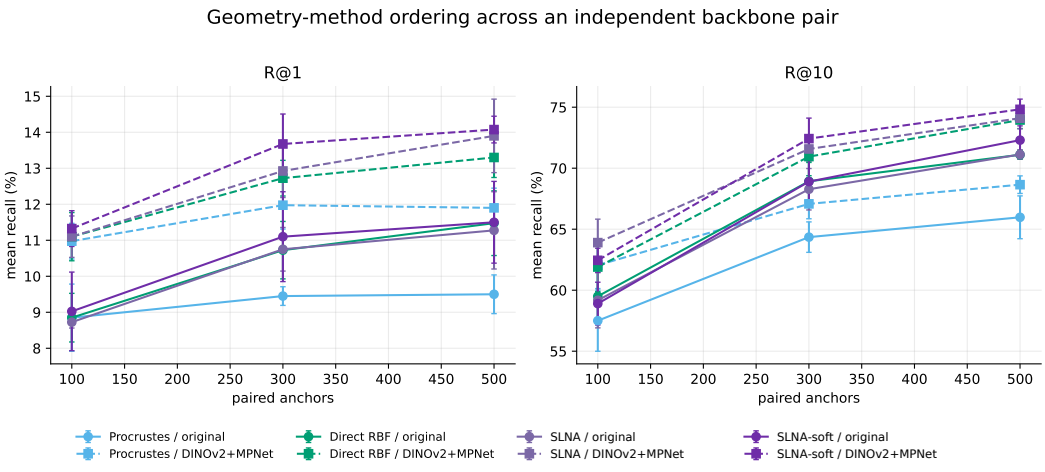



Figure 8: **Alignment with an alternative encoder pair.** Five-seed mean R@1 and R@10 are reported for the primary MSCOCO representations and independently re-encoded DINOv2-S/14 image and all-mpnet-base-v2 text features.

## D STRUCTURAL AND SYNTHETIC DIAGNOSTICS

Figure 9 further examines the structural premise of sparse landmark alignment. We standardize the coordinates of each modality separately and construct Gaussian-weighted $k = 15$ neighborhood graphs. Verified correspondences show substantially smaller structural discrepancy than 1,000 shuffled correspondences on both datasets, while the estimated intrinsic dimensions remain well below the ambient dimension of 30. These results indicate that independently learned spectral spaces retain non-random and locally low-dimensional cross-modal structure, providing a suitable basis for landmark-based alignment.

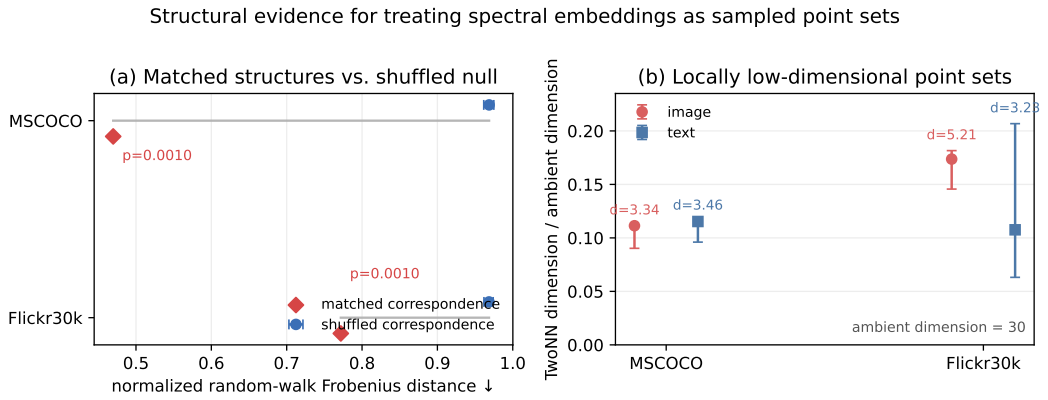



Figure 9: **Full structural diagnostic.** Verified correspondence is compared with 1,000 shuffled index mappings, and TwoNN estimates local dimension within the 30-dimensional spectral coordinates.

Figure 10 further verifies the benefit of the global–local decomposition under a known non-rigid deformation. As landmark coverage increases, Procrustes approaches a rigid error floor, whereas adding local residual modeling continues to reduce the correspondence error and improve neighborhood recovery. At 300 landmarks, the residual model reduces RMSE from 0.2612 to 0.0322 and increases $k$NN retention from 0.8640 to 0.9454. This confirms that the local residual field can recover deformation that cannot be captured by global rigid alignment alone.

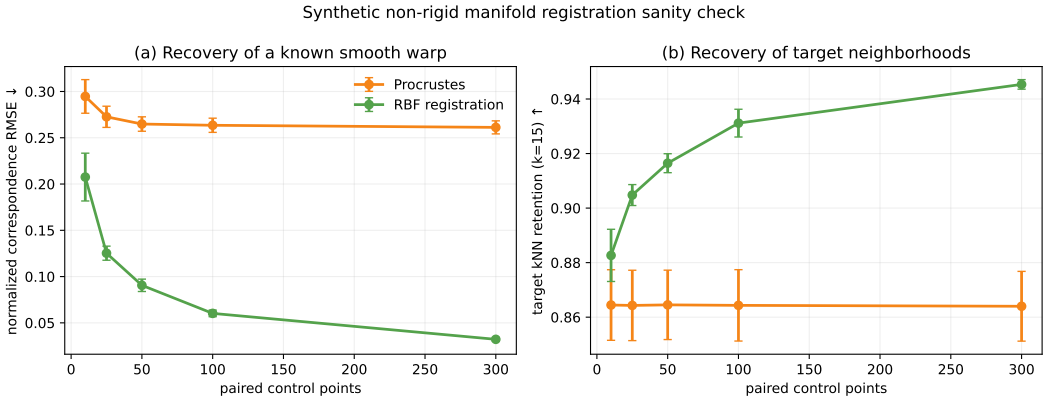



Figure 10: **Recovery of a known smooth warp.** Curves show mean and standard deviation over five synthetic seeds for correspondence RMSE and target-neighborhood retention.

## E Robustness and Efficiency

Figures 11–13 examine the sensitivity of SLNA to the number of unverified candidates, landmark quality, and the structural-cost weight. Increasing the candidate pool does not produce monotonic improvements, while replacing verified landmarks with mismatches consistently degrades retrieval performance. In contrast, performance changes only mildly across the tested structural-cost weights, particularly on MSCOCO. These results show that the main retrieval benefit relies on reliable landmark-based global–local alignment, while the optional posterior correction is less sensitive to its auxiliary configuration.

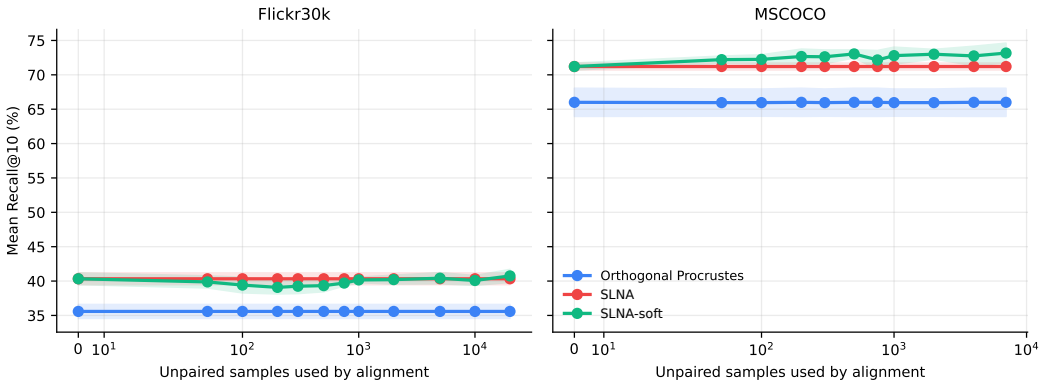



Figure 11: **Sensitivity to the number of unpaired candidates.** Bidirectional mean R@10 for jointly fitted Soft under the fixed-configuration protocol. Flickr30k largely plateaus as the pool grows, while MSCOCO shows small non-monotonic changes.

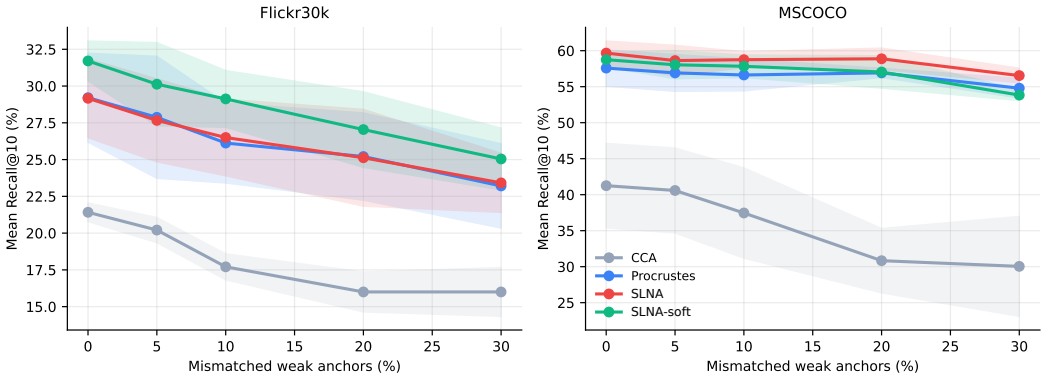



Figure 12: **Anchor-corruption stress test.** Bidirectional mean R@10 as 0–30% of verified correspondences are replaced by mismatches. Robustness varies across datasets and alignment methods.

Figure 14 compares retrieval performance with the recorded alignment cost. Procrustes has the lowest fitting cost, while SLNA introduces additional computation for residual-field estimation and posterior correction. This additional offline computation is accompanied by higher retrieval accuracy for SLNA and SLNA-soft. After fitting, gallery representations can be precomputed, so the alignment procedure does not need to be repeated for individual retrieval queries.

Figure 15 further compares SLNA-soft with matched posterior-control variants. The relatively small differences among the full, feature-only, and shuffled-structure variants indicate that the additional Soft gain is not dominated by a single structural component. This is consistent with the main experiments, where SLNA-soft acts as an optional refinement over SLNA rather than the primary source of the retrieval improvement.

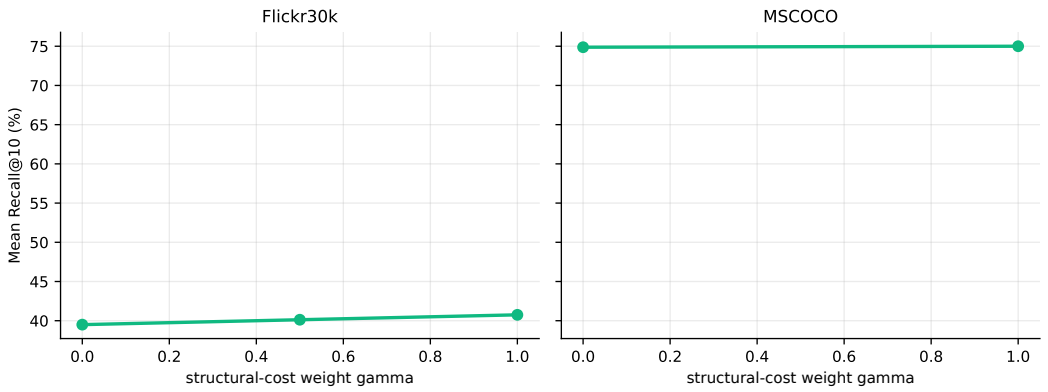



Figure 13: **Sensitivity to the structural-cost weight.** Fixed-configuration retrieval varies only weakly with the weight, particularly on MSCOCO.

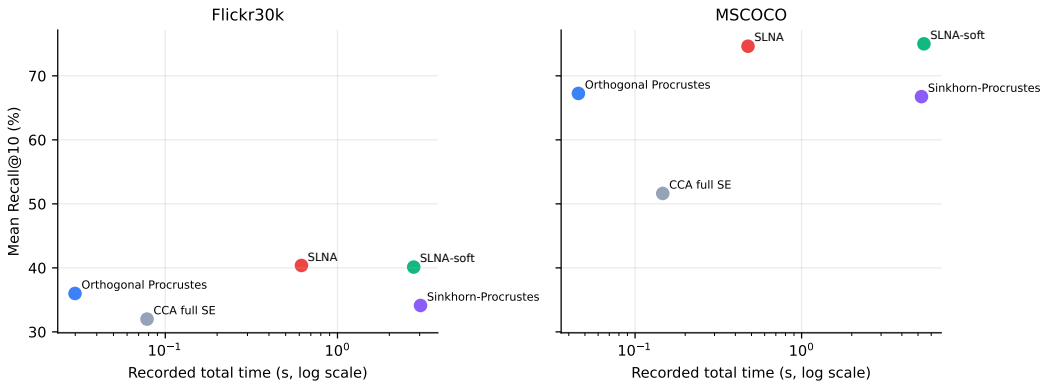



Figure 14: **Recorded registration time versus retrieval performance.** Each point shows bidirectional mean R@10 and total recorded alignment time for a representative fixed-configuration run. Timings remain implementation- and environment-dependent.

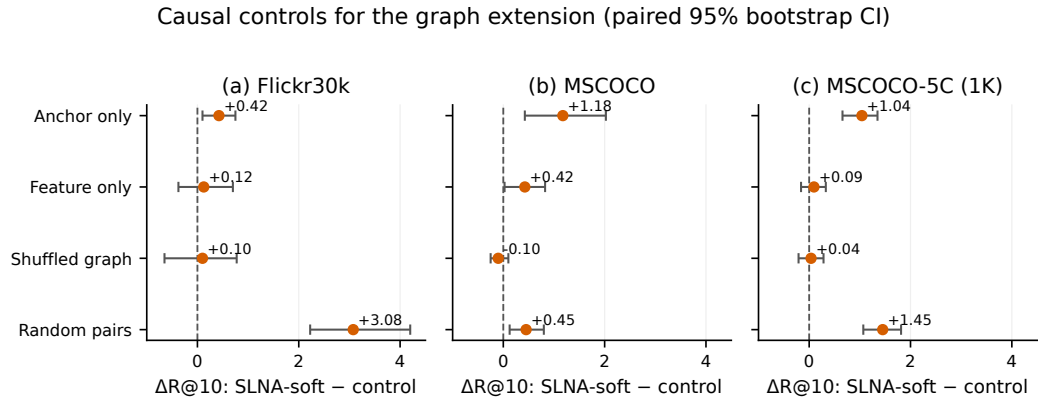



Figure 15: **Seed-wise Soft effects relative to matched controls.** The fixed-configuration full, feature-only, and shuffled-graph variants produce similar results, leaving the contribution of graph topology unresolved.

## F Correspondence Diagnostics

Figure 16 compares the posterior barycentric targets used by SLNA-soft with two discrete alternatives. Row-wise argmax independently chooses the highest-probability target and can therefore assign the same target to multiple source samples. Hungarian matching enforces a one-to-one assignment, but it can introduce larger source-to-target displacements. The barycentric construction instead preserves posterior uncertainty and provides a smoother residual signal for the optional correction.

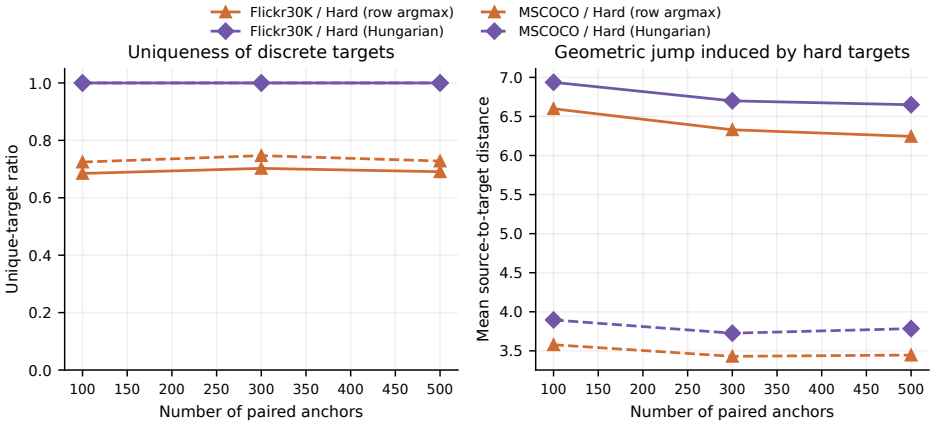



Figure 16: **Diagnostics for discrete targets.** Hungarian matching enforces uniqueness but induces larger geometric jumps; row-wise argmax permits repeated targets.

Table 6 and Figure 17 compare these correspondence choices in retrieval. The barycentric Soft target achieves the highest mean R@10 on Flickr30k and consistently exceeds Hungarian assignment across the reported settings. On MSCOCO, its performance remains close to row-wise argmax, with only small differences between the two. These results indicate that posterior barycenters provide a stable alternative to hard assignment, although the additional retrieval benefit remains dataset- and protocol-dependent.

Table 6: **R@10 for soft and hard targets at 500 anchors.** Fixed-configuration protocol, five-seed mean $\pm$ sample standard deviation.

| Protocol | Soft | Row argmax | Hungarian |
|---|---|---|---|
| Flickr30k single | $41.48 \pm 0.90$ | $40.60 \pm 1.02$ | $41.05 \pm 0.82$ |
| MSCOCO single | $72.30 \pm 1.17$ | $72.40 \pm 0.95$ | $72.22 \pm 1.07$ |
| MSCOCO 5K/1K-fold | $38.19 \pm 0.40$ | $38.36 \pm 0.62$ | $38.08 \pm 0.39$ |

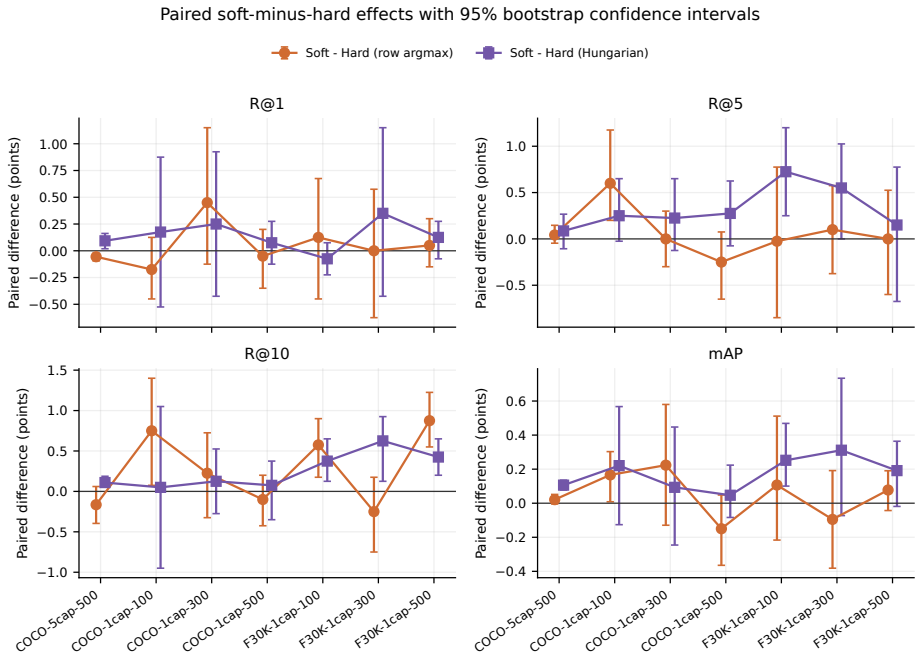



Figure 17: **Paired soft-minus-hard effects.** Positive values favor barycentric targets. The advantage is more consistent against Hungarian assignment than against row-wise argmax.

## G Geometry and Qualitative Visualization

Figure 18 further examines the balance between retrieval performance and geometric preservation. Procrustes preserves the original neighborhood structure exactly but provides lower retrieval accuracy. In contrast, SLNA and SLNA-soft improve R@10 while retaining most of the original high-dimensional neighborhoods. This supports the view that SLNA improves retrieval through moderate local geometric correction rather than large-scale distortion of the original spectral structure.

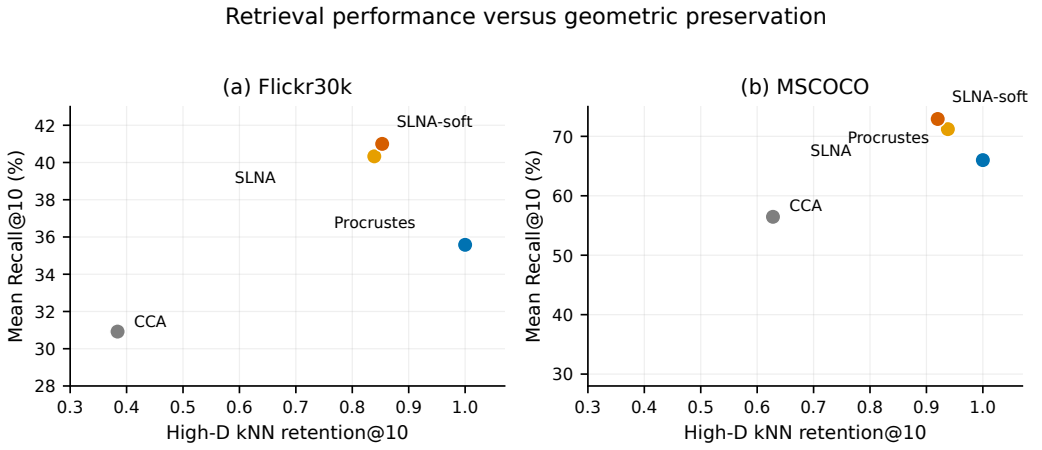



Figure 18: **Retrieval versus high-dimensional neighborhood retention.** Retention averages image and text $k$NN overlap with $k = 10$. The one-way RBF maps leave text coordinates unchanged; higher mean retention indicates less average geometric change.

Figures 19 and 20 provide qualitative views of the cross-modal organization used by SLNA. After alignment, corresponding image and text representations show closer local organization than in the

original independent spectral spaces. The local semantic neighborhood also illustrates the coarse correspondence information available to the optional SLNA-soft correction. These visualizations complement the quantitative retrieval results by showing how local semantic structure can support cross-modal alignment.

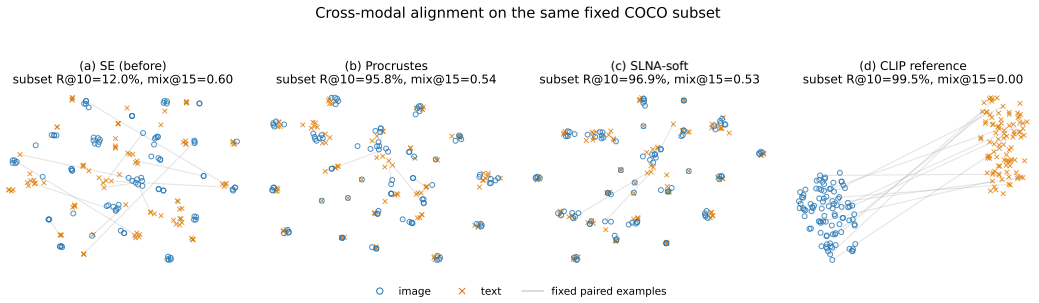



Figure 19: **Cross-modal visualization on a fixed COCO subset.** Each t-SNE panel is fitted independently and shows the local arrangement within that representation.

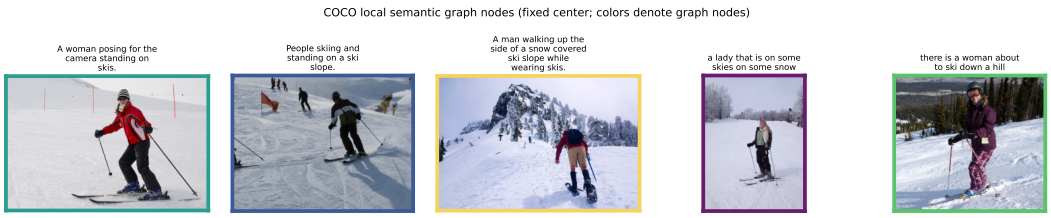



Figure 20: **Fixed local semantic neighborhood in the unpaired pools.** The panel illustrates the coarse local evidence available to the optional correspondence branch.

Figures 21–25 further analyze how SLNA changes the geometry of the two spectral spaces. The distance and retrieval-margin analyses show that SLNA reduces cross-modal mismatch while preserving useful within-modality structure. The shared-reference PCA provides a common view of the transformation, while the semantic PCA and UMAP visualizations show the local organization of image and text samples after alignment. Together, these results provide additional evidence that the retrieval gains arise from structured geometric correction rather than unconstrained deformation.

Figures 26–28 provide qualitative evidence of the semantic information preserved after alignment. The retrieval examples show that SLNA-soft can recover images with the correct coarse semantic content, while remaining errors mainly involve finer attributes, counts, and scene composition. The caption-prototype experiment further shows that the aligned representations retain category-level cross-modal information on held-out samples. This supervised prototype analysis is used only as a diagnostic and is separate from the retrieval evaluation.

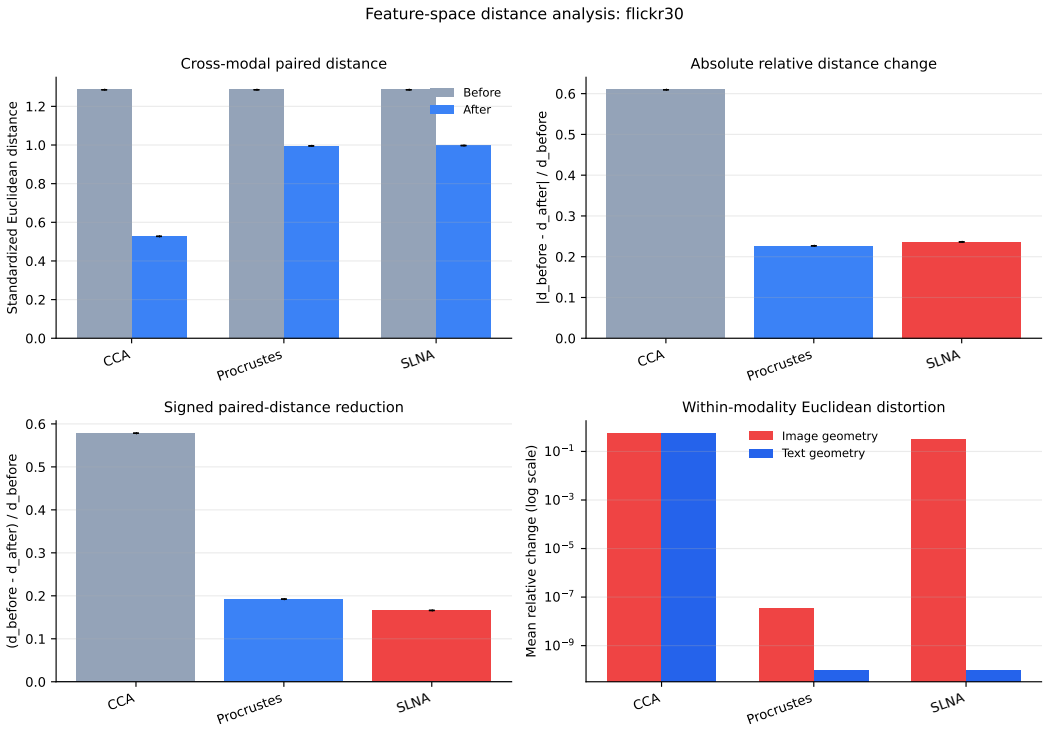



Figure 21: **Paired-distance and within-modality changes on Flickr30k.** Distance contraction measures transformation magnitude; retrieval and neighborhood retention indicate whether the resulting map is useful.

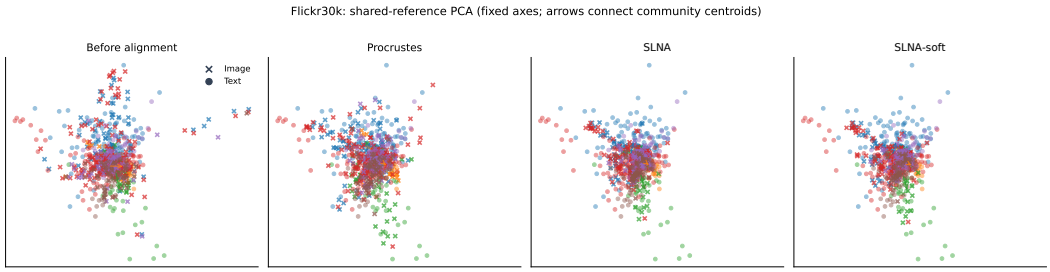



Figure 22: **Shared-reference PCA on Flickr30k.** A single basis fitted to the pre-alignment spectral point sets is reused in every panel, allowing the projected transformations to be viewed in a common frame.

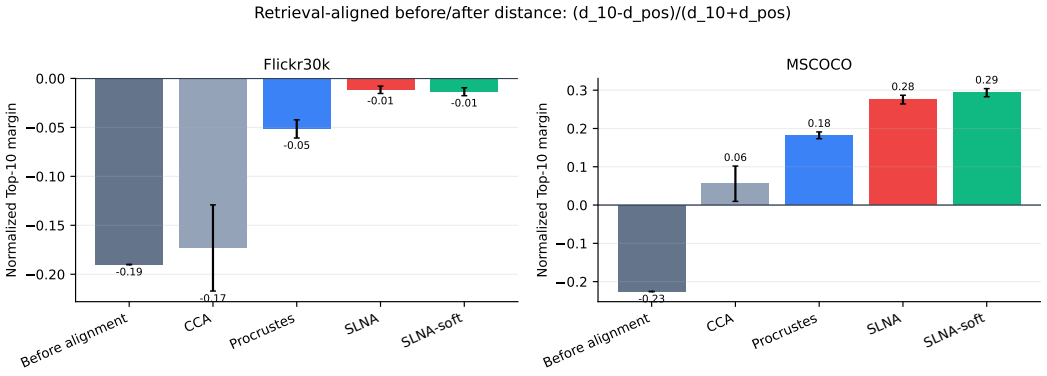



Figure 23: **Normalized Top-10 retrieval margin.** Larger values indicate better separation of a positive match from the Top-10 boundary. Remaining negative margins on Flickr30k expose unresolved hard-negative separation.




Figure 24: **Semantic PCA on a fixed balanced MSCOCO sample.** Each method panel uses its own fitted basis. Local organization can be inspected within panels, but global coordinates are not directly comparable.

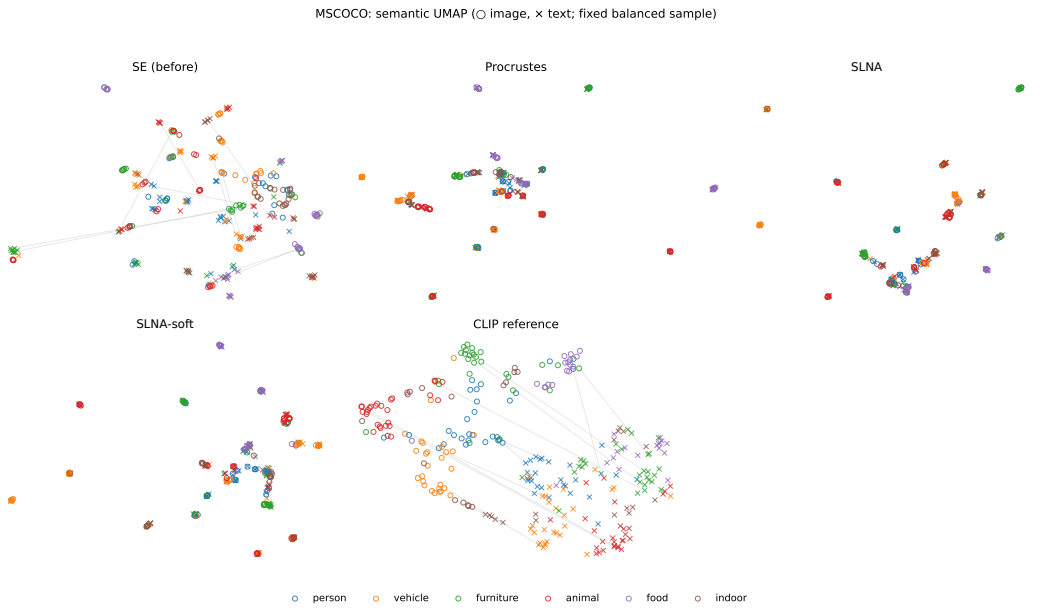



Figure 25: **Semantic UMAP on the same fixed MSCOCO sample.** Each panel is embedded independently and illustrates the local arrangement of the sampled categories.

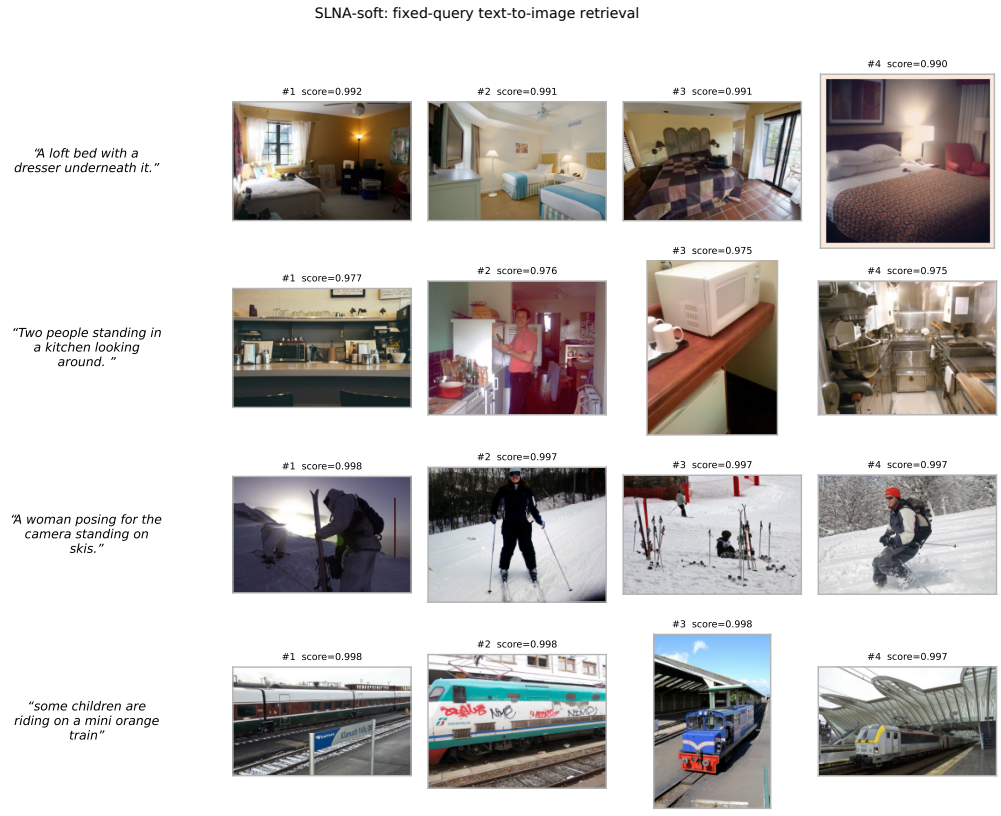



Figure 26: **Fixed-query text-to-image retrieval examples for fixed-configuration RBF-graph.** Query indices were fixed before visual inspection. The examples show both coarse semantic successes and failures involving counts, attributes, and scene composition.




Figure 27: **Additional caption-driven retrieval examples for fixed-configuration RBF-graph.** The outputs illustrate both coarse semantic agreement and fine-grained mismatches.

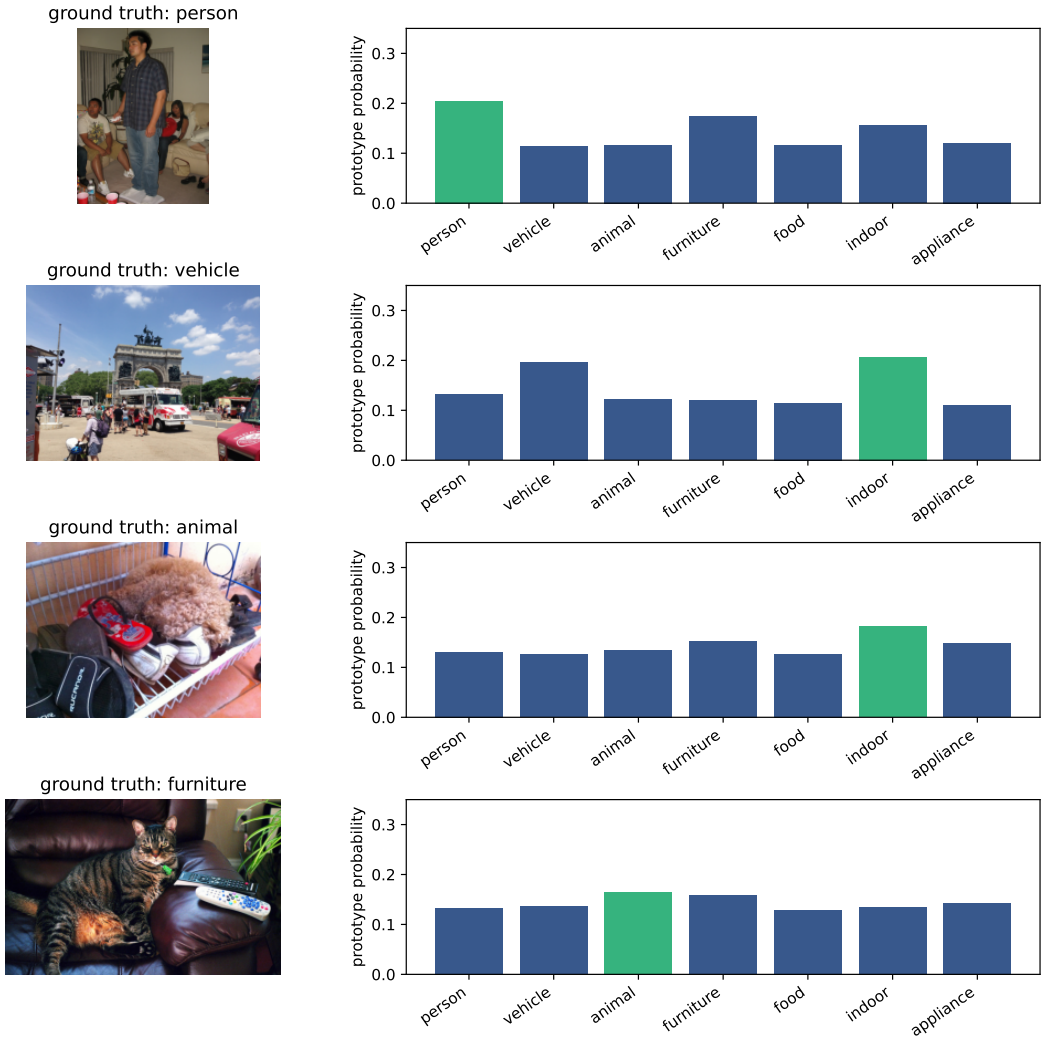



Figure 28: **Cross-modal caption prototypes on MSCOCO.** Supervised prototypes from seven coarse categories are evaluated on held-out images.

## H CSA DIAGNOSTIC

Figure 29 examines the sensitivity of CSA to the number of retained canonical dimensions. Its retrieval performance varies with both the retained dimension and the available landmark budget, showing that this baseline is sensitive to this configuration. We therefore keep the reference setting fixed at $s = 8$ throughout the main comparisons rather than selecting the best dimension from this diagnostic sweep. This provides a consistent CSA baseline for evaluating the retrieval gains of SLNA under different landmark budgets.

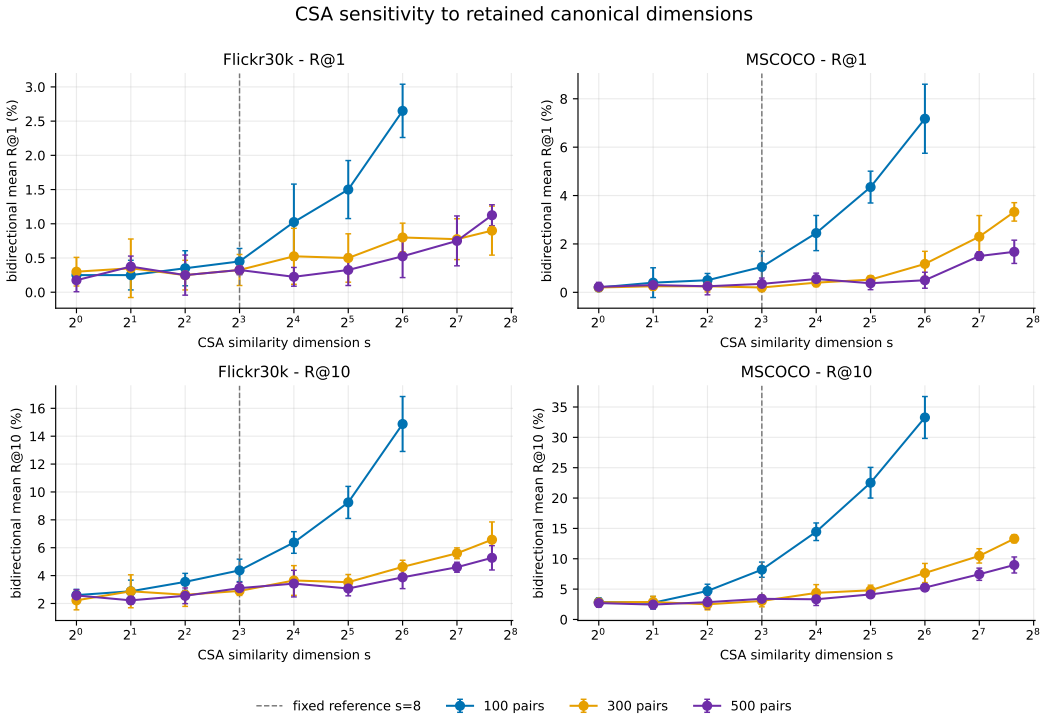



Figure 29: **CSA sensitivity to the number of retained canonical dimensions.** Bidirectional mean R@1 and R@10 are shown for 100, 300, and 500 paired anchors on Flickr30k and MSCOCO over five matched seeds. The dashed line marks the pre-fixed local reference $s = 8$.

## I AI USE STATEMENT

Generative AI tools were used for language editing, translation, figure-layout assistance, and consistency checking during manuscript preparation. The authors checked the reported values against the experiment files and remain responsible for the paper's claims, equations, references, and final wording.