RankAlign: Vision-Language Representation Alignment via Rank Transformation

Enzhe Zhao[2]* **Marco Fiorucci**[1]* **Lamberto Ballan**[1]

[1] University of Padova, Italy
[2] Harbin Institute of Technology, China

Project page: marcofiorucci.github.io/rankalign/

Abstract

While the Platonic Representation Hypothesis posits that vision and language models converge toward compatible latent geometries, recovering class-level correspondences between unpaired modalities without cross-modal supervision remains challenging. Existing relational matching methods compare within-modality similarity kernels directly, although the same semantic structure may be represented through different numerical parameterizations across modalities. We propose RankAlign, a rank-based normalization strategy for class-level unpaired vision-language matching. RankAlign transforms each within-modality similarity kernel into an ordinal relational representation before permutation matching, while leaving the correspondence space and matching solver unchanged. The resulting representation is invariant to strictly increasing reparameterizations of similarity values while preserving the ordinal structure used for matching. We provide exact-recovery and stability guarantees that characterize correspondence recovery under both exact and partial ordinal agreement. Experiments across diverse frozen vision and language encoders and multiple datasets demonstrate improved and more stable class-level correspondence recovery, including under label corruption and domain shift.

1 Introduction

Recent advances in foundation models have fundamentally redefined representation learning across diverse modalities. While self-supervised vision encoders like DINO [1] and large-scale language models [3, 17] have matured within their respective domains, multimodal contrastive frameworks such as CLIP [15] and SigLIP2 [22] have further bridged the two modalities. These architectures induce sophisticated latent representations that capture intricate structural and semantic nuances. Yet, despite this progress, the geometric alignment between visual and language representation spaces remains poorly understood when encoders are trained independently on disparate datasets. This poses a significant challenge for class-level unpaired cross-modal matching, where category structure is available within each modality, while the correspondence between visual and textual categories remains unknown and must be recovered.

Representation similarity and the Platonic hypothesis. Independently trained visual and language models frequently exhibit a striking convergence toward compatible representational structures. This phenomenon is formalized by the Platonic Representation Hypothesis, which posits that scaling and data diversity drive neural networks toward shared latent geometries that encode the statistical regularities of the world [6]. In vision-language settings, visual and textual encoders learn compatible relational structures despite disparate training objectives [10, 19]. Although methods such as CCA [5],

*Equal contribution. Correspondence to: marco.fiorucci@unipd.it.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

SVCCA [16], and CKA [8] provide quantitative measures of representational similarity, they rely on direct numerical comparisons that are inherently sensitive to modality-specific scale distortions and biases.

Class-level unpaired vision-language matching. The emergence of a shared embedding space raises the fundamental question: can explicit correspondences between vision and language representations be recovered when independently trained encoders receive no paired supervision? Traditional alignment relies on image–caption pairs or manually verified correspondences to learn a cross-modal mapping. We instead study class-level unpaired vision-language matching with known within-modality groups and an unordered candidate set of textual concepts, where visual groups and a set of candidate textual concepts are available but their cross-modal correspondence is unknown. Supervision is confined to constructing class-level representations within each modality, with no paired image–text samples, cross-modal anchors, or verified correspondences provided to the matching procedure. When reliable cross-modal anchors are available, supervised alignment is often simpler and more practical. Our goal is to determine whether independently trained encoders contain sufficient relational structure to recover correspondence without such anchors. This zero-anchor setting can arise when visual groups and a candidate textual ontology are collected or maintained independently, so that both sides are available but their correspondence has not been verified. In this regime, correspondence recovery can reduce the need for manually established cross-modal anchors and also serves as a diagnostic of structural compatibility between independently trained representation spaces. More broadly, it tests whether cross-modal correspondence can emerge from relational structure alone, without shared training data or explicit cross-modal supervision. Recent work frames this problem as matching pairwise relationships between embeddings [19]. Such formulations arise naturally in kernel alignment [2], graph matching via quadratic assignment [23], and structural optimal transport, most notably the Gromov-Wasserstein ($\mathcal{GW}$) distance [14, 24].

Limitations of value-based relational alignment. Blind matching compares the within-modality similarity kernels K_x and K_y and searches for the class permutation π under which their relational structures agree best (Section 2.1). Existing relational alignment methods already operate at the level of pairwise relations rather than raw embeddings, but they still compare the numerical entries of the two kernels directly [11, 19, 12]. Even when vision and language encoders learn similar embeddings, the numerical entries of their within-modality kernels need not be directly comparable across modalities. The two modalities may agree on the ordering of concept similarities, while still assigning different kernel values to the same relations. Consequently, alignment objectives that rely directly on kernel-entry comparison can become weakly discriminative [8, 16].

This numerical mismatch weakens the discriminative value of direct kernel-entry comparison. At the same time, recent evidence suggests that cross-modal agreement can be more stable in relative and local relational structure than in absolute similarity values [4]. These observations raise a key question: can cross-modal correspondence be recovered from relational order when the numerical kernel values are not directly comparable across modalities?

Rank-based relational alignment. To address this issue, we introduce an rank-based formulation of relational alignment. Instead of matching kernels based on raw values, we align them by comparing their rank-based transformation, which uses the ordering induced by these entries. The goal is therefore not to compare $K_x(i, j)$ and $K_y(\pi(i), \pi(j))$ as numbers, but to ask whether the two modalities induce consistent relational order over concept pairs. Concretely, we replace each off-diagonal entry of a within-modality kernel with its rank among the off-diagonal entries and evaluate the same matching objective on the transformed kernels. The correspondence problem itself remains unchanged, while the matching signal is determined by relational order rather than raw kernel values.

Contributions. We study class-level unpaired vision-language matching with known categories. Class-level representations are available within each modality, while their cross-modal correspondence is unknown. Our contributions are summarized as follows:

- **Solver-agnostic ordinal relational representation.** We introduce a rank-based representation that plugs into existing blind-matching objectives and QAP solvers without changing the correspondence space or optimization procedure. RankAlign removes arbitrary strictly increasing reparameterizations of similarity values while retaining the ordinal structure used for matching.
- **Theory of ordinal correspondence recovery.** We formalize cross-modal agreement up to modality-specific monotone reparameterizations and show that the resulting ordinal repre-

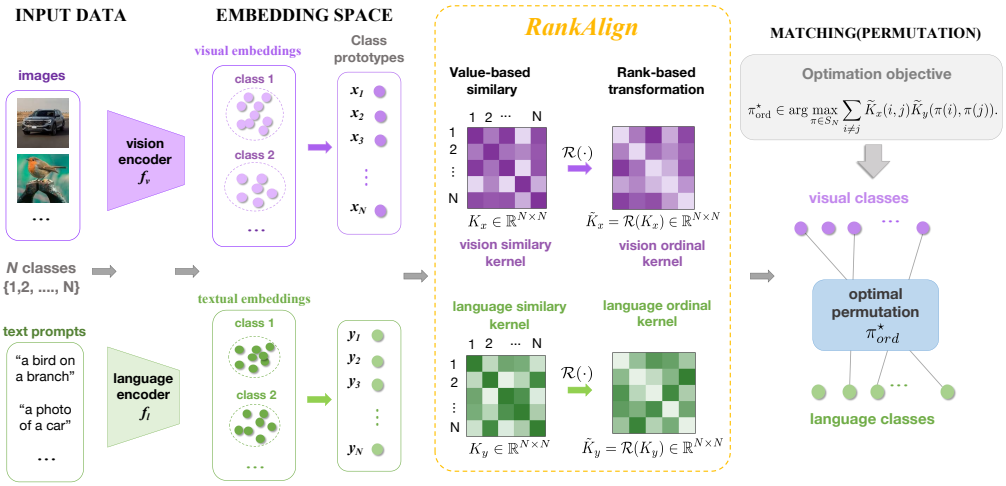

Figure 1: Visual-language alignment pipeline. The RankAlign module provides the core ordinal representation of each within-modality similarity kernel. These representations are refined using local neighbor order before permutation matching, without changing the matching space or solver.

sentation is invariant to such distortions. We further provide exact-recovery and stability guarantees that separate cross-modal ordinal disagreement from the intrinsic distinguishability of class relational profiles.

- **Empirical validation.** Experiments across diverse vision and language encoders and datasets demonstrate that RankAlign achieves more stable and reliable cross-modal correspondence alignment.

2 Method

Figure 1 summarizes our visual-language alignment pipeline, which illustrates its key steps: *i*) learning visual and textual embeddings and computing the corresponding prototypes; *ii*) RankAlign transforms the visual and textual similarity kernels into rank-based similarity kernels; and *iii*) permutation matching for cross-modal alignment. In the following sections, we will describe each component in detail.

2.1 Preliminary

We adopt the blind-matching setting of Schnaus et al. [19]. Let f_v and f_l denote frozen visual and language encoders, respectively. We consider N semantic classes shared across the two modalities. We construct class-level prototypes as

$$x_i = \frac{1}{|\mathcal{I}_i|}\sum_{I\in\mathcal{I}_i} f_v(I), \qquad y_i = \frac{1}{|\mathcal{T}_i|}\sum_{T\in\mathcal{T}_i} f_l(T). \tag{1}$$

Then, we center the prototypes independently within each modality and renormalize them before constructing the relational kernels

$$\hat{x}_i = \frac{x_i - \bar{x}}{\|x_i - \bar{x}\|}, \qquad \bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j, \tag{2}$$

and analogously for the language prototypes. This operation uses no cross-modal correspondence information and is permutation equivariant.

For notational simplicity, we use x_i and y_i below to denote the resulting normalized prototypes. The within-modality similarity kernels are

$$K_x(i,j) = \langle x_i, x_j\rangle, \qquad K_y(i,j) = \langle y_i, y_j\rangle. \tag{3}$$

A candidate cross-modal correspondence is represented by a permutation $\pi \in S_N$. Given a distortion function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, define

$$D_\ell(K_x, K_y; \pi) = \sum_{i \neq j} \ell(K_x(i,j), K_y(\pi(i), \pi(j))) \,. \tag{4}$$

Blind relational matching seeks

$$\pi^\star \in \arg\min_{\pi \in S_N} D_\ell(K_x, K_y; \pi). \tag{5}$$

Equivalently, we use the agreement score

$$f(\pi) = \sum_{i \neq j} K_x(i,j) K_y(\pi(i), \pi(j)). \tag{6}$$

Lemma 1 (Distortion and agreement)**.** *For $\ell(a,b) = -ab$, minimizing D_ℓ is equivalent to maximizing f. For the squared distortion $\ell(a,b) = (a-b)^2$,*

$$D_\ell(K_x, K_y; \pi) = C_x + C_y - 2f(\pi), \tag{7}$$

where C_x and C_y do not depend on π. Hence both formulations have identical optimizers.

Hence, the objective function can be written as $\pi^\star \in \arg\max_{\pi \in S_N} f(K_x, K_y; \pi)$. And we use the maximization form throughout the remainder of the paper. The resulting problem is a quadratic assignment problem (QAP). A proof of Lemma 1 is given in Appendix A.1.

2.2 Limitations of value-based relational alignment

The objective in Eq. 6 compares numerical kernel values directly. This implicitly assumes that visual and language similarities are sufficiently calibrated across modalities. Such calibration is not guaranteed for independently trained encoders.

To formalize this distinction, suppose both modalities encode a shared latent relational structure $S = \{s_{ij}\}$ through modality-specific strictly increasing maps,

$$K_x(i,j) = \phi_x(s_{ij}), \qquad K_y(\pi^\star(i), \pi^\star(j)) = \phi_y(s_{ij}), \tag{8}$$

where $\pi^\star$ denotes the ground-truth correspondence.

Even under the correct correspondence, the numerical similarities in the two modalities may differ substantially whenever ϕ_x and ϕ_y have different parameterizations. Consequently, a value-based objective can penalize a correct correspondence despite the two modalities encoding the same underlying relational structure.

By contrast, strictly increasing transformations preserve relational order. For any two unordered class pairs e, e',

$$K_x(e) < K_x(e') \quad \Longleftrightarrow \quad K_y(\pi^\star(e)) < K_y(\pi^\star(e')). \tag{9}$$

Thus, the two modalities may share the same ordinal relational geometry without sharing the same numerical parameterization. This suggests representing relational structure through order rather than absolute similarity values, thereby making the matching objective insensitive to modality-specific monotone reparameterizations.

2.3 Rank-based relational alignment

Because the similarity kernels are symmetric, we rank unordered class pairs. Let

$$\mathcal{E} = \{\{i,j\} : 1 \leq i < j \leq N\}, \qquad M = |\mathcal{E}| = \frac{N(N-1)}{2}. \tag{10}$$

For a symmetric kernel K, we define the empirical cumulative distribution function of its off-diagonal similarities as

$$F_K(t) = \frac{1}{M} \sum_{\{i,j\} \in \mathcal{E}} \mathbf{1}[K(i,j) \leq t], \tag{11}$$

and the corresponding rank transformation as

$$R(K)(i,j) = \begin{cases} F_K(K(i,j)), & i \neq j, \\ 0, & i = j. \end{cases} \tag{12}$$

The transformation replaces the numerical spacing of pairwise similarities by their relative positions within each modality. If all M unordered off-diagonal similarities are distinct, the transformed values are $1/M, \ldots, 1$. If ties are present in the original kernel, they remain tied after the transformation. Rank transformation preserves the available relational ordering rather than creating additional relational information.

Lemma 2 (Ordinal invariance)**.** *For every strictly increasing function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$R(\phi \circ K) = R(K). \tag{13}$$

Moreover, for every permutation $\sigma \in S_N$,

$$R(K \circ \sigma) = R(K) \circ \sigma, \tag{14}$$

where $(K \circ \sigma)(i,j) = K(\sigma(i), \sigma(j))$.

Lemma 2 establishes the central invariance of the rank representation. Modality-specific strictly monotone reparameterizations of similarity values are removed without changing the underlying correspondence structure. The transformation also preserves every strict pairwise ordering induced by the original kernel. The proof is provided in Appendix A.1.

Replacing the kernels in Eq. 6 by their rank representations gives the ordinal matching objective

$$f_{\mathrm{ord}}(\pi) = \sum_{i \neq j} \widetilde{K}_x(i,j) \widetilde{K}_y(\pi(i), \pi(j)), \qquad \widetilde{K}_x = R(K_x), \quad \widetilde{K}_y = R(K_y). \tag{15}$$

The correspondence problem itself is unchanged. RankAlign changes only the relational representation supplied to the matching objective. The recovery analysis below focuses on the core rank representation $R(K)$, while the full ordinal refinement used in the experiments is described in Appendix A.1.

2.4 Recovery guarantees

We next give conditions under which ordinal relational matching recovers the ground-truth correspondence. We first establish exact recovery for the basic rank objective in Eq. 15, which isolates the contribution of ordinal information, and then study stability under partial ordinal disagreement.

Assumption 1 (Strict ordinal consistency)**.** *Let $\pi^\star$ denote the ground-truth correspondence. Assume that the unordered off-diagonal values of K_x are distinct and that the two modalities induce the same strict ordering over unordered class pairs under $\pi^\star$. For any distinct $e, e' \in \mathcal{E}$,*

$$K_x(e) < K_x(e') \quad \Longleftrightarrow \quad K_y(\pi^\star(e)) < K_y(\pi^\star(e')). \tag{16}$$

Assumption 1 requires agreement in relational order, rather than agreement in numerical similarity values. The noise-free latent model of Eq. 8 satisfies this assumption whenever the latent similarities s_{ij} are distinct, independently of the strictly increasing modality-specific maps ϕ_x and ϕ_y.

Proposition 1 (Exact recovery)**.** *Under Assumption 1 and $N \geq 3$, $\pi^\star$ is the unique maximizer of f_{ord} over S_N. Moreover, for every $\pi \neq \pi^\star$,*

$$f_{\mathrm{ord}}(\pi^\star) - f_{\mathrm{ord}}(\pi) \geq \frac{2}{M^2}, \qquad M = \frac{N(N-1)}{2}, \tag{17}$$

when the objective is summed over ordered off-diagonal pairs. Equivalently, the corresponding margin over unordered pairs is at least $1/M^2$.

Proposition 1 shows that strict ordinal consistency is sufficient for unique correspondence recovery. The guarantee depends only on the relational ordering and the problem size, rather than on the numerical spacing of either kernel. The complete proof is given in Appendix A.2.

Strict ordinal consistency is an idealized condition. In practice, some relational orders may disagree across modalities. We therefore state the stability result in a representation-independent form. Let

$Z(K)$ denote any permutation-equivariant ordinal representation, with $Z_x = Z(K_x)$ and $Z_y = Z(K_y)$. We define the cross-modal residual as

$$E_Z = Z_y \circ \pi^\star - Z_x, \tag{18}$$

and the structural discrepancy induced by a permutation τ as

$$d_Z(\tau) = \|Z_x - Z_x \circ \tau\|_F. \tag{19}$$

Proposition 2 (Stability under partial ordinal disagreement)**.** *For any candidate permutation π, let $\tau = (\pi^\star)^{-1} \circ \pi$. Then*

$$f_Z(\pi^\star) - f_Z(\pi) \geq d_Z(\tau) \left(\frac{1}{2} d_Z(\tau) - \|E_Z\|_F \right), \tag{20}$$

where $f_Z(\pi) = \sum_{i \neq j} Z_x(i,j) Z_y(\pi(i), \pi(j))$. Consequently, every maximizer $\widehat{\pi}$ satisfies

$$d_Z(\widehat{\tau}) \leq 2\|E_Z\|_F, \qquad \widehat{\tau} = (\pi^\star)^{-1} \circ \widehat{\pi}. \tag{21}$$

The result separates two sources of difficulty. The residual $\|E_Z\|_F$ measures cross-modal ordinal disagreement, while $d_Z(\tau)$ measures how distinguishable the relational structure is under an incorrect relabeling. Further specialization to the practical RankAlign representation, together with bounds on mismatched classes and approximate solver outputs, is provided in Appendix A.2.

2.5 Computational complexity

We finally analyze the additional computational cost introduced by RankAlign. Constructing a dense similarity kernel from N class prototypes of dimension d requires $O(N^2 d)$ time and $O(N^2)$ memory. The kernel-level rank transformation sorts $M = N(N-1)/2 = O(N^2)$ pairwise similarities, while the anchor-wise transformation ranks N rows of length $N-1$. Both operations require $O(N^2 \log N)$ time. The score transformation, symmetrization, and standardization are all $O(N^2)$.

Consequently, excluding feature extraction, RankAlign representation can be constructed in $O(N^2 d + N^2 \log N)$ time and $O(N^2 + Nd)$ memory. Since the $O(N^2 d)$ kernel construction is already required by value-based relational matching, the additional asymptotic preprocessing overhead of RankAlign is $O(N^2 \log N)$ time and $O(N^2)$ memory. RankAlign does not change the subsequent correspondence problem. If $C_{\mathcal{S}}(N)$ denotes the running time of a chosen QAP solver $\mathcal{S}$, the complete matching pipeline requires $C_{\mathcal{S}}(N) + O(N^2 d + N^2 \log N)$ time. Further details are provided in Appendix A.3.

3 Experiments

We measured performance using matching accuracy, defined as the fraction of correctly recovered class correspondences. A matching accuracy of 1.00 corresponds to exact recovery of the ground-truth class permutation. We compared against BlindMatch [19] using both Gromov–Wasserstein ($(\mathcal{GW})$) and Mutual kNN. RankAlign uses the same permutation solvers as the baselines but replaces the raw within-modality similarity kernels with the ordinal representation defined in Section 2.3 before matching. To isolate the contribution of this representation, we additionally compare against z-score normalization, quantile normalization, and row-wise ranking under an identical protocol. The controlled ablation in Appendix B.2 shows that the gains cannot be explained by generic normalization or local ranking alone. All experiments were conducted on a single NVIDIA RTX A6000 GPU with 48GB memory.

Pre-trained models. We evaluated RankAlign using a variety of frozen pre-trained visual and textual encoders to cover diverse representation geometries and supervision paradigms, including DINO/DINOv2 [1, 13], DeiT [21], and vision-language models such as CLIP [15], SigLIP [25], and SigLIP2 [22]. For language side, we primarily used SentenceTransformers models [17].

Prompt construction. Given an image classification dataset with N visual groups, we form class-level prototypes independently in the two modalities. On the visual side, each image is embedded with the frozen visual encoder and the embeddings are averaged within each visual group. On the language side, we construct an unordered candidate set of N textual concepts. For each candidate concept, we generate a small collection of semantically equivalent textual descriptions, embed them

with the frozen language encoder, and average them to obtain a textual prototype. The correspondence between the visual groups and textual prototypes is withheld from the matching algorithm. The algorithm receives only the two unordered sets of class-level prototypes and recovers their cross-modal permutation from within-modality relational structure.

Dataset We conducted our experiments on CIFAR-100 [9], ImageNet-100 [18], the medical imaging dataset CheXpert [7], and a pseudo-class dataset derived from the LAION-Aesthetic subset [20]. More details are provided in Appendix B.1.

3.1 Computational Cost

Because encoders were frozen, no end-to-end training was required. RankAlign introduces only an ordinal preprocessing step before permutation matching. As analyzed in Section 2.5, constructing the complete RankAlign representation requires $O(N^2 \log N)$ additional preprocessing time and $O(N^2)$ memory beyond the similarity-kernel construction. The transformed representations remain $N \times N$, so the overall computational cost is primarily determined by the QAP solver.

Table 1: Runtime and matching accuracy on CIFAR-100 across different problem sizes using CLIP as the vision encoder and Sentence-T as the language encoder. Results are reported as mean $\pm$ standard deviation over five random seeds. * indicates that the solver reaches the 300s time limit. Actual wall-clock runtime may be slightly higher due to termination and post-processing overhead.

Solver	Method	$N = 50$	$N = 70$	$N = 100$
		Matching accuracy		
FAQ	BLINDMATCH	0.26 $\pm$ 0.10	0.14 $\pm$ 0.09	0.04 $\pm$ 0.02
	RANKALIGN	0.98 $\pm$ 0.02	0.79 $\pm$ 0.05	0.66 $\pm$ 0.21
OT	BLINDMATCH	0.04 $\pm$ 0.01	0.02 $\pm$ 0.02	0.01 $\pm$ 0.01
	RANKALIGN	0.98 $\pm$ 0.02	0.79 $\pm$ 0.05	0.69 $\pm$ 0.18
FHG	BLINDMATCH	0.77 $\pm$ 0.02	0.42 $\pm$ 0.07	0.15 $\pm$ 0.03
	RANKALIGN	0.98 $\pm$ 0.02	0.91 $\pm$ 0.04	0.79 $\pm$ 0.02
		Runtime (s)		
FAQ	BLINDMATCH	0.54 $\pm$ 0.18	0.70 $\pm$ 0.12	1.44 $\pm$ 0.21
	RANKALIGN	0.36 $\pm$ 0.02	1.10 $\pm$ 0.03	2.70 $\pm$ 0.31
OT	BLINDMATCH	0.12 $\pm$ 0.03	0.28 $\pm$ 0.01	0.82 $\pm$ 0.01
	RANKALIGN	0.42 $\pm$ 0.04	1.20 $\pm$ 0.03	2.80 $\pm$ 0.28
FHG	BLINDMATCH	300.00*	300.00*	300.00*
	RANKALIGN	300.00*	300.00*	300.00*

Table 1 reports runtime and matching accuracy on CIFAR-100 with respect to the problem size N. When the same solver is used, RankAlign introduces only a modest runtime overhead while providing substantially higher matching accuracy. The OT-based solver remains computationally efficient across all tested problem sizes. In contrast, FHG reaches the runtime limit at larger problem sizes in our implementation. These results indicate that the main scalability bottleneck comes from permutation optimization rather than the rank transformation itself.

3.2 Scalability

We tested scalability on CIFAR-100 and ImageNet-100 by increasing the problem size N, where larger N requires finer-grained semantic discrimination and yields a more challenging matching problem. Following [19], we selected class subsets using a p-dispersion strategy, which favors subsets whose relational structures are well aligned across modalities. Specifically, let $S \subseteq \{1, \ldots, L\}$ be a subset of classes with $|S| = N$. We can get a subset score

$$A(S) = \sum_{i,j \in S} \ell(K_x(i,j), K_y(i,j)) \,,$$

which measures the agreement between visual and language relational kernels restricted to S. We then select subsets by approximately solving $S^\star \in \arg\max_{S:|S|=N} A(S)$, which is an instance of the p-dispersion-sum problem. This procedure yielded class subsets where cross-modal relational structure

is more consistent, making non-trivial matching feasible. For each problem size N, i.e., the number of classes to be matched, we constructed matching instances from the selected subsets and reported results across different visual encoders. We use the same solver of FHG [19] for a fair comparison. Fig. 2 shows that the matching problem becomes progressively more challenging as N increases, particularly at larger problem sizes. At smaller N, RANKALIGN reaches a matching accuracy of 1 for several backbone settings. Since matching accuracy measures the fraction of correctly recovered class correspondences, a value of 1 means that the recovered permutation exactly matches the ground-truth permutation, i.e., $\hat{\pi} = \pi^\star$. This should be interpreted as exact correspondence recovery for the given matching instance rather than as an indication of overfitting. This observation is consistent with Proposition 1, which shows that under strict ordinal consistency, $\pi^\star$ is the unique maximizer of the ordinal matching objective. As N grows, RANKALIGN consistently outperforms BLINDMATCH($\mathcal{GW}$), with the advantage becoming increasingly pronounced at larger problem sizes. Most notably, at $N = 100$, RANKALIGN retains approximately $30\% - 60\%$ matching accuracy across the representative backbones on CIFAR-100 and ImageNet-100, while BLINDMATCH($\mathcal{GW}$) falls to near-chance levels. This yields an improvement of roughly $45\% - 70\%$ at the largest evaluated problem size.

These observations are consistent with the recovery analysis in Section 2.4. Ordinal consistency yields unique recovery of the ground-truth permutation, while under partial disagreement recovery depends on the balance between cross-modal ordinal agreement and the distinguishability of class relational profiles. The widening performance gap suggests that ordinal relational structure remains informative as correspondence recovery requires increasingly fine-grained discrimination among classes. For clarity, Fig. 2 reports representative results for CLIP, DINOv2, and SigLIP2, while results across all evaluated vision backbones are provided in Appendix B.4. Additional comparisons in Appendix B.3 show that classical Orthogonal Procrustes and vanilla Gromov–Wasserstein optimal transport remain substantially below RankAlign under the same representations and matching instances.

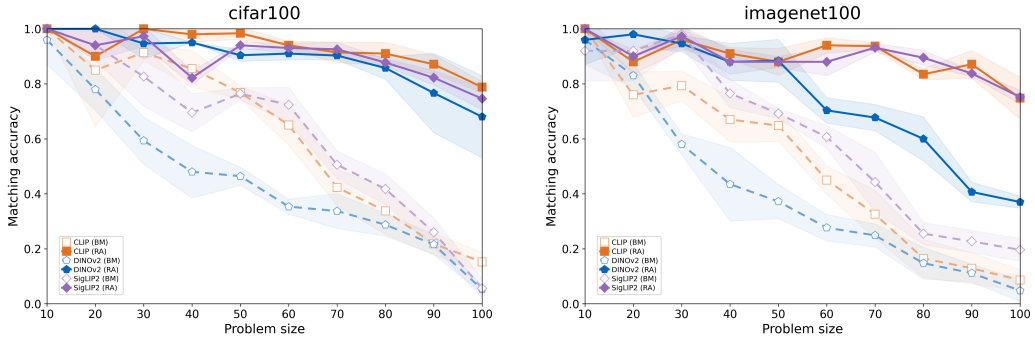

Figure 2: Matching accuracy as a function of problem size on CIFAR-100 and ImageNet-100. Each color corresponds to a vision backbone. For each backbone, the darker solid line denotes RANKALIGN, while the lighter dashed line denotes BLINDMATCH with $\mathcal{GW}$.

3.3 Robustness to label corruption

Setup. We conducted this experiment on CIFAR-100 and ImageNet-100 with $N = 100$ classes. For each corruption rate $r \in [0, 1]$, we randomly assigned a fraction r of image embeddings within each class to incorrect classes. Let S_c be the set of samples originally assigned to class c, and let $m_c = \lfloor r|S_c| \rfloor$ be the number of corrupted samples from class c. This produces corrupted labels $\tilde{y}_n$, while the language prototypes $\{t_c\}_{c=1}^{C}$ remain unchanged. Using the corrupted labels, we recomputed the visual class centroids as

$$\tilde{v}_c = \frac{1}{|\{n : \tilde{y}_n = c\}|} \sum_{n:\tilde{y}_n=c} z_n,$$

where z_n is the visual embedding of image n. We performed cross-modal matching between the corrupted visual prototypes $\{\tilde{v}_c\}_{c=1}^{C}$ and the clean language prototypes.

Let π denote the recovered permutation. We report the class-level matching accuracy

$$\mathrm{Acc}_{\mathrm{match}} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{1}[\pi(c) = c]\,.$$

Each experiment was repeated over multiple random seeds and we report mean and standard deviation. Robustness is quantified by the relative drop from the clean setting:

$$\mathrm{Drop}(r) = \frac{\mathrm{Acc}(0) - \mathrm{Acc}(r)}{\mathrm{Acc}(0)} \times 100\%,$$

where $\mathrm{Acc}(r)$ denotes the mean matching accuracy at corruption rate r.

Table 2: Robustness under visual-side label corruption on CIFAR-100 and ImageNet-100. We report matching accuracy (%) as mean $\pm$ standard deviation over five random seeds and relative performance drop (%) with respect to the clean setting.

Dataset	Metric	0.00	0.10	0.20	0.30	0.40
CIFAR-100	Acc. (%)	78.7 ± 7.3	78.0 ± 5.2	77.6 ± 5.8	73.1 ± 15.9	68.9 ± 6.4
	Drop (%)	0.0	0.9	1.4	7.1	12.5
ImageNet-100	Acc. (%)	74.4 ± 5.4	71.0 ± 3.9	71.4 ± 5.4	56.6 ± 12.0	41.5 ± 10.3
	Drop (%)	0.0	4.6	4.0	23.9	44.2

Table 2 reports robustness under increasing visual-side label corruption on CIFAR-100 and ImageNet-100. As the corruption rate increases, matching accuracy decreases on both datasets, reflecting the impact of noisy class assignments on prototype quality. The degradation remains gradual across a wide range of corruption levels, indicating that our method maintains reasonable performance even under moderate noise. Similar trends are observed on both datasets, suggesting that the approach is consistently robust across different data distributions.

3.4 LAION Pseudo-Class Matching

We further evaluate RankAlign on the LAION pseudo-class dataset, which provides a noisy image–text setting without manually defined class annotations. We compare RANKALIGN against BLINDMATCH($\mathcal{GW}$) across different visual backbones.

Table 3: Matching accuracy on the LAION pseudo-class dataset. We compare BLINDMATCH($\mathcal{GW}$) with RANKALIGN across different visual backbones.

Vision model	BLINDMATCH	RANKALIGN
DINOv2	0.59 $\pm$ 0.29	0.96 $\pm$ 0.09
DeiT	0.72 $\pm$ 0.05	0.76 $\pm$ 0.09
CLIP	0.77 $\pm$ 0.13	0.68 $\pm$ 0.05
SigLIP	0.54 $\pm$ 0.13	0.79 $\pm$ 0.15
SigLIP2	0.61 $\pm$ 0.09	0.74 $\pm$ 0.11

As shown in Table 3, RANKALIGN achieves stronger overall matching performance than BLINDMATCH($\mathcal{GW}$) across the evaluated visual backbones. The results indicate that rank-based relational alignment is generally more robust in this noisy pseudo-class setting, where class definitions are derived from weakly supervised image–text signals rather than manually curated labels. The consistent gains across different backbone families further suggest that the benefit of RankAlign is not tied to a specific visual encoder.

3.5 Medical Domain Evaluation

We also evaluate RankAlign on CheXpert, which introduces a substantial domain gap between medical images and language descriptions, together with subtle inter-class differences and class imbalance. Compared with natural-image benchmarks, CheXpert presents a more challenging cross-modal matching setting because different pathological categories may exhibit highly similar visual

patterns, while the corresponding language descriptions encode fine-grained clinical distinctions. Moreover, the uneven class distribution further perturbs class-level prototype estimation, making reliable correspondence recovery more difficult. Table 4 reports matching accuracy across different visual backbones. We compare RANKALIGN against BLINDMATCH($\mathcal{GW}$) under the same evaluation setting. Across all evaluated backbones, RANKALIGN achieves substantially higher matching accuracy, indicating that the rank-based relational representation remains effective under pronounced domain shift and challenging medical-image semantics.

Table 4: Matching performance on the CheXpert subset. We report matching accuracy across different vision backbones for BLINDMATCH($\mathcal{GW}$) and RANKALIGN.

Vision model	BLINDMATCH	RANKALIGN
DeiT	0.08 $\pm$ 0.01	0.65 $\pm$ 0.01
DINOv2	0.06 $\pm$ 0.01	0.62 $\pm$ 0.30
CLIP	0.08 $\pm$ 0.01	0.63 $\pm$ 0.26
SigLIP	0.10 $\pm$ 0.02	0.45 $\pm$ 0.22
SigLIP2	0.08 $\pm$ 0.00	0.23 $\pm$ 0.04

4 Conclusions

This work revisits class-level unpaired vision-language matching from an ordinal perspective and introduces RankAlign, a rank-based normalization strategy for relational correspondence. Rather than modifying the underlying permutation-matching framework, RankAlign changes the relational representation supplied to the matching objective. By emphasizing relational order over absolute similarity magnitudes, it provides a monotone-invariant signal for recovering cross-modal correspondences without cross-modal supervision. Beyond its empirical effectiveness, these findings offer strong support for the Platonic Representation Hypothesis, demonstrating that cross-modal convergence does not require rigid metric equivalence, but rather emerges naturally through shared relational topologies across independently trained representations. A fundamental property of rank-based normalization lies in its complete invariance to strictly increasing reparameterizations, providing provable guarantees against modality-specific scale distortions, domain shifts, and monotonic noise. These theoretical properties translate directly into empirical robustness: across diverse frozen encoders and matching settings, RankAlign delivers stable class-level correspondence recovery even under severe corruption conditions. By formulating cross-modal correspondence in terms of relational order rather than absolute similarity magnitudes, the approach ensures invariant alignment without relying on cross-modal anchors or paired supervision.

Limitations RankAlign relies on informative class-level prototypes in both modalities. When visual groups are estimated from noisy clustering rather than reliable class labels, prototype errors can distort the induced relational structure and reduce ordinal agreement, and our experiments suggest that prototype quality is a major source of degradation, although larger and more fine-grained matching problems may introduce additional difficulty. The permutation-based matching stage can also become computationally expensive as the number of classes grows. Moreover, RankAlign focuses on class-level zero-anchor correspondence and intentionally discards absolute similarity magnitudes, so instance-level, open-vocabulary, and hybrid ordinal–magnitude alignment remain directions for future work.

Broader Impact RankAlign may reduce the need for manually verified cross-modal correspondences in settings where paired multimodal supervision is limited or expensive. Potential applications include representation analysis, class-level retrieval, and domain-specific multimodal matching. As with other alignment methods, unreliable prototypes or biases in pretrained models may lead to incorrect correspondences, so applications in sensitive domains should include appropriate validation.

Acknowledgments

We are grateful to Francis Engelmann for carefully reviewing the manuscript and providing valuable suggestions that helped improve the clarity of the paper.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 6

[2] Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. On kernel-target alignment. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2001. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 1

[4] Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Proc. of the International Conference on Machine Learning (ICML)*, 2026. 2

[5] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. 1

[6] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proc. of the International Conference on Machine Learning (ICML)*, 2024. 1

[7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2019. 7, 17

[8] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proc. of the International Conference on Machine Learning (ICML)*, 2019. 2

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 7, 17

[10] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[11] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2023. 2

[12] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transaction of Machine Learing Research (TMLR)*, 2024. 6

[14] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proc. of the International Conference on Machine Learning (ICML)*, 2016. 2

[15] 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 *Proc. of the International Conference on Machine Learning (ICML)*, 2021. 1, 6

[16] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[17] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 1, 6

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 7, 17

[19] Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3, 6, 7, 8

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 7, 17

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021. 6

[22] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Henaff, 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. 1, 6

[23] 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), 2015. 2

[24] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov-Wasserstein Learning for Graph Matching and Node Embedding. In *Proc. of the International Conference on Machine Learning (ICML)*, 2019. 2

[25] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

A Technical appendices and supplementary material

In this appendix, we provide further theoretical results and detailed proofs in §A. Additionally, we report supplementary empirical results and analyses in §B.

This appendix provides proofs and additional details for the theoretical results in the main text. We first establish the relation between relational distortion and agreement and prove the invariance properties of the rank representation. We then describe the full ordinal refinement used in the experiments. Next, we prove the exact-recovery and stability results and connect them to this refined representation and approximate solver outputs. Finally, we provide additional computational complexity details.

A.1 Properties of the ordinal representation

We first prove the relation between the distortion formulation and the agreement score introduced in Section 2.1.

Proof of Lemma 1. For $\ell(a, b) = -ab$,

$$D_\ell(K_x, K_y; \pi) = -\sum_{i \neq j} K_x(i, j) K_y(\pi(i), \pi(j)) = -f(\pi).$$

Hence minimizing D_ℓ is equivalent to maximizing f.

For $\ell(a, b) = (a - b)^2$,

$$D_\ell(K_x, K_y; \pi) = \sum_{i \neq j} \left(K_x(i, j) - K_y(\pi(i), \pi(j))\right)^2 \tag{22}$$

$$= \sum_{i \neq j} K_x(i, j)^2 + \sum_{i \neq j} K_y(\pi(i), \pi(j))^2 \tag{23}$$

$$- 2 \sum_{i \neq j} K_x(i, j) K_y(\pi(i), \pi(j)). \tag{24}$$

Because $(i, j) \mapsto (\pi(i), \pi(j))$ is a bijection over the ordered off-diagonal pairs,

$$\sum_{i \neq j} K_y(\pi(i), \pi(j))^2 = \sum_{i \neq j} K_y(i, j)^2.$$

The first two terms are therefore independent of π, which proves the equivalence.

When the off-diagonal entries of both kernels have zero mean and unit variance, define

$$\rho(\pi) = \frac{1}{N(N-1)} \sum_{i \neq j} K_x(i, j) K_y(\pi(i), \pi(j)).$$

Then

$$f(\pi) = N(N-1)\rho(\pi),$$

and

$$D_\ell(K_x, K_y; \pi) = 2N(N-1)(1 - \rho(\pi)).$$

Thus maximizing the agreement score is equivalent to maximizing the correlation between the two relational structures.

Proof of Lemma 2. Let ϕ be strictly increasing. For any observed kernel entries,

$$\phi(K(k, l)) \leq \phi(K(i, j))$$

if and only if

$$K(k, l) \leq K(i, j).$$

Therefore

$$F_{\phi \circ K}(\phi(K(i, j))) = F_K(K(i, j)),$$

which implies

$$R(\phi \circ K) = R(K).$$

Now let $\sigma \in S_N$. Relabeling the classes induces a bijection on the unordered-pair set $\mathcal{E}$ and therefore leaves the multiset of off-diagonal kernel values unchanged. Hence

$$R(K \circ \sigma)(i,j) = R(K)(\sigma(i), \sigma(j)),$$

and therefore

$$R(K \circ \sigma) = R(K) \circ \sigma.$$

The rank transformation also preserves strict pairwise order. For $e, e' \in \mathcal{E}$,

$$K(e) < K(e') \quad \Longleftrightarrow \quad R(K)(e) < R(K)(e').$$

If all $M = N(N-1)/2$ unordered off-diagonal values are distinct, the rank values are $1/M, 2/M, \ldots, 1$. If ties occur, equal values remain tied. Rank transformation therefore preserves the ordinal information already present in the kernel rather than creating additional relational distinctions.

The theoretical analysis in the main text is stated for the basic rank representation $R(K)$. In the experiments, we use a fixed ordinal refinement $T(K) = h(R(K))$, which combines kernel-level ranks with anchor-wise neighbor order. For ranks computed over m candidate relations, we use the monotone reweighting

$$g_m(u) = -\log\left(1 - u + \frac{1}{2m}\right). \tag{25}$$

Because g_m is strictly increasing, it changes the relative weighting of rank positions without changing their ordering. For each anchor i, define the within-row rank

$$R_{\mathrm{row}}(K)(i,j) = \frac{1}{N-1} \sum_{l \neq i} \mathbf{1}\left[R(K)(i,l) \leq R(K)(i,j)\right], \qquad i \neq j, \tag{26}$$

with zero diagonal. Then let $\mathcal{C}(A)$ denote standardization over the off-diagonal entries of a symmetric matrix A,

$$\mathcal{C}(A)(i,j) = \frac{A(i,j) - \mu_A}{s_A}, \qquad i \neq j, \tag{27}$$

with zero diagonal, where μ_A and s_A are the mean and standard deviation of the off-diagonal entries. We define

$$G(K) = \mathcal{C}(g_M(R(K))), \tag{28}$$

and

$$L(K) = \mathcal{C}(\mathrm{sym}\, g_{N-1}(R_{\mathrm{row}}(K))), \tag{29}$$

where $\mathrm{sym}(A) = \frac{A + A^\top}{2}$.

The RankAlign representation is then $T(K) = \mathcal{C}(G(K) + L(K))$.

The two components are standardized before aggregation so that their relative contribution is not determined solely by numerical scale. Importantly, all operations in this refinement depend only on ordinal information. Consequently, $T(\phi \circ K) = T(K)$ for every strictly increasing ϕ, and $T(K \circ \sigma) = T(K) \circ \sigma$ for every permutation $\sigma \in S_N$. Thus, the refined representation inherits the monotone invariance and permutation equivariance of the basic rank representation without reintroducing the original similarity magnitudes.

A.2 Recovery guarantees and approximate optimization

We first prove the exact-recovery result for the basic rank objective in Section 2.4.

Proof of Proposition 1. Let $X = R(K_x)$, Under Assumption 1, the two kernels induce the same strict ordering under $\pi^\star$. Hence

$$R(K_y) \circ \pi^\star = X.$$

For an arbitrary candidate permutation π, define

$$\tau = (\pi^\star)^{-1} \circ \pi.$$

Then

$$R(K_y) \circ \pi = X \circ \tau.$$

Therefore,

$$f_{\mathrm{ord}}(\pi^\star) - f_{\mathrm{ord}}(\pi) = \|X\|_F^2 - \langle X, X \circ \tau \rangle \tag{30}$$

$$= \frac{1}{2}\|X - X \circ \tau\|_F^2. \tag{31}$$

The right-hand side is non-negative for every π, so $\pi^\star$ is a maximizer.

We next show uniqueness and the explicit margin. The permutation τ induces a permutation of the unordered-pair set $\mathcal{E}$. If every unordered pair were fixed, then for any i and any two distinct $j, k \neq i$,

$$\tau(i) \in \{i, j\} \cap \{i, k\} = \{i\},$$

so $\tau = \mathrm{id}$. Thus every non-identity class permutation with $N \geq 3$ changes at least one non-trivial cycle of unordered pairs.

Under Assumption 1, the unordered rank values are $1/M, \ldots, 1$ and are pairwise distinct. A non-trivial cycle contains at least two unordered pairs, and each changed rank differs by at least $1/M$. Since every unordered pair contributes twice to the ordered off-diagonal objective,

$$\|X - X \circ \tau\|_F^2 \geq \frac{4}{M^2}.$$

Hence

$$f_{\mathrm{ord}}(\pi^\star) - f_{\mathrm{ord}}(\pi) \geq \frac{2}{M^2}.$$

Therefore $\pi^\star$ is the unique maximizer. If the objective is instead written over unordered pairs only, each relation appears once, giving the corresponding margin $1/M^2$.

The exact-recovery result above is stated for the basic rank representation. Under Assumption 1, the two modalities differ only by a strictly increasing reparameterization under the ground-truth correspondence. Since the refined representation $T(K)$ is monotone invariant and permutation equivariant,

$$T(K_y) \circ \pi^\star = T(K_x). \tag{32}$$

Hence $\pi^\star$ also maximizes the practical RankAlign objective. It is unique whenever the transformed representation has no non-trivial permutation symmetry. We now prove the stability result in the representation-independent form used in the main text. Let

$$Z_x = Z(K_x), \qquad Z_y = Z(K_y),$$

and define $E_Z = Z_y \circ \pi^\star - Z_x$. For $\tau = (\pi^\star)^{-1} \circ \pi$, let $d_Z(\tau) = \|Z_x - Z_x \circ \tau\|_F$.

Proof of Proposition 2. Since $Z_y \circ \pi^\star = Z_x + E_Z$, we have $Z_y \circ \pi = (Z_x + E_Z) \circ \tau$. Therefore,

$$f_Z(\pi^\star) - f_Z(\pi) = \langle Z_x, Z_x + E_Z \rangle - \langle Z_x, (Z_x + E_Z) \circ \tau \rangle \tag{33}$$

$$= \|Z_x\|_F^2 - \langle Z_x, Z_x \circ \tau \rangle + \langle Z_x - Z_x \circ \tau^{-1}, E_Z \rangle. \tag{34}$$

The first two terms satisfy

$$\|Z_x\|_F^2 - \langle Z_x, Z_x \circ \tau \rangle = \frac{1}{2} d_Z(\tau)^2.$$

By Cauchy–Schwarz,

$$\langle Z_x - Z_x \circ \tau^{-1}, E_Z \rangle \geq -d_Z(\tau)\|E_Z\|_F.$$

Hence

$$f_Z(\pi^\star) - f_Z(\pi) \geq d_Z(\tau)\left(\frac{1}{2} d_Z(\tau) - \|E_Z\|_F\right).$$

If $\hat{\pi}$ maximizes f_Z, then we have

$$f_Z(\pi^\star) - f_Z(\hat{\pi}) \leq 0,$$

and

$$d_Z(\hat{\tau}) \leq 2\|E_Z\|_F, \qquad \hat{\tau} = (\pi^\star)^{-1} \circ \hat{\pi}.$$

The stability result also yields a bound on the number of mismatched classes. Define

$$\mathrm{err}(\pi) = |\{i : \pi(i) \neq \pi^\star(i)\}|$$

and

$$\kappa(Z_x) = \min_{\tau \neq \mathrm{id}} \frac{\|Z_x - Z_x \circ \tau\|_F^2}{|\{i : \tau(i) \neq i\}|}.$$

Then

$$\kappa(Z_x)\,\mathrm{err}(\hat{\pi}) \leq d_Z(\hat{\tau})^2,$$

which gives

$$\mathrm{err}(\hat{\pi}) \leq \frac{4\|E_Z\|_F^2}{\kappa(Z_x)}. \tag{35}$$

For the standardized representation T, define

$$\rho^\star = \frac{\langle T_x, T_y \circ \pi^\star \rangle}{N(N-1)}.$$

Because the off-diagonal entries are standardized to zero mean and unit variance,

$$\|E_T\|_F^2 = 2N(N-1)(1-\rho^\star).$$

Substituting this expression into Eq. 35 gives

$$\mathrm{err}(\hat{\pi}) \leq \frac{8N(N-1)(1-\rho^\star)}{\kappa(T_x)}.$$

This separates cross-modal ordinal disagreement, measured by $\rho^\star$, from the intrinsic distinguishability of the relational profiles, measured by $\kappa(T_x)$.

Finally, the previous result extends to an approximate QAP solver. Let $\hat{\pi}$ denote an arbitrary solver output and define its objective deficit relative to the ground truth as

$$\eta = \max\{0, f_Z(\pi^\star) - f_Z(\hat{\pi})\}.$$

From the stability inequality,

$$d_Z(\hat{\tau}) \left(\frac{1}{2} d_Z(\hat{\tau}) - \|E_Z\|_F \right) \leq \eta.$$

Solving the resulting quadratic inequality gives

$$d_Z(\hat{\tau}) \leq \|E_Z\|_F + \sqrt{\|E_Z\|_F^2 + 2\eta}.$$

Hence,

$$\mathrm{err}(\hat{\pi}) \leq \frac{\left(\|E_Z\|_F + \sqrt{\|E_Z\|_F^2 + 2\eta} \right)^2}{\kappa(Z_x)}.$$

The analysis therefore separates three sources of difficulty. Cross-modal disagreement is captured by $\|E_Z\|_F$, structural ambiguity by $\kappa(Z_x)$, and optimization error by η.

A.3 Computational complexity

We finally provide a detailed accounting of the computational cost of RankAlign. Let N denote the number of classes and d the embedding dimension. The number of unordered class pairs is

$$M = \frac{N(N-1)}{2}.$$

Prototype centering and normalization require $O(Nd)$ time. Constructing each dense similarity kernel requires $O(N^2 d)$ time and $O(N^2)$ memory. The kernel-level rank transformation sorts $M = O(N^2)$ values and requires $O(M \log M) = O(N^2 \log N)$ time.

The anchor-wise transformation ranks N rows of length $N-1$ and also requires $O(N^2 \log N)$ time. The score transformation, symmetrization, and standardization each operate over at most $O(N^2)$ entries and therefore require $O(N^2)$ time. Consequently, the complete practical RankAlign

representation can be constructed in $O(N^2d + N^2 \log N)$ time and $O(N^2 + Nd)$ memory, excluding feature extraction.

The $O(N^2d)$ kernel construction is already required by conventional value-based relational matching. The additional asymptotic preprocessing overhead introduced by RankAlign is $O(N^2 \log N)$ time and $O(N^2)$ memory. Let $C_{\mathcal{S}}(N)$ denote the running time of a chosen QAP solver $\mathcal{S}$. The complete pipeline requires $C_{\mathcal{S}}(N) + O(N^2d + N^2 \log N)$ time.

The transformed representation remains an $N \times N$ matrix. RankAlign therefore leaves the input structure of the QAP unchanged and can be combined with the same solver used by value-based relational matching. Evaluating the complete matching objective for one permutation requires $O(N^2)$ operations. If a local-search solver swaps two assignments, only terms involving the affected classes change, so the corresponding objective update involves $O(N)$ pairwise terms. We use an iterated local search (ILS) procedure to obtain a strong initial solution before applying the selected QAP solver. The ILS optimizes the rank-based objective and its solution is used to initialize FAQ, entropic OT, or FHG. The same initialization procedure is applied to BlindMatch for a fair comparison.

B Additional experiments

In this section, we provide supplementary experimental results supporting the main paper.

B.1 Datasets

We evaluate RankAlign on diverse benchmarks spanning web-scale pseudo-label data, object recognition datasets at different semantic granularities, and a domain-specific medical imaging dataset.

CIFAR-100 and ImageNet-100. To study scalability, we evaluated the matching accuracy as a function of the number of classes N using subsets of CIFAR-100 [9] and ImageNet-100 [18]. We varied N across matching instances to assess performance under increasing task difficulty.

LAION pseudo-class dataset. We evaluated RankAlign on a pseudo-class dataset derived from the LAION-Aesthetic subset [20]. Since LAION lacks predefined class labels, we defined $N = 20$ visual categories and assign pseudo-labels via caption matching. To ensure label quality, we classified each image using both its original caption (`TEXT`) and a `Moondream`-generated auxiliary caption, retaining only samples with a unique, consistent top-scoring assignment. The dataset contains approximately 7,000 samples, with 200–500 images per class. Visual and language prototypes are constructed by averaging embeddings within each pseudo-class and its associated captions, respectively.

Medical imaging dataset. We evaluated RankAlign on CheXpert [7], a large-scale chest X-ray benchmark with 220,000 radiographs. We extracted a 12-class single-label subset by retaining samples with one positive finding, totaling about 62,000 images. This setting is challenging due to subtle inter-class differences, class imbalance, and the medical image-text domain gap.

B.2 Kernel Transformation Ablation

To determine whether the gains of RankAlign can be explained by generic normalization or by a single form of ranking, we conduct a controlled ablation on CIFAR-100. All variants use identical visual and language prototypes, class subsets, similarity kernels, encoders, and permutation solvers. Only the relational representation supplied to the solver is changed.

We compare the raw kernel without transformation (none), z-score normalization, row-wise ranking, quantile normalization, and the full RankAlign representation. Z-score normalization removes global mean and scale differences while preserving the relative spacing of kernel values. Row-wise ranking preserves local neighbor order but does not define a single kernel-level ordinal ordering. Quantile normalization aligns marginal value distributions.

Table 5 shows that the improvement cannot be explained by normalization alone. Z-score normalization provides almost no gain over the raw value-based kernel, consistent with the fact that affine normalization does not alter the optimizer of the inner-product matching objective. Quantile normalization also yields limited improvements. Row-wise ranking performs substantially better than value-based normalization, confirming that local ordinal information is useful. However, its advantage decreases as the problem size grows. RankAlign consistently achieves the highest matching accuracy

Table 5: Kernel transformation ablation on CIFAR-100. All variants use the same prototypes, class subsets, encoders, kernels, and permutation solver, only the kernel transformation differs.

Transformation	$N = 30$	$N = 50$	$N = 70$	$N = 100$
None	0.03	0.01	0.00	0.00
Z-score	0.03	0.02	0.01	0.00
Row-wise ranking	0.32	0.14	0.09	0.06
Quantile normalization	0.06	0.07	0.04	0.00
RankAlign	1.00	0.98	0.91	0.79

across all problem sizes, indicating that the combination of kernel-level ordinal normalization and anchor-wise neighborhood order is more effective than generic normalization or local ranking alone.

B.3 Comparison with Classical Structural Alignment Methods

To examine whether simpler classical alignment methods can recover the same cross-modal correspondence without RankAlign's ordinal formulation, we compare against unsupervised Orthogonal Procrustes and standard entropic Gromov–Wasserstein optimal transport. All methods use the same class-level visual and language representations and are evaluated on the same CIFAR-100 matching instances, so the comparison isolates differences in the alignment objective rather than differences in the underlying encoders or prototypes.

Table 6: Comparison with classical structural alignment methods on CIFAR-100. All methods use the same class-level representations and matching instances. We report matching accuracy across problem sizes.

Model	Method	$N = 30$	$N = 50$	$N = 70$	$N = 100$
CLIP	Procrustes	0.01 ± 0.02	0.03 ± 0.01	0.02 ± 0.02	0.00 ± 0.00
	Vanilla GW-OT	0.03 ± 0.03	0.03 ± 0.02	0.03 ± 0.03	0.03 ± 0.00
	RankAlign	1.00 ± 0.00	0.98 ± 0.02	0.91 ± 0.04	0.79 ± 0.02
DeiT	Procrustes	0.04 ± 0.04	0.02 ± 0.01	0.02 ± 0.01	0.01 ± 0.00
	Vanilla GW-OT	0.07 ± 0.04	0.05 ± 0.03	0.07 ± 0.05	0.05 ± 0.00
	RankAlign	0.93 ± 0.01	0.95 ± 0.05	0.72 ± 0.05	0.36 ± 0.16
DINOv2	Procrustes	0.05 ± 0.04	0.02 ± 0.02	0.01 ± 0.01	0.04 ± 0.00
	Vanilla GW-OT	0.01 ± 0.02	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.00
	RankAlign	0.95 ± 0.06	0.90 ± 0.02	0.90 ± 0.04	0.68 ± 0.15

Table 6 shows that neither classical baseline approaches the performance of RankAlign across the evaluated backbones and problem sizes. Both Orthogonal Procrustes and vanilla GW-OT remain close to chance as N increases, whereas RankAlign retains substantially higher matching accuracy. Because all methods operate on the same class-level representations, this difference cannot be attributed to stronger encoders or higher-quality prototypes. These results indicate that a global linear alignment or direct matching of the original metric structures is insufficient to recover the cross-modal correspondence in these experiments, and RankAlign is not replacing a comparably accurate simpler baseline.

B.4 Full larger-scale matching results

In the main paper, we report representative larger-scale matching results for a subset of vision backbones to improve readability. We provide the full results across all evaluated backbones, including CLIP, DINOv2, DeiT, SigLIP, and SigLIP2.

Figure 3 shows matching accuracy as a function of the problem size on CIFAR-100 and ImageNet-100. The overall trend is consistent with the representative results in the main text, RANKALIGN consistently outperforms BLINDMATCH with the $\mathcal{GW}$ objective across different backbone families, and its performance degrades more gradually as the number of matched classes increases. These results indicate that the improvement is not tied to a particular visual encoder, but holds across both vision-language and visual representation models.

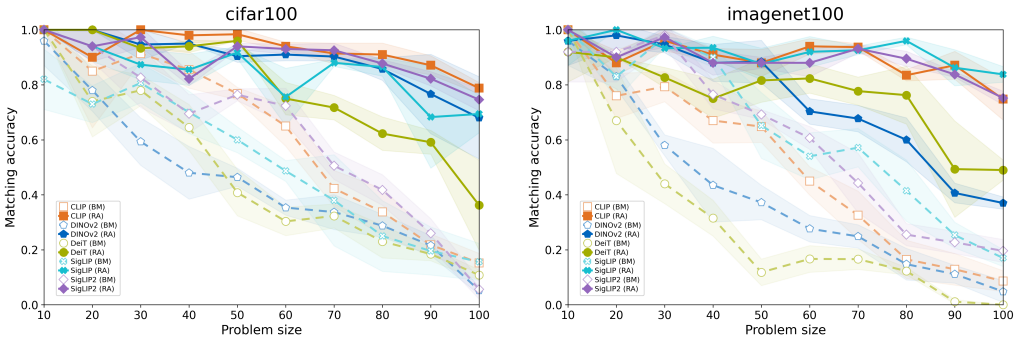

Figure 3: Full larger-scale matching results on CIFAR-100 and ImageNet-100 across all evaluated vision backbones.Each color corresponds to a vision backbone. The darker solid line denotes RANKALIGN, while the lighter dashed line denotes BLINDMATCH with the $\mathcal{GW}$.

B.5 Existing Assets and Licenses

This work uses existing public datasets and pretrained models, including LAION-Aesthetic, CIFAR-100, ImageNet-100, CheXpert, DINO/DINOv2, DeiT, CLIP, SigLIP/SigLIP2, and SentenceTransformers. We cite the original sources for these assets and use them only for research evaluation. We will follow the corresponding licenses and terms of use for each dataset and pretrained model, and will include detailed license information in the released code and supplemental documentation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the scope of the paper, namely unsupervised vision-language representation alignment via rank-based relational transformation. The main claims are supported by the theoretical analysis in Section 2 and the empirical results in Section 3, including experiments on LAION dataset, CIFAR-100, ImageNet-100, and CheXpert.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations subsection that discusses the dependence of RankAlign on reliable class-level prototypes, the scalability of permutation-based matching, and the current focus on class-level alignment. It also identifies instance-level and open-vocabulary alignment as directions for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for the theoretical results in the corresponding theorem and proposition statements, including the random permutation setting and the finite value-diversity setting. Complete proofs are provided in Appendix A for Theorem 1 and Proposition 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the datasets, frozen vision and language encoders, prototype construction procedure, matching objectives, evaluation metric, and reporting protocol used for the main experiments. Additional experimental details and supplementary analyses are provided in the appendix to support reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We publicly release the code at https://github.com/zez356/RankAlign, along with detailed instructions for reproducing the main experimental results. All datasets used in our experiments are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, pretrained vision and language encoders, class-level prototype construction, matching objectives, baselines, evaluation metrics, and reporting protocol used in the experiments. Additional settings, including robustness experiments, solver comparisons, subset construction, and retrieval/classification protocols, are described in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports mean and standard deviation across multiple random seeds for the main experimental results. Tables and figures include variability information where applicable, and the text states that the reported uncertainty corresponds to standard deviation over repeated runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the compute environment used for the experiments. It also clarifies that the main computational cost comes from feature extraction, class-level prototype construction, kernel computation, and permutation-based matching rather than end-to-end model training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research uses existing datasets and pretrained models for unsupervised vision-language representation alignment and does not involve human-subject experiments, private data collection, or deployment of high-risk systems. We have reviewed the NeurIPS Code of Ethics and believe that the work conforms to its requirements.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses broader impacts in the appendix, including potential benefits in reducing reliance on paired image-text annotations, lowering data collection costs, and enabling better use of unpaired or weakly annotated data. It also notes that applications in sensitive domains require careful validation due to possible unreliable correspondences or biases inherited from datasets and pretrained models.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new data or models that pose a high risk for misuse, such as pretrained language models, image generators, or newly scraped datasets. The work is methodological and evaluates RankAlign using existing datasets and pretrained encoders.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources of the existing datasets and pretrained models used in the experiments, including LAION-Aesthetic, CIFAR-100, ImageNet-100, CheXpert, and the evaluated vision and language encoders. The appendix also states that these assets are used for research evaluation and that their corresponding licenses and terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, pretrained models, or other new assets. The experiments are conducted using existing datasets and pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. All experiments are conducted using existing datasets and pretrained models.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. Therefore, IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development does not involve LLMs as important, original, or non-standard components. Any use of LLMs, if applicable, was limited to writing, editing, or formatting and did not affect the methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.