ORTHOGONAL MAPS ARE SUFFICIENT TO ALIGN MULTIMODAL CONTRASTIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As models and data scale, independently trained networks often learn similar representations. Yet establishing similarity is weaker than identifying an explicit correspondence between representation spaces, especially for multimodal models, where consistency must hold both within modalities and across the learned coupling between modalities. We therefore ask whether two *independently* trained multimodal contrastive models, with encoders (f, g) and $(\tilde{f}, \tilde{g})$, trained on different data distributions and with different architectures, exhibit a systematic geometric relationship between their embedding spaces. Across model families including CLIP, SigLIP, FLAVA and CLAP, we find that this relationship is well approximated by a single orthogonal transformation, i.e., there exists an orthogonal map Q, with $Q^\top Q = I$, such that $\tilde{f}(x) \approx Qf(x)$ for paired images (audio) x. Strikingly, the *same* map also aligns the text encoders, so that $\tilde{g}(y) \approx Qg(y)$ for text inputs y. Theoretically, we prove that if two models agree on the multimodal kernel over a small finite anchor set, meaning $\langle f(x), g(y)\rangle \approx \langle \tilde{f}(x), \tilde{g}(y)\rangle$, then their image (audio) and text representations must be related by a *single* orthogonal map. We validate this premise, showing strong qualitative and quantitative kernel agreement. This geometric alignment enables backward-compatible model upgrades without costly re-embedding, and has implications for the privacy of learned representations.

1 INTRODUCTION

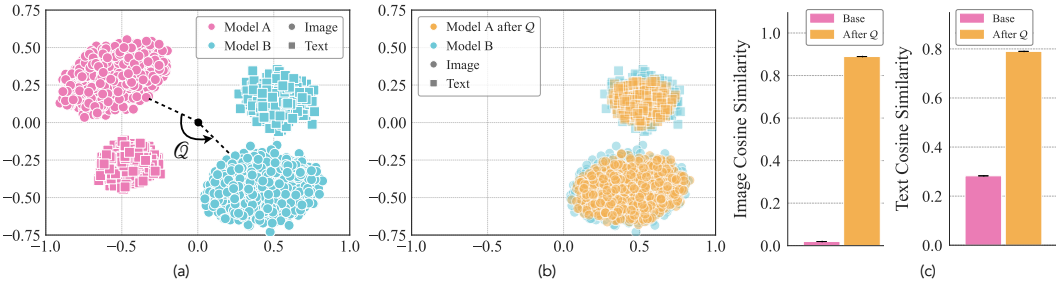

Figure 1: *A single orthogonal map aligns two independent contrastive models across modalities.* (a) Illustrative PCA schematic on synthetic embeddings from two models (A and B) shows that the spaces are a priori incomparable; (b) A single orthogonal map $\mathcal{Q}$, fit only using image embeddings of CLIP (OpenAI) and CLIP (LAION), almost perfectly aligns image embeddings of the two models (c, left) and simultaneously aligns their text embeddings, as evidenced by a large gain in text-text pointwise cosine similarity (c, right).

Whether independently trained neural networks recover similar internal representations is among the earliest questions in representation learning. Early studies of visual selectivity (Hubel & Wiesel, 1959), sparse coding (Olshausen & Field, 1996), and deep visual features (Krizhevsky et al., 2012) suggest that distinct learning systems can arrive at similar internal features. At the scale of foundation models, networks exhibit shared similarity structure over inputs or contain universal neurons and "circuits" (Huh et al., 2024; Merullo et al., 2022; Gupta et al., 2026; Chughtai et al., 2023; Gurnee et al., 2024; Zimmermann et al., 2021; Raghu et al., 2017; Kornblith et al., 2019). This observation

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.

underlies the *Platonic Representation Hypothesis* (PRH), which posits that, at sufficient scale, learned embeddings converge toward a shared statistical model of the world (Huh et al., 2024). Recent work, however, has called this into question. Gröger et al. (2026) argues that careful calibration leaves evidence for local neighborhood agreement rather than global convergence, while Koepke et al. (2026) finds that global agreement across models can weaken in larger and more realistic multimodal settings. Rather than asking whether all models recover a universal representation, we instead ask *what form that convergence takes when it does occur, and what geometry governs it.*

In this work, we focus on multimodal contrastive models, which couple image (or audio) and text through a contrastive objective while keeping the two modalities at arm's length in the learned space, a phenomenon often referred to as the *modality gap*. To make this precise, let $\mathcal{M} := (f, g)$ and $\tilde{\mathcal{M}} := (\tilde{f}, \tilde{g})$ denote two distinct multimodal models, with image (or audio, more generally) encoders $f, \tilde{f}$ and text encoders $g, \tilde{g}$ mapping inputs to embedding spaces. Here, it no longer suffices to ask whether representations from one modality i.e. f and $\tilde{f}$ of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ converge in isolation, or whether g and $\tilde{g}$ do. Instead, the central question becomes:

> *Given two independently trained multimodal contrastive models, what geometric relationship exists between their embeddings, and how does it differ across modalities?*

We find that this relationship is remarkably simple. Concretely, across model families such as CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), FLAVA (Singh et al., 2022) and CLAP (Elizalde et al., 2023), we find that the inter-model relationship is well-approximated by a *single orthogonal* map (up to a global mean shift) i.e., there exists an orthogonal map Q where $Q^\top Q = I$ such that $\tilde{f}(x) \approx Qf(x)$ for paired images x.[1] Moreover, the *same* Q, estimated only from image (audio) embeddings, simultaneously aligns text encoders: $\tilde{g}(y) \approx Qg(y)$ for texts y (as shown in Figure 1). This induces a commuting correspondence between encoders; once Q is learned, embeddings from either modality can be mapped into the other model's coordinate system. The resulting aligned space also preserves downstream behavior, with mapped source embeddings reproducing the target model's nearest-neighbor classification and retrieval performance, indicating that Q preserves semantic and task-relevant geometry. The map is also data-efficient as it can be learned from a small fraction of the data, transfers across datasets without re-fitting, and composes consistently across model triplets. Together, these properties suggest a global correspondence between models.

Theoretically, we show that the multimodal contrastive objective constrains the *multimodal kernel* , i.e., the matrix of image-text similarities. Under mild data curation assumptions, this kernel is aligned across models at the population level, and we empirically find that such kernel agreement holds closely in practice. We further show that this population condition can be relaxed to a *finite* set of anchors, where agreement of the multimodal kernel forces the joint image and text embeddings to align through a shared orthogonal map. Finally, we extend the result for approximate kernel agreement.To summarize, the key contributions of our work are:

- We show that independently trained multimodal contrastive models are related by a *single orthogonal* map. Further, this map is shared across modalities, i.e., estimating the map from images alone aligns text, and vice versa.
- Theoretically, we prove that shared orthogonal alignment follows from a mild finite-anchor condition, and establish stability guarantees under approximate kernel agreement.
- We validate these claims across eight benchmarks and model pairs, with extensive ablations showing that this map transfers across datasets without re-fitting and remains consistent under composition, yielding the most reliable cross-model, cross-modal transfer.

2 Related Work

For an extended discussion of related work, see Section A.

Representational Convergence and Functional Interoperability. A central question in deep learning is whether independently trained models converge to identical representations. The Platonic

[1] Equivalently, Q can be learned from either modality: estimating it from paired text aligns images as well.

Representation Hypothesis (Huh et al., 2024) argues that large models across different modalities are converging towards the same representations Standard embedding similarity tools such as CKA (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), and shape metrics (Williams et al., 2021) measure this convergence up to broad equivalence classes (e.g., invertible linear maps), but do not provide an explicit mapping. A stronger operational test is *model stitching* and anchor-based latent translation, which connect representations via simple learnable maps (Lenc & Vedaldi, 2015; Bansal et al., 2021; Merullo et al., 2022; Moschella et al., 2023; Maiorca et al., 2023; Lähner & Moeller, 2024). However, these approaches are limited to coarse transfer metrics (e.g., image captioning), whereas we measure a stronger notion of alignment—tight pointwise agreement via cosine similarity—while simultaneously verifying that semantic structure is preserved through retrieval performance. Prior alignment results largely operate on *unimodal* marginals, including word embeddings (Mikolov et al., 2013; Dev et al., 2021; Conneau et al., 2018) and vision features (Maystre et al., 2025; Merullo et al., 2022), or move information across models without identifying the original joint geometry (Achara et al., 2026; Norelli et al., 2023; Roth et al., 2024). Aligning marginals, however, does not in general identify the *joint* distribution. We therefore study a strictly stronger question, whether the *joint* image-text geometry of two multimodal models is identifiable up to a *single, rigid* isometry.

The Modality Gap and Geometry of Contrastive Representations. Our study is grounded in the geometry of contrastive multimodal models (Radford et al., 2021; Jia et al., 2021). Empirically, these models exhibit a *modality gap*, where image and text embeddings cluster in distinct cones (Liang et al., 2022; Udandarao, 2022; Shi et al., 2023). This separation renders alignment non-trivial, as naive mappings can easily collapse the gap or distort the intra-modal structure. Theoretically, prior work has analyzed conditions under which contrastive objectives identify the underlying latent factors up to linear or affine transformations (Zimmermann et al., 2021; Roeder et al., 2021). While general linear maps include rotations, they also permit shear and anisotropic scaling, which are poorly constrained and undesirable for preserving semantic structure. In contrast, our setting restricts attention to *isometries*, which preserve angles, norms, and neighborhood relations. Even though images and text remain separated within each model, we prove that multimodal kernel agreement on a *small* anchor set suffices to recover a *shared* isometry Q that aligns the two models across *both* modalities. As a result, instead of retraining a model or learning complex transfer operators, one can anchor a new model to a reference model using a small fraction of the data and obtain transfer to another modality *for free*.

3 Problem Formulation

We consider the standard dual-encoder framework where data consists of co-occurring pairs (x, y) (e.g., images and text). A contrastive model consists of two encoders, $f : \mathcal{X} \to \mathbb{S}^{d-1}$ and $g : \mathcal{Y} \to \mathbb{S}^{d-1}$, which map inputs to the unit hypersphere in $\mathbb{R}^d$. The training objective maximizes the cosine similarity $\langle f(x), g(y) \rangle$ for semantically matched pairs while minimizing it for mismatched ones.

Consider two such models, trained in complete isolation on different datasets, with different architectures, initializations, and modeling choices: a *source* model $\mathcal{M} = (f, g)$ and a *target* model $\tilde{\mathcal{M}} = (\tilde{f}, \tilde{g})$ mapping to dimensions d and $\tilde{d}$ respectively. Due to optimization stochasticity and training differences, the embedding spaces of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are a priori incomparable. Even with identical data, the contrastive objective is invariant to a joint orthogonal transform of image and text embeddings, since it depends only on within-model dot products $\langle f(x), g(y) \rangle$. Differences in data, architecture, and optimization further amplify this ambiguity (Saunshi et al., 2022; Robinson et al., 2021). Cross-model dot products or nearest-neighbor queries are therefore ill-defined without first aligning the spaces. We seek a map $\psi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ that transports embeddings from $\mathcal{M}$ into $\tilde{\mathcal{M}}$. The key question is whether a consistent geometric relationship exists between the two models.

4 Modality Gap in Contrastive Representations

This question is nontrivial due to the intrinsic geometry of contrastive representations. If matched image-text pairs collapsed to approximately the same point on the hypersphere (i.e., $f(x) \approx g(y)$), then aligning the image manifold would be equivalent to aligning the text manifold. In practice, however, contrastive models exhibit a pronounced *modality gap* where image and text embeddings occupy largely disjoint regions of the sphere (Liang et al., 2022; Shi et al., 2023; Udandarao, 2022).

Prior work also suggests that naïvely "closing" this gap can harm downstream performance and fairness (Liang et al., 2022), as shown in Figure 2(a).

Since the gap reflects systematic modality-specific structure rather than noise (Schrodi et al., 2024), it is *not* obvious that a map ψ learned solely on images will extend correctly to text. In principle, infinitely many maps can agree on the image manifold yet behave arbitrarily on the text manifold. We argue that the alignment problem is indeed solvable because *relative* geometry is remarkably stable. As shown in Figure 2(b) and (c), the angular arrangement of the texts with respect to the images remains consistent. Mathematically, this means that the *multimodal kernels* are approximately preserved across models: $\langle f, g\rangle \approx \langle \tilde{f}, \tilde{g}\rangle$. This observation can be viewed as a multimodal analogue of the Platonic Representation Hypothesis (Huh et al., 2024) that posits that models converge to similar *unimodal* kernels ($\langle f, f\rangle \approx \langle \tilde{f}, \tilde{f}\rangle$). In the next section, we prove that this preservation of multimodal kernels is a sufficient condition to constrain the functional form of ψ, forcing it to be an isometry.

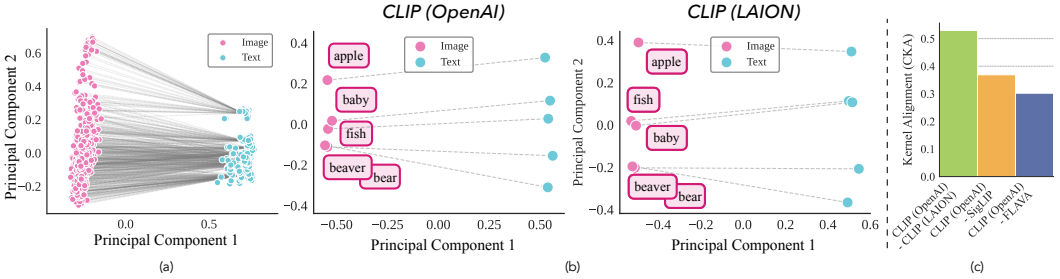

Figure 2: (a) PCA visualization of embeddings from CLIP (OpenAI) on CIFAR 100, showing a large systematic modality gap; (b) Across CLIP variants, the multimodal kernel $\langle f(x), g(y)\rangle$ (relative angles between image and text embeddings) is strongly preserved (dashed lines), unlike the unimodal kernel $\langle f(x), \tilde{f}(x')\rangle$; (c) CKA on multimodal kernels shows high alignment across models.

5 Theoretical Perspectives

In this section, we formalize the above intuition, showing that contrastive models, even when trained on different distributions, can recover the same multimodal kernels up to a constant (Section 5.2, as shown in Figure 2(b,c). We then ask how much of this population structure is actually needed. Remarkably, agreement along finitely many anchor-induced kernel sections identifies a linear map (Section 5.3), which becomes an isometry under the hypersphere constraint (Section 5.4). We further show that purely finite anchors approximate these sections when it sufficiently covers the representation domains. Finally, we extend these guarantees to the approximate setting (Section C.6). For a schematic overview of our theoretical analysis, refer to Figure 8 in the appendix.

5.1 Contrastive Representation Learning

Building on the dual-encoder framework defined in Section 3, we assume access to data characterized by co-occurring pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (e.g., images and text) drawn from a joint distribution p_{XY} with marginals p_X and p_Y. Here, the contrastive objective aims to learn functions, $f : \mathcal{X} \to \mathbb{S}^{d-1}$ and $g : \mathcal{Y} \to \mathbb{S}^{d-1}$, which map inputs to a shared unit hypersphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ where similarity of a pair (x, y) is measured by the score function $s(x, y) := \langle f(x), g(y)\rangle/\tau + b$, where $\tau > 0$ is a temperature parameter and $b \in \mathbb{R}$ is a scalar bias. This alignment is achieved by optimizing the symmetric InfoNCE objective (Oord et al., 2018) $\mathcal{L}_N$, which minimizes the cross-entropy of identifying the correct positive pair relative to the negatives.

5.2 Optimal Critic for Family of Contrastive Learners

We first characterize the score function $s(x, y)$ that minimizes the population InfoNCE loss. Let $r(x, y) := \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$ denote the pointwise density ratio, and $\log r(x, y)$ be the Pointwise Mutual Information (PMI). For a fixed training distribution P_{XY}, any global minimizer of the symmetric objective $\mathcal{L}_N$ induces an optimal score $s^*(x, y) = \log r(x, y) + C$, i.e. the pointwise mutual

information (PMI) up to a global constant. We note that this has been partially studied in prior work (Huh et al., 2024; Oord et al., 2018), with a detailed derivation provided in Section C.2 and Section C.3. Consequently, any two globally optimal models trained on the *same* distribution—or on distributions related to P_{XY} by a bijective reparameterization—induce kernels that differ only by a constant (Corollary C.6). Here, we go beyond the fixed-distribution setting and prove that this invariance persists even for models trained on *distinct* internet-scale corpora.

Platonic distribution and Dataset Curation. Let $P^\star_{XY}$ denote an underlying "reality" distribution with density $p^\star_{XY}$ and marginals $p^\star_X, p^\star_Y$ (satisfying positivity on the domain of interest). We model each training corpus as a *curation* of $P^\star_{XY}$ i.e. for dataset $a \in \{1, 2\}$, there exist measurable weights $u_a : \mathcal{X} \to (0, \infty)$ and $v_a : \mathcal{Y} \to (0, \infty)$ such that $p^{(a)}_{XY}(x, y) \propto u_a(x)\, v_a(y)\, p^\star_{XY}(x, y)$. Here u_a and v_a capture modality-specific curation (e.g., image quality and text language/safety etc.) that can substantially change the marginals while preserving the underlying semantic information. We further assume that curation acts *independently across modalities* in the sense that the expected text acceptance does not depend on the image it is paired with, and vice versa i.e. $\mathbb{E}^\star[v_a(Y) \mid X = x] = \mathbb{E}^\star[v_a(Y)]$ a.e. in x and $\mathbb{E}^\star[u_a(X) \mid Y = y] = \mathbb{E}^\star[u_a(X)]$ a.e. in y. where $\mathbb{E}^\star$ corresponds to expectation under $p^\star$ and a.e. stands for almost everywhere, causing these expectation terms to appear in the PMI only as additive constants.

Theorem 5.1. *If s^*_1 and s^*_2 are Bayes-optimal scores for the contrastive models trained on two distinct distributions defined above, then there exists a constant Δ' such that $s^*_1(x, y) = s^*_2(x, y) + \Delta'$ for $P^\star_X \otimes P^\star_Y$-a.e. (x, y).*

Theorem 5.1 establishes that even when trained on different distributions, contrastive learners can converge to optimal similarity scores that agree up to an additive constant. We note that our result holds for several widely used contrastive objectives such as softmax InfoNCE (Oord et al., 2018) and pairwise sigmoid objectives in SigLIP (Zhai et al., 2023). Since contrastive models approximate this target using dot products, any two models converging to the same score must implicitly align their kernels: $\langle f(x), g(y)\rangle \approx \langle \tilde{f}(x), \tilde{g}(y)\rangle$.[2] For detailed proofs and discussion, refer to Section C.4.

5.3 Linear Alignment of Contrastive Models

So far, we showed that independently trained contrastive models can induce multimodal kernels that agree up to an additive constant, despite differences in data and modeling choices. But recovering geometry *does not* require this population-level agreement. It turns out that agreement on a *finite* set of anchors, such as a downstream image-text dataset, is enough. Under this condition, our first main result shows that the alignment between the two embedding spaces is forced to be linear.

Let the contrastive model pairs (f, g) and $(\tilde{f}, \tilde{g})$ map inputs to $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and $\mathbb{S}^{\tilde{d}-1} \subset \mathbb{R}^{\tilde{d}}$ respectively, where $d \le \tilde{d}$, without loss of generality. We fix a set of *image anchors* $\{\bar{x}_j\}_{j=1}^d \subset \Omega_X$ and *text anchors* $\{\bar{y}_i\}_{i=1}^{\tilde{d}} \subset \Omega_Y$ and collect their embeddings into the following matrices: $G \coloneqq [g(\bar{y}_1)\ \cdots\ g(\bar{y}_{\tilde{d}})] \in \mathbb{R}^{d\times\tilde{d}}$, $\tilde{G} \coloneqq [\tilde{g}(\bar{y}_1)\ \cdots\ \tilde{g}(\bar{y}_{\tilde{d}})] \in \mathbb{R}^{\tilde{d}\times\tilde{d}}$, $F \coloneqq [f(\bar{x}_1)\ \cdots\ f(\bar{x}_d)] \in \mathbb{R}^{d\times d}$.

Assumption 5.2. The multimodal kernels coincide on the set of anchors:

$$\langle f(x), g(\bar{y}_i)\rangle = \langle \tilde{f}(x), \tilde{g}(\bar{y}_i)\rangle \ \ \forall x \in \Omega_X,\ \forall i \in \{1, \dots, \tilde{d}\},$$
$$\langle f(\bar{x}_j), g(y)\rangle = \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle \ \ \forall y \in \Omega_Y,\ \forall j \in \{1, \dots, d\}.$$

Theorem 5.3. *(Linear Identifiability, proof in Section C.5.1). Suppose $\tilde{G}$ and F are invertible. Then there exists a linear map A such that $\tilde{f}(x) = Af(x)\ \forall x \in \Omega_X$. Further, if A has full column rank, then for every $y \in \Omega_Y$, $\mathrm{Proj}_{\mathrm{Im}(A)}\, \tilde{g}(y) = A(A^\top A)^{-1} g(y)$. If $\tilde{d} = d$, then $\tilde{g}(y) = A^{-\top} g(y)$.*

5.4 Orthogonal Alignment of Contrastive Models

Theorem 5.3 shows that kernel matching identifies the representation up to a linear map A. But, contrastive encoders normalize embeddings to the unit hypersphere $\mathbb{S}^{d-1}$, forcing

[2] Exact equality implies the target PMI respects the contrastive parameterization constraints. Since CLIP scores are bounded and low-rank, they act as a low-rank approximation of the generally unbounded, high-rank population target.

$\|\tilde{f}(x)\|_2 = \|Af(x)\|_2 = 1$ everywhere. This forces A to be an isometry ($A^\top A = I_d$) only if the data is sufficiently diverse to probe the matrix in all directions. We formalize via the following condition.

Definition 5.4. *($\mathrm{Sym}(d)$-spanning)* A set of vectors $S \subset \mathbb{R}^d$ is $\mathrm{Sym}(d)$-*spanning* if the rank-one matrices $\{xx^\top : x \in S\}$ span the space of symmetric matrices $\mathrm{Sym}(d)$. Equivalently: if $M \in \mathrm{Sym}(d)$ and $x^\top Mx = 0$ for all $x \in S$, then $M = 0$.

Theorem 5.5. *(Orthogonal Identifiability, proof in Section C.5.2). Assume the conditions of Theorem 5.3 hold. If the set of image embeddings $\{f(x) : x \in \Omega_X\}$ contains a $\mathrm{Sym}(d)$-spanning subset, then the linear map A has orthonormal columns ($A^\top A = I_d$). Consequently, $\tilde{f}(x) = Qf(x) \quad \forall x \in \Omega_X$, where $Q := A$ satisfies $Q^\top Q = I_d$. Furthermore, for the other modality: $\mathrm{Proj}_{\mathrm{Im}(Q)}\, \tilde{g}(y) = Qg(y) \quad \forall y \in \Omega_Y$. If $\tilde{d} = d$, Q is orthogonal i.e. $Q \in O(d)$ and $\tilde{g}(y) = Qg(y)$.*

In Section C.5.3, we extend Theorem 5.5 to the case where $\mathrm{span}\{f(x) : x \in \Omega_X\}$ lies in low-dimensions. In Section C.6, we further relax Assumption 5.2 to an ϵ-approximate bound: $|\langle f(x), g(y)\rangle - \langle \tilde{f}(x), \tilde{g}(y)\rangle| \leq \epsilon$ and prove that the map Q becomes an approximate isometry. Finally, Proposition C.27 extends these guarantees to finite, sufficiently representative image and text anchor sets.

6 The Procrustes Algorithm

Guided by the theoretical guarantees in Section 5, we translate the alignment problem into an optimization procedure. We align the source and target manifolds of one modality (say images) using a set of N unlabelled anchor images $\{x_i\}_{i=1}^N$. Let $X \in \mathbb{R}^{d\times N}$ and $\tilde{X} \in \mathbb{R}^{\tilde{d}\times N}$ be the data matrices containing the centered, normalized embeddings $f(x_i)$ and $\tilde{f}(x_i)$ as columns. We solve for the optimal isometry $\hat{Q}$ by minimizing the transport cost subject to an orthogonality constraint: $\hat{Q} = \mathrm{argmin}_{Q^\top Q = I} \|\tilde{X} - QX\|_F^2$. This is the classic *Orthogonal Procrustes Problem*, which has a closed-form solution. For additional details and pseudocode, refer to Section D.4.1

7 Experimental Results

In this section, we empirically evaluate our central claim across various benchmarks and configurations, leading to three main takeaways: (i) a single orthogonal map $\mathcal{Q}$ accurately captures the relationship between independently trained contrastive models and, crucially, the same $\mathcal{Q}$ applies to image, audio and text representations (Section 7.2); (ii) $\mathcal{Q}$ is data-efficient and a small fraction of the data suffices to estimate it (Section 7.3); and (iii) although more expressive linear or non-linear maps can increase pointwise similarity on the fitted domain, they fail to transfer to the second modality and distort task-relevant geometry, degrading downstream retrieval; enforcing orthogonality in contrast yields the most reliable transfer across models and modalities (Section 7.4). Finally, we show that the learned map approximately commutes with cross-modal retrieval across models i.e. direct image alignment and text-mediated alignment recover the same semantic neighborhoods(Section 7.5).

7.1 Training Protocol

We report each metric both before alignment and after applying the learned orthogonal map $\mathcal{Q}$, and average results over three random seeds. We describe the evaluated model pairs, datasets, and metrics below and in detail in Section D.4.

Models. We evaluate five independently trained vision-language and audio-language pairs: (i) CLIP ViT-B/32 (`OpenAI`) and CLIP ViT-B/32 trained on `LAION-400M`; (ii) CLIP ViT-L/14 (`OpenAI`) and SigLIP; (iii) CLIP ViT-L/14 (`OpenAI`) and FLAVA (Radford et al., 2021; Schuhmann et al., 2021; Zhai et al., 2023; Singh et al., 2022) (iv) LAION-630K-best and LAION-630K-AudioSet-best and (v) LAION-CLAP (HTSAT+RoBERTa) and Microsoft-CLAP (HTSAT+GPT-2).We ℓ_2-normalize all embeddings such that dot products equal cosine similarity.

Datasets. We report results on Oxford-IIIT Pets (Parkhi et al., 2012), CIFAR-100 (Krizhevsky et al., 2009), Caltech-101 (Fei-Fei et al., 2004), STL10 (Coates et al., 2011), DTD (Cimpoi et al., 2014), ESC-50 (Piczak, 2015), Flickr30k (Plummer et al., 2015), and MSCOCO (Lin et al., 2014). For

more information about creating text prompts, refer to Section D.3.1. We report results only for Oxford Pets, ESC-50, and Flickr30k in the main paper and defer results on the rest to Section E.

Training. We learn the alignment across models using the standard orthogonal Procrustes solution described in Section 6. In practice, the two models can differ by a constant offset in embedding space due to finite-sample effects and dataset mismatch. We therefore fit and apply $\mathcal{Q}$ on centered embeddings, and then re-add the target mean i.e. $z \mapsto Q(z - \mu_A^{(\cdot)}) + \mu_B^{(\cdot)}$, where $\mu_A^{(\cdot)}$ and $\mu_B^{(\cdot)}$ are modality-specific means of the source and target embeddings (image or text, as appropriate). Centering isolates the rotational relationship by removing this offset while preserving the orthogonal correspondence in the centered space[3] Theoretically, this mean offset vanishes when the two models are exactly related by an orthogonal map, as discussed in Remark C.15.

Performance Metrics. We report five evaluation metrics: (1) paired-instance cosine similarity between aligned and target embeddings; (2) class-level image-image and text-text top-1 retrieval; and zero-shot classification or retrieval for (3) aligned images against native target text, (4) native target images against aligned text, and (5) aligned images against aligned text. All scores use cosine similarity. Full definitions are in Section D.4.2. To isolate the effect of Q, we further compare matched and mismatched pairs in Section E.5.

7.2 Multimodal Contrastive Models Differ by an Orthogonal Map

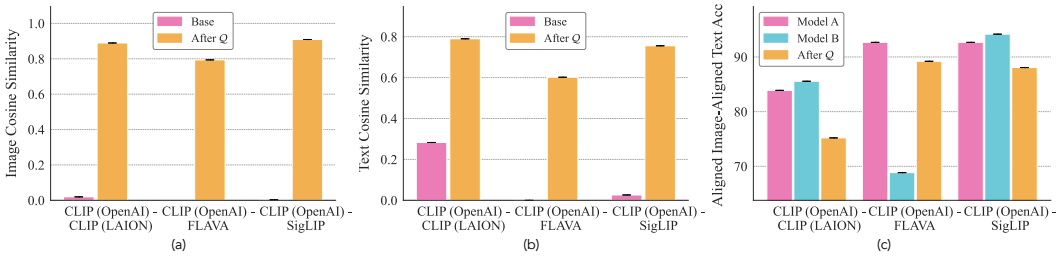

Figure 3: *Inter-model alignment on Oxford-Pets before and after fitting a single orthogonal map $\mathcal{Q}$ from paired images.* (a) Image-image cosine similarity; (b) Text-text cosine similarity; (c) Aligned Image and Aligned text class retrieval accuracy. Here, Model A and Model B denote each model's within-model image-to-text baseline. $\mathcal{Q}$ aligns images across models with a single orthogonal map, and the same $\mathcal{Q}$ learned only from image embeddings transfers to text, boosting text-text cosine from near-chance to near-oracle, all while preserving strong image classification accuracy.

An Orthogonal Map Aligns Different Models. First, from Figure 3(a), we find that a *single orthogonal map* almost perfectly aligns *image* embeddings across distinct multimodal contrastive models, improving the image-image cosine similarity from near zero to $\sim 0.8 - 0.9$. We observe analogous findings for *text* embeddings in Figure 3(b), indicating that independently trained contrastive models are related by an approximately orthogonal map. The same trend holds on the remaining image-text datasets (Section E.3). After alignment, matched pairs are well separated from mismatched pairs (Section E.5). These findings also extend to audio-text models (Figure 4; Section E.4).

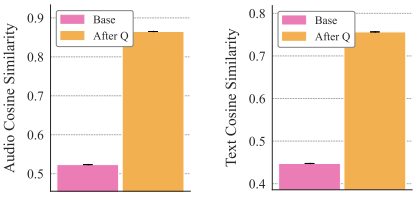

Figure 4: Orthogonal map also aligns independent audio-text contrastive models. Refer to Section E.4 for additional results.

This Map Transfers Across Modalities. Second, and more importantly, this map is *modality-invariant*. Figure 3(b) (and Figure 5) shows that the same orthogonal map $\mathcal{Q}$ fit using paired images (audio) sharply improves *text* alignment, significantly improving text-text cosine similarity across model pairs. Finally, as shown in Figure 3(c), aligned-

[3]We find that mean-centering has a negligible effect on class-level retrieval and decision geometry; it primarily changes pointwise cosine agreement (see Section E.14). Thus, a pure orthogonal map on raw embeddings suffices for semantic alignment and preserves decision geometry.

image-to-aligned-text classification remains high, showing that $\mathcal{Q}$ preserves task-relevant geometry while eliminating any need to compute the second model's text embeddings.

Additionally, in some cases, $\mathcal{Q}$ effectively transfers model A's stronger decision geometry into model B's space, matching or even exceeding model B's native performance. Aligned-to-native transfer, i.e. $\langle Qf, \tilde{g} \rangle$ and $\langle \tilde{f}, Qg \rangle$, is also reported across datasets and model pairs in Section E.3. Our findings extend beyond classification, yielding strong gains on fine-grained retrieval as shown in Figure 5 on Flickr30k and Table 4 on MS-COCO (Section E.6).

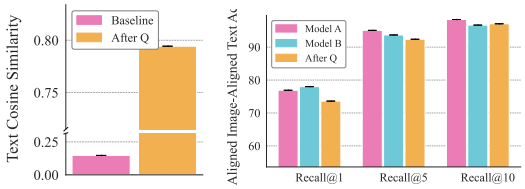

Figure 5: Orthogonal alignment also improves fine-grained retrieval. Refer to Section E.6 for additional results.

We extend these findings to mismatched embedding dimensions in Section E.12 and also show the reverse direction, fitting $\mathcal{Q}$ on text to align images, in Section E.10. Finally, Section E.14 ablates mean-centering and find that it mainly raises pointwise cosine and not class-level retrieval. Mean shift alone fails to restore cross-modal performance, while a joint image-text oracle is no better than fitting Q on images alone.

7.3 Only a Few Data Points Are Needed to Learn the Orthogonal Map

Theorem 5.5 states that if the multimodal kernels induced by two contrastive models agree on a sufficiently rich but *finite* set of d anchors, a single global orthogonal map aligns their representations across both modalities. As shown in Figure 6(a), we empirically validate this on Oxford-Pets, fitting $\mathcal{Q}$ using paired images from only N classes and evaluating transfer on the remaining unseen classes. Here, Model A and Model B denote each model's within-model image-to-text baseline. Performance on both seen and unseen classes improves quickly with just a few anchor classes (around 10-15 classes on Oxford Pets), after which additional anchors provide little benefit. Additional sample-complexity results are in Section E.7. Section E.8 shows that diverse anchors matter more than using more samples from the same class, and Section E.9 shows that a map fit on one dataset transfers to another without refitting.

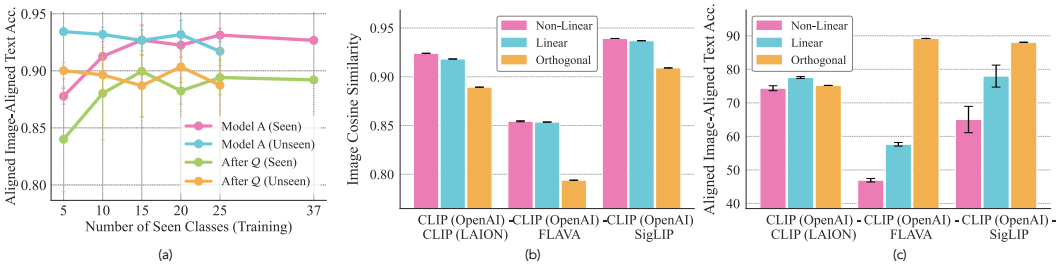

Figure 6: (a) Generalization of orthogonal alignment under limited supervision (Oxford Pets; CLIP (OpenAI) and FLAVA) as measured in aligned-image-to-aligned text accuracy. Comparison of alignment strategies on Oxford Pets across model pairs in terms of (b) image cosine similarity and (c) aligned-image-to-aligned-text accuracy. Unlike orthogonal maps, linear and non-linear maps distort task-relevant geometry and do not transfer to text, leading to poor downstream performance.

7.4 Evaluating Alternative Alignment Maps Than The Orthogonal Mapping

Here, we ablate the alignment design by comparing three maps of increasing expressiveness: (i) an orthogonal map Q, (ii) a linear map, and (iii) a non-linear MLP. As shown in Figure 6(b) and (c), more expressive maps improve *pointwise* image-image cosine similarity. However, these maps transfer poorly to the text modality and distort the image-text geometry. In contrast, the orthogonal map consistently performs best on both pointwise text cosine similarity and geometry-sensitive downstream metrics. Extended results for additional datasets, including the linear and MLP training setup, are provided in Section E.13.

Additional Ablations. We further test cycle and composition consistency in Section E.11. When holding the training data fixed and varying only the architecture, transfer is even stronger (Section E.15). We also show that a map fit on one dataset transfers to another without refitting (Section E.9), that fitting Q on text aligns images (Section E.10), and that Q preserves fine-grained attributes beyond class labels (Section E.17).

7.5 Qualitative Evidence of Commutativity of Image-Text Alignment Paths



Figure 7: *Qualitative k-NN retrieval under two alignment paths for Oxford Pets.* We compare two routes from a source image x into the target model's image space after alignment by $\mathcal{Q}$: a direct image-alignment path (top row) and a text-mediated path (bottom row). Both routes yield highly consistent neighborhoods, showing that the two alignment paths approximately close as a commuting diagram.

In this section, we test whether $\mathcal{Q}$ induces a consistent, modality-invariant geometry by comparing two retrieval routes from a source image x. *Direct* (Figure 7 (pink)): map the image embedding with $\mathcal{Q}$ and retrieve its top-5 nearest target images. *Text-mediated* (Figure 7 (blue)): retrieve the nearest source text for x, map it with $\mathcal{Q}$, retrieve the top-1 nearest target text, then retrieve the top-5 target images associated with that text.

Figure 7 (and Appendix Figure 58) shows that both routes recover essentially the same semantic neighborhood. In terms of the commuting diagram, transporting x by $\mathcal{Q}$ and then applying target-space retrieval agrees with first retrieving through the source image-to-text operator, transporting via $\mathcal{Q}$, and then retrieving back to images. This indicates that $\mathcal{Q}$ approximately commutes with the cross-modal nearest-neighbor operators on this domain.

8 Discussion and Conclusion

Conclusion. In this work, we show a rigid form of geometric convergence in multimodal contrastive models: across independently trained systems (with different data and modeling choices), a single orthogonal map learned in one modality can approximately align both image and text representations, inducing a shared coordinate system. Moreover, estimating this map requires only a small fraction of the data from a single modality (image or text). Theoretically, we characterize conditions under which agreement of the multimodal similarity kernel forces such a shared isometry.

Discussion and Implications. Our results have several practical and scientific implications. In large embedding systems, switching models typically triggers full re-embedding, often infeasible at modern scale (billions of vectors) (Jayaram Subramanya et al., 2019; Johnson et al., 2019) and costly in both time and compute (OpenAI, 2024). We show that a small fraction of the data can recover the orthogonal map that restores compatibility across models. Since it preserves inner products, it supports model upgrades without re-encoding while keeping the embedding geometry intact. Finally, models often specialize differently; one might have a stronger vision tower, while another has a stronger or multilingual text tower. Our approach lets practitioners swap and combine towers while preserving image-text geometry. For an extended discussion, refer to Section B.

Limitations and Future Work. While we already test fine-grained attribute transfer qualitatively in Section E.17, a natural next step is to measure this with trained decoders rather than nearest-neighbor edits. We study image-text and audio-text contrastive encoders; extending to other modalities (e.g., video) is an important direction.

AI USE STATEMENT

We used generative AI tools only to assist in polishing the language in this manuscript. All AI-assisted text was reviewed by the authors, who take responsibility for the final content.

REPRODUCIBILITY STATEMENT

We provide our code for reproducibility in the supplementary material. All datasets are publicly available, and for experiments, we detail the setup, training protocols, and algorithm implementations in Section D. We provide complete proofs and relevant background for all theoretical results in Section C.

REFERENCES

Akshit Achara, Tatiana Gaintseva, Mateo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey, Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06205.

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. URL https://arxiv.org/abs/1902.09229.

Kiril Bangachev, Guy Bresler, Iliyas Noman, and Yury Polyanskiy. Global minimizers of sigmoid contrastive loss. *arXiv preprint arXiv:2509.18552*, 2025.

Kiril Bangachev, Guy Bresler, and Yury Polyanskiy. Representation alignment rests on linear structure, 2026. URL https://arxiv.org/abs/2605.28870.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.07682.

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. Stationary representations: Optimally approximating compatibility and implications for improved model replacements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28793–28804, June 2024. doi: 10.1109/CVPR52733.2024.02720. URL https://openaccess.thecvf.com/content/CVPR2024/html/Biondi_Stationary_Representations_Optimally_Approximating_Compatibility_and_Implications_for_Improved_Model_CVPR_2024_paper.html.

Enric Boix-Adserà, Hannah Lawrence, George Stepaniants, and Philippe Rigollet. GULP: a prediction-based metric between representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7115–7127, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f0435cffef91068ced08d7c7d8e643e-Abstract-Conference.html.

Rhea Chowers, Oshri Naparstek, Udi Barzelay, and Yair Weiss. Is the modality gap a bug or a feature? a robustness perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30288–30298, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chowers_Is_the_Modality_Gap_a_Bug_or_a_Feature_A_CVPR_2026_paper.html.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, pp. 6243–6267. PMLR, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(28):795–828, 2012. URL https://jmlr.org/papers/v13/cortes12a.html.

Nello Cristianini, Andre Elisseeff, John Shawe-Taylor, and Jaz Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, 2002. URL https://papers.neurips.cc/paper/1946-on-kernel-target-alignment.pdf.

Sunipa Dev, Safia Hassan, and Jeff M Phillips. Closed form word embedding alignment. *Knowledge and Information Systems*, 63(3):565–588, 2021.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. *International Conference on Learning Representations (ICLR)*, 2023.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. *International Conference on Learning Representations (ICLR)*, 2026.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Jonas Herzog and Yue Wang. Reevaluating the intra-modal misalignment hypothesis in CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24310–24319, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Herzog_Reevaluating_the_Intra-Modal_Misalignment_Hypothesis_in_CLIP_CVPR_2026_paper.html.

(First name unknown) Hu et al. Learning backward compatible embeddings. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2022.

David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. URL https://arxiv.org/abs/2405.07987.

Young Kyun Jang and Ser-nam Lim. Towards cross-modal backward-compatible representation learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Jang_Towards_Cross-modal_Backward-compatible_Representation_Learning_for_Vision-Language_Models_ICCV_2025_paper.pdf.

Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in neural information processing Systems*, 32, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

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

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008/full.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Zorah Lähner and Michael Moeller. On the direct alignment of latent spaces. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pp. 158–169. PMLR, 2024. URL https://proceedings.mlr.press/v243/lahner24a.html.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Lenc_Understanding_Image_Representations_2015_CVPR_paper.html.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55394–55414, 2023. doi: 10.52202/075280-2418. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ad5fa03c906ca15905144ca3fbf2a768-Abstract-Conference.html.

Lucas Maystre et al. When embedding models meet: Procrustes bounds and first-order equivalences for embedding interoperability. *arXiv preprint arXiv:2510.13406*, 2025. URL https://arxiv.org/abs/2510.13406.

Qiang Meng, Chixiang Zhang, Xiaoqiang Xu, and Feng Zhou. Learning compatible embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9939–9948, 2021.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013. URL https://arxiv.org/abs/1309.4168.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25037–25060. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/moayeri23a.html.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1806.05759.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15303–15319, 2023. doi: 10.52202/075280-0673. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3186591903d9db31770ad131adb5ceb4-Abstract-Conference.html.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

OpenAI. Api pricing. https://openai.com/api/pricing/, 2024. Accessed: January 2026.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, 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. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.05806.

J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra. Can contrastive learning avoid shortcut solutions? In *Neural Information Processing Systems (NeurIPS)*, 2021.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Karsten Roth, Lukas Thede, Almut Sophia Koepke, Oriol Vinyals, Olivier Hénaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qIdmXmg2JJ.

N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning (ICML)*, 2022.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. *arXiv preprint arXiv:2404.07983*, 2024.

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

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2020.

Peiyang Shi, Michael C Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in clip. In *ICLR 2023 workshop on multimodal representation learning: perks and pitfalls*, 2023.

Amanpreet Singh et al. FLAVA: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2022. URL https://arxiv.org/abs/2112.04482.

Vishaal Udandarao. Understanding and fixing the modality gap in vision-language models. *Master's thesis, University of Cambridge*, 32, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, June 2022. doi: 10.1109/CVPR52688.2022.01759. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_LiT_Zero-Shot_Transfer_With_Locked-Image_Text_Tuning_CVPR_2022_paper.html.

Xiaohua Zhai et al. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. URL https://arxiv.org/abs/2303.15343.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zimmermann21a.html.

Liu Ziyin and Isaac Chuang. Proof of a perfect platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025.

APPENDIX

A Additional Related Works **19**

B Additional Discussion and Implications **20**

C Proof of Theoretical Results **21**

C.1 Setup and Notations 21

C.2 Bayes Posterior and Optimal Critic for Unimodal Contrastive Learners 22

C.3 Optimal Critic for Multimodal Contrastive Learners 23

C.4 Bayes-Optimal Scores Are Identifiable Up to a Constant 23

C.5 Guarantees Under Exact Alignment 25

C.5.1 Linear Alignment of Independent Contrastive Models 25

C.5.2 Isometric Alignment of Independent Contrastive Models 26

C.5.3 Isometric Alignment of Independent Contrastive Models When One Modality Lies in Low Dimension 26

C.5.4 Alignment Upto Classification Boundaries 27

C.6 Guarantees Under Approximate Alignment 28

C.6.1 Approximate Linear Alignment of Independent Contrastive Models 28

C.6.2 Approximate Isometric Alignment of Independent Contrastive Models . . . 29

C.6.3 Finite Anchor-Anchor Alignment 30

C.6.4 Approximate Isometric Alignment of Independent Contrastive Models When One Modality Lies in Low Dimension 30

D Supplementary Experimental Details and Assets Disclosure **31**

D.1 Assets 31

D.2 Hardware and setup 31

D.3 Datasets 31

D.3.1 Benchmarks 31

D.3.2 Constructing Text Templates 31

D.4 Training and Evaluation Protocol 33

D.4.1 Learning An Orthogonal Transformation 33

D.4.2 Performance Metrics at Evaluation 35

E Additional Experiments **35**

E.1 Modality Gap in Contrastive Models 35

E.2 Visualization of Multimodal Kernels 36

E.3 Independently Trained Image-Text Contrastive Models Differ by an Orthogonal Map That Is Shared Across Modalities 36

E.4 Extension to Audio Modality 42

E.5 Matched and Mismatched Cross-Model Similarity 42

E.6 Extension to Fine-Grained Retrieval 43
E.7 Only a Few Data Points Are Needed to Learn the Orthogonal Map 44
E.8 Sensitivity To The Choice And Diversity Of Anchors 51
E.9 The Learned Orthogonal Map Generalizes Broadly 51
E.10 Learning the Orthogonal Map from Text Instead of Images Transfers to Images . . 56
E.11 Cycle Consistency and Consistency Under Composition 59
E.12 Alignment Across Embedding Dimensions. 66
E.13 Evaluating Alternative Alignment Maps Than The Orthogonal Mapping 69
E.14 Orthogonal Alignment With and Without Centering 72
E.15 Isolating Architecture Effects Under Fixed Training Data 76
E.16 Qualitative Evidence of Commutativity of Image-Text Alignment Paths 78
E.17 Fine-grained Semantic Preservation Under Orthogonal Maps. 78

A Additional Related Works

Representational Convergence. A long-standing theme in representation learning is whether independently trained networks converge to comparable internal representations. This question is frequently studied through representational similarity analyses that compare activations up to broad equivalence classes, including RSA (Kriegeskorte et al., 2008), CCA-based methods such as SVCCA (Raghu et al., 2017) and its refinements (Morcos et al., 2018), kernel-based comparisons such as CKA (Kornblith et al., 2019), and shape metrics defined modulo chosen transformation groups (Williams et al., 2021). Earlier work also probed convergence by explicitly matching units or subspaces across independently trained networks (e.g., via neuron matching and sparse prediction), highlighting that models may learn similar *subspaces* even when individual coordinates do not align. These tools have been influential for documenting empirical convergence trends and motivating hypotheses such as the Platonic Representation Hypothesis (PRH) (Huh et al., 2024). PRH formalizes convergence at the level of induced similarity structure: two representations are considered aligned when their (co-occurrence) kernels agree on corresponding inputs. Recent evaluations, however, question how robust this claim remains after calibration and at realistic multimodal scale (Gröger et al., 2026; Koepke et al., 2026). By construction, many similarity measures are invariant to broad classes of transformations (e.g., invertible linear maps) or compare induced kernels rather than producing an explicit coordinate-level correspondence. Relatedly, GULP defines a prediction-based distance that is invariant to orthogonal transformations (Boix-Adserà et al., 2022), while recent work shows that centering can remove a dominant first-order source of cross-model misalignment (Bangachev et al., 2026). These works characterize similarity or improve calibration, but do not identify a map that must be shared across modalities. Our work targets this stronger notion of convergence: whether independently trained *multimodal embedding spaces* agree in (or can be brought into) a shared coordinate system via a simple global map.

Model Stitching and Functional Interoperability. Beyond similarity indices, several works test whether independently trained representations become *interchangeable* under simple transformations such as linear or orthogonal maps. Relative representations use similarities to fixed anchors to obtain invariance across latent spaces and enable zero-shot communication and stitching (Moschella et al., 2023). Anchor-based linear and orthogonal maps are also well established for word translation (Mikolov et al., 2013; Conneau et al., 2018) and for translation between pretrained latent spaces (Maiorca et al., 2023; Lähner & Moeller, 2024; Moayeri et al., 2023; Merullo et al., 2022; Maystre et al., 2025). In particular, Maiorca et al. (2023) translate across architectures and domains, but fit a separate map for each observed source-target pair and evaluate heterogeneous transfer mainly by downstream classification rather than pointwise embedding agreement. Lähner & Moeller (2024) directly compare rigid and unconstrained linear maps and argue that a strict isometry can be insufficient. Multi-way alignment extends this setting to more than two models through a shared reference space and finds that orthogonal alignment best preserves geometry, while CCA-style methods can be stronger for retrieval (Achara et al., 2026). Thus, neither Procrustes alignment, anchor-based estimation, nor zero-shot stitching is itself new.

Our contribution is the *cross-modal coupling*, where a map fit on one modality aligns a second modality for which no anchors are observed, together with conditions under which one shared isometry must exist. This differs from fitting a separate map for each observed source-target pair (Maiorca et al., 2023), constructing a new anchor-relative multimodal space from paired image-text data as in ASIF (Norelli et al., 2023), or transferring complementary knowledge through distillation (Roth et al., 2024). These methods address related forms of interoperability but do not identify a single map between two complete multimodal embedding spaces that transfers to an unseen modality.

Symmetries, Global Minima, and Landscape Connectivity. The existence of many apparently distinct solutions is also consistent with known symmetries of neural networks and the geometry of the loss landscape. Empirically, SGD solutions are often connected by low-loss paths (mode connectivity) (Garipov et al., 2018; Draxler et al., 2018), and accounting for permutation symmetries of hidden units can further reduce apparent barriers between independently trained models (Entezari et al., 2021). Recent work makes this operational by explicitly *rebasing* one model to another via permutation alignment to enable weight-space merging (Ainsworth et al., 2022). In this context, our results can be viewed as a representation-space analogue: while weights admit large symmetry groups, we find that multimodal embedding spaces often differ primarily by an (approximately) orthogonal transform that is shared across modalities.

Vision-Language Contrastive Pretraining and the Modality Gap. Large-scale vision-language representation learning is dominated by dual-encoder contrastive objectives, as exemplified by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), with many variants exploring alternative losses, scaling strategies, and training recipes (e.g., Zhai et al., 2023; Singh et al., 2022). LiT takes a related training-time approach by freezing the image encoder while learning a compatible text encoder (Zhai et al., 2022). A recurring geometric observation in such models is the *modality gap*: image and text embeddings can occupy systematically shifted regions even after normalization, with consequences for optimization and downstream behavior (Liang et al., 2022; Udandarao, 2022; Shi et al., 2023; Chowers et al., 2026). Recent work further shows that exact cross-modal similarities can determine the corresponding unimodal Gram matrices under suitable rank and normalization conditions (Herzog & Wang, 2026). Our work asks a different inter-model question: when cross-modal kernels agree across independently trained models, does this structure identify one coordinate map shared by both modalities?

Theory of Contrastive Learning and Kernel Alignment. Theoretical analyses of contrastive objectives have clarified what is identifiable from paired data and how representation geometry emerges. Key work includes contrastive predictive coding and InfoNCE-style objectives (Oord et al., 2018), general theoretical frameworks and guarantees for contrastive representation learning (Arora et al., 2019; HaoChen et al., 2021; Bangachev et al., 2025), and geometric decompositions into alignment and uniformity (Wang & Isola, 2020). Gupta et al. (2023) show that minimizing a loss which preserves unimodal kernels across augmentations induces an orthogonally equivariant structure in the contrastive embedding space. Other analyses connect contrastive learning to inversion of latent generative structure under suitable assumptions (Zimmermann et al., 2021). More generally, Roeder et al. (2021) establish representation identifiability up to an invertible linear transformation. In contrast, our result shows when multimodal contrastive structure reduces this ambiguity to one shared orthogonal map.

Separately, *kernel alignment* formalizes agreement between similarity functions and targets (Cristianini et al., 2002; Cortes et al., 2012), and underlies modern representational comparisons such as CKA (Kornblith et al., 2019). PRH also emphasizes kernel-level convergence, and recent theoretical work proves “perfect" PRH in simplified (deep linear) settings where representations align up to an orthogonal transformation (Ziyin & Chuang, 2025). Our theoretical results operate at this interface: we provide minimal conditions under which agreement of cross-modal similarity structure on a finite anchor set forces the existence of a shared orthogonal transform coupling modalities across models, and we establish stability guarantees translating approximate kernel agreement into reliable retrieval transfer.

B Additional Discussion and Implications

Our results have several practical and scientific implications. First, they point to a simple mechanism for backward-compatible upgrades in large embedding systems. Upgrading embedding models typically forces full re-embedding and ANN index rebuilds, which are prohibitively costly at modern scale (hundreds of millions to billions of vectors) (Jayaram Subramanya et al., 2019; Johnson et al., 2019) and often require multi-hour rebuilds and six-figure re-embedding budgets (OpenAI, 2024). This has motivated prior work on backward-compatible and interoperable embeddings, where new models are explicitly trained to preserve compatibility with deployed representations or learn stationary compatible representations (Jang & Lim, 2025; Meng et al., 2021; Hu et al., 2022; Shen et al., 2020; Biondi et al., 2024). Our findings show that a small fraction of the data can often restore cross-model comparability for multimodal contrastive systems. Because orthogonal maps preserve inner products, this can enable upgrades without re-encoding stored corpora and often without rebuilding indexes.

Second, a shared coordinate system enables mix-and-match multimodal pipelines. Different models often excel in different components, such as stronger vision encoders or stronger and more multilingual text encoders. When representations are aligned by a single orthogonal map, image and text towers from different models can be combined into a common space, enabling retrieval across heterogeneous encoders. This perspective complements recent work on representation compatibility and model stitching, which studies when independently trained networks can be connected via lightweight alignment layers (Bansal et al., 2021).

Finally, the ability to align text representations without accessing text has implications for data governance and security. In many deployments, raw text may be unavailable due to privacy, licensing, or retention constraints, even though embeddings are stored. Our results show that, in multimodal systems, text-space alignment can be recovered without using text, given a small fraction of the data from another modality. At the same time, easy cross-model transferability raises security considerations: if embeddings across models and modalities are easily transformable, then stored embeddings may encode more transferable semantic information than anticipated, reinforcing the need to treat embeddings as sensitive artifacts rather than model-specific byproducts.

C Proof of Theoretical Results

For a schematic overview of our theoretical analysis, refer to Figure 8.

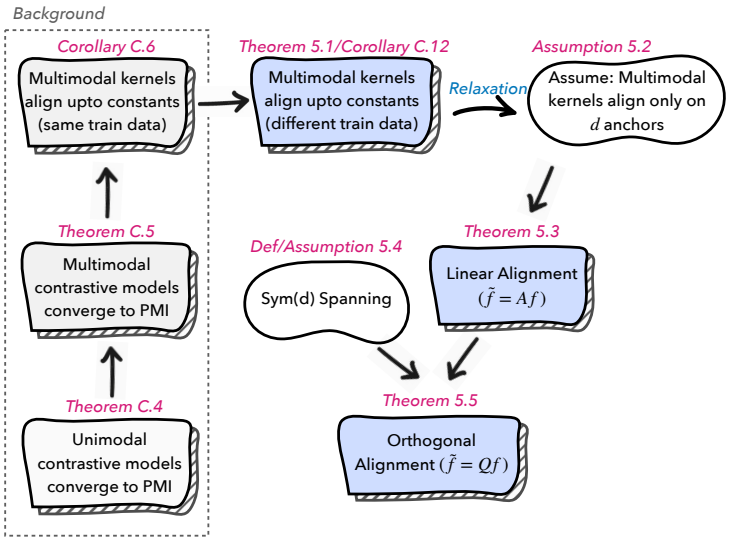

Figure 8: High-level overview of our theoretical results. Background results establish multimodal kernel alignment across models at a population level; We show that this full population agreement is stronger than necessary: kernel agreement on a finite anchor set already forces linear alignment, and under the hypersphere constraint, this alignment becomes orthogonal.

C.1 Setup and Notations

Let (X, Y) be random variables on measurable spaces $\mathcal{X}, \mathcal{Y}$ with joint density p_{XY} and marginals p_X, p_Y.

Assumption C.1 (Positivity on the domain)**.** On the domain of interest, assume $p_{XY}(x, y) > 0$ and $p_X(x)p_Y(y) > 0$.

Remark C.2. The above assumptions can be expressed without densities by replacing ratios of densities with Radon-Nikodym derivatives. Specifically, on the domain of interest, assume $P_{XY} \ll P_X \otimes P_Y$ (equivalently, $P_{Y|X=x} \ll P_Y$ for P_X-a.e. x), and let $r(x, y) := \frac{dP_{XY}}{d(P_X \otimes P_Y)}(x, y)$ denote the density ratio. Optionally, assume $r(x, y) > 0$ a.e. if $\log r$ is required to be finite. For clarity, we'll use densities.

Definition C.3 (Density ratio and Pointwise Mutual Information)**.** Define the mutual density ratio $r : \mathcal{X} \times \mathcal{Y} \to (0, \infty)$ by

$$r(x, y) := \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}, \qquad K^*(x, y) := \log r(x, y).$$

A contrastive model is a pair of measurable maps $f : \mathcal{X} \to \mathbb{S}^{d-1}, g : \mathcal{Y} \to \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$, and a temperature $\tau > 0$, defining a score

$$s(x, y) := \frac{1}{\tau} \langle f(x), g(y) \rangle + b, \ b \in \mathbb{R}.$$

In practice, contrastive learning constructs a finite candidate set by pairing each query with one positive and several independent negatives; the InfoNCE loss is exactly the cross-entropy for identifying the positive within this list. Below, we formalize the standard finite-sample sampling procedure underlying InfoNCE (Oord et al., 2018) principle.

InfoNCE Sampling Model. Fix $N \geq 2$. Sample $X \sim P_X$, draw one *positive* $Y^+ \sim P_{Y|X}(\cdot \mid X)$, and draw $N - 1$ *negatives* $Y^{(1)}, \ldots, Y^{(N-1)} \overset{\text{iid}}{\sim} P_Y$ independently of (X, Y^+). Sample $J \sim \text{Unif}\{0, \ldots, N-1\}$ and place Y^+ in slot J to form the candidate list $(Y_0, \ldots, Y_{N-1})$. The learner observes $(X, Y_{0:N-1})$ but not J.

Given a measurable score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, define

$$q_s(j \mid x, y_{0:N-1}) := \frac{\exp(s(x, y_j))}{\sum_{k=0}^{N-1} \exp(s(x, y_k))}, \qquad \mathcal{L}_N^{(X)}(s) := \mathbb{E}\big[-\log q_s(J \mid X, Y_{0:N-1}) \big].$$

Define $\mathcal{L}_N^{(Y)}$ analogously by swapping the roles of X and Y, and set

$$\mathcal{L}_N(s) := \mathcal{L}_N^{(X)}(s) + \mathcal{L}_N^{(Y)}(s).$$

This gives us the objective of a multimodal contrastive learner. We now characterize the minimizers of $\mathcal{L}_N$.

C.2 Bayes Posterior and Optimal Critic for Unimodal Contrastive Learners

Theorem C.4 (Bayes posterior and characterization of minimizers)**.** *Under Assumption C.1,*

1. *For almost every $(x, y_{0:N-1})$ under the InfoNCE sampling model,*

$$q^*(j \mid x, y_{0:N-1}) := \Pr(J = j \mid X = x, Y_{0:N-1} = y_{0:N-1}) = \frac{r(x, y_j)}{\sum_{k=0}^{N-1} r(x, y_k)}.$$

2. *Any measurable s^* achieving the infimum of $\mathcal{L}_N^{(X)}$ must satisfy*

$$s^*(x, y) = \log r(x, y) + c(x)$$

for some measurable $c : \mathcal{X} \to \mathbb{R}$, holding for almost every (x, y) with respect to $P_X \otimes P_Y$.

Proof. (i) Conditioning on $(J = j, X = x)$ gives

$$p(y_{0:N-1} \mid J = j, X = x) = p_{Y|X}(y_j \mid x) \prod_{i \neq j} p_Y(y_i).$$

Since J is uniform, Bayes' rule yields

$$\Pr(J = j \mid x, y_{0:N-1}) = \frac{p_{Y|X}(y_j \mid x) \prod_{i \neq j} p_Y(y_i)}{\sum_{k=0}^{N-1} p_{Y|X}(y_k \mid x) \prod_{i \neq k} p_Y(y_i)}.$$

Dividing numerator and denominator by $\prod_{i=0}^{N-1} p_Y(y_i)$ gives (i), since $\frac{p_{Y|X}(y|x)}{p_Y(y)} = r(x, y)$.

(ii) For each realization $(x, y_{0:N-1})$, the conditional risk is

$$\mathbb{E}\big[-\log q_s(J \mid x, y_{0:N-1}) \mid x, y_{0:N-1} \big] = H(q^*(\cdot \mid x, y_{0:N-1}),\ q_s(\cdot \mid x, y_{0:N-1})),$$

so $\mathcal{L}_N^{(X)}(s)$ is minimized iff $q_s(\cdot \mid x, y_{0:N-1}) = q^*(\cdot \mid x, y_{0:N-1})$ a.e. On this full-measure set, for any $j \neq k$,

$$\exp\big(s^*(x, y_j) - s^*(x, y_k)\big) = \frac{q_{s^*}(j \mid x, y_{0:N-1})}{q_{s^*}(k \mid x, y_{0:N-1})} = \frac{q^*(j \mid x, y_{0:N-1})}{q^*(k \mid x, y_{0:N-1})} = \frac{r(x, y_j)}{r(x, y_k)}.$$

Integrating out $Y_2, \ldots, Y_{N-1}$ (Fubini–Tonelli), gives $\exp(s^*(x, y) - s^*(x, y')) = r(x, y)/r(x, y')$ for $(P_X \otimes P_Y \otimes P_Y)$-a.e. (x, y, y'). Thus $s^*(x, y) - \log r(x, y)$ is (for P_X-a.e. x) independent of y on a P_Y-full-measure set, i.e. $s^*(x, y) = \log r(x, y) + c(x)$ for some measurable c, holding $(P_X \otimes P_Y)$-a.e. □

C.3 Optimal Critic for Multimodal Contrastive Learners

Theorem C.5. *Under the Assumption C.1, any measurable s^* achieving the infimum of $\mathcal{L}_N$ must satisfy*

$$s^*(x, y) = \log r(x, y) + C$$

for some constant $C \in \mathbb{R}$, holding for almost every (x, y) with respect to $P_X \otimes P_Y$.

Proof. Let $\ell_X := \inf_s \mathcal{L}_N^{(X)}(s)$ and $\ell_Y := \inf_s \mathcal{L}_N^{(Y)}(s)$. By Theorem C.4(ii), the score $s_0(x, y) := \log r(x, y)$ simultaneously achieves ℓ_X and ℓ_Y (up to an additive constant), hence $\inf_s \mathcal{L}_N(s) = \ell_X + \ell_Y$.

If s^* achieves the infimum of $\mathcal{L}_N$, then

$$\mathcal{L}_N^{(X)}(s^*) + \mathcal{L}_N^{(Y)}(s^*) = \ell_X + \ell_Y.$$

Since each term is bounded below by its own infimum, we must have $\mathcal{L}_N^{(X)}(s^*) = \ell_X$ and $\mathcal{L}_N^{(Y)}(s^*) = \ell_Y$. Applying Theorem C.4(ii) in each direction yields

$$s^*(x, y) = \log r(x, y) + c(x) = \log r(x, y) + d(y)$$

$(P_X \otimes P_Y)$-a.e. for measurable c and d. Thus $c(x) = d(y)$ on a full $(P_X \otimes P_Y)$-measure set Let $E \subseteq \mathcal{X} \times \mathcal{Y}$ be a full $(P_X \otimes P_Y)$-measure set where $c(x) = d(y)$. By Fubini-Tonelli's theorem, there exists y_0 such that the section $E_{y_0} := \{x : (x, y_0) \in E\}$ has $P_X(E_{y_0}) = 1$. Then for $x \in E_{y_0}$, $c(x) = d(y_0)$, so c is P_X-a.e. constant. Call this constant C. Plugging back gives $s^*(x, y) = \log r(x, y) + C$ for $(P_X \otimes P_Y)$-a.e. (x, y). □

Corollary C.6. *Let (f, g) and $(\tilde{f}, \tilde{g})$ be two contrastive models trained on the same distribution that achieve the global infimum of the symmetric objective $\mathcal{L}_N$. Then there exists a constant $\Delta \in \mathbb{R}$ such that*

$$\langle f(x), g(y) \rangle = \langle \tilde{f}(x), \tilde{g}(y) \rangle + \Delta$$

for almost every (x, y) with respect to $P_X \otimes P_Y$. If the temperatures differ, the relation becomes affine in $\langle \tilde{f}(x), \tilde{g}(y) \rangle$.

Remark C.7. On a finite domain $\{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_m\}$, the CLIP score matrix $S \in \mathbb{R}^{n \times m}$ with entries $S_{ij} = \tau^{-1} \langle f(x_i), g(y_j) \rangle + b$ can be written as $S = \tau^{-1} F^\top G + b\, \mathbf{1}_n \mathbf{1}_m^\top$, where $F = [f(x_1) \ \cdots \ f(x_n)]$ and $G = [g(y_1) \ \cdots \ g(y_m)]$. Hence $\operatorname{rank}(S) \leq d + 1$. Therefore, exact equality $S_{ij} = K^*(x_i, y_j) + C$ would require a strong low-rank structure of the PMI matrix on that domain. This is *not* assumed in Theorem C.4 and Theorem C.5 as it becomes relevant only when one tries to realize the Bayes-optimal critic within a dot-product parameterization.

C.4 Bayes-Optimal Scores Are Identifiable Up to a Constant

So far, Theorem C.5 and Corollary C.6 show that *for a fixed training distribution P_{XY}*, any two global minimizers of the symmetric InfoNCE objective induce the same kernel (up to a global additive constant). We now give a simple, formal setting in which this conclusion can *also* hold for two *different* training distributions.

Dataset Curation. Let $P_{XY}^\star$ be a ground truth distribution on $\mathcal{X} \times \mathcal{Y}$ with density $p_{XY}^\star$ and marginals $p_X^\star, p_Y^\star$, satisfying positivity on the domain of interest (cf. Assumption C.1). For each dataset $a \in \{1, 2\}$, let $P_{XY}^{(a)}$ be a training distribution with density $p_{XY}^{(a)}$ and marginals $p_X^{(a)}, p_Y^{(a)}$. Let $r^\star, K^\star$ and $r^{(a)}, K^{(a)}$ denote the density ratios and PMIs defined earlier, applied to $p_{XY}^\star$ and $p_{XY}^{(a)}$, respectively.

Assumption C.8. For each dataset $a \in \{1, 2\}$, there exist measurable weights $u_a : \mathcal{X} \to (0, \infty)$ and $v_a : \mathcal{Y} \to (0, \infty)$ such that

$$Z_a := \mathbb{E}^\star[u_a(X)v_a(Y)] < \infty, \qquad p^{(a)}_{XY}(x, y) = \frac{u_a(x)v_a(y)}{Z_a}\, p^\star_{XY}(x, y).$$

In Assumption C.8, we view large-scale training corpora as distinct *curations* of a common underlying distribution, where each modality (image/text) is filtered by its own criteria (e.g., quality, safety, language) independently of the other.

Lemma C.9. *Under Assumption C.8, for each $a \in \{1, 2\}$ and for $P^\star_X \otimes P^\star_Y$-a.e. (x, y),*

$$r^{(a)}(x, y) = r^\star(x, y) \cdot \frac{Z_a}{\mathbb{E}^\star[v_a(Y) \mid X = x]\, \mathbb{E}^\star[u_a(X) \mid Y = y]}.$$

Equivalently,

$$K^{(a)}(x, y) = K^\star(x, y) + \log Z_a - \log \mathbb{E}^\star[v_a(Y) \mid X = x] - \log \mathbb{E}^\star[u_a(X) \mid Y = y].$$

Proof. By Assumption C.8, $p^{(a)}_{XY}(x, y) = \frac{u_a(x)v_a(y)}{Z_a} p^\star_{XY}(x, y)$. Thus

$$p^{(a)}_X(x) = \int p^{(a)}_{XY}(x, y)\, dy = \frac{u_a(x)p^\star_X(x)}{Z_a}\, \mathbb{E}^\star[v_a(Y) \mid X = x],$$

and similarly

$$p^{(a)}_Y(y) = \frac{v_a(y)p^\star_Y(y)}{Z_a}\, \mathbb{E}^\star[u_a(X) \mid Y = y].$$

Plugging into $r^{(a)}(x, y) = \frac{p^{(a)}_{XY}(x,y)}{p^{(a)}_X(x)p^{(a)}_Y(y)}$ gives the stated identity, and taking logs yields the PMI form. □

Lemma C.9 shows that curation perturbs the density ratio by a multiplicative factor governed by the conditional expectations $\mathbb{E}^\star[v(Y) \mid X = x]$ and $\mathbb{E}^\star[u(X) \mid Y = y]$. Below, we impose a mild condition on these terms, requiring that dataset curation acts independently across modalities i.e the expected acceptance rate of texts does not depend on the image they are paired with, and vice versa.

Assumption C.10. For each dataset $a \in \{1, 2\}$,

$$\mathbb{E}^\star[v_a(Y) \mid X = x] = \mathbb{E}^\star[v_a(Y)] \quad \text{for } P^\star_X\text{-a.e. } x, \qquad \mathbb{E}^\star[u_a(X) \mid Y = y] = \mathbb{E}^\star[u_a(X)] \quad \text{for } P^\star_Y\text{-a.e. } y.$$

Theorem C.11. *Under Assumption C.8 and Assumption C.10, for each $a \in \{1, 2\}$ there exists a constant $C_a \in \mathbb{R}$ such that $K^{(a)}(x, y) = K^\star(x, y) + C_a$ for $P^\star_X \otimes P^\star_Y$-a.e. (x, y). Consequently, there exists a constant $\Delta \in \mathbb{R}$ such that*

$$K^{(1)}(x, y) = K^{(2)}(x, y) + \Delta$$

for $P^\star_X \otimes P^\star_Y$-a.e. (x, y).

Proof. Under Assumption C.10, the conditional expectations in Lemma C.9 are constants: $\mathbb{E}^\star[v_a(Y) \mid X = x] = \mathbb{E}^\star[v_a(Y)]$ and $\mathbb{E}^\star[u_a(X) \mid Y = y] = \mathbb{E}^\star[u_a(X)]$ a.e. Hence $r^{(a)}(x, y) = \alpha_a\, r^\star(x, y)$ a.e., where

$$\alpha_a := \frac{Z_a}{\mathbb{E}^\star[u_a(X)]\, \mathbb{E}^\star[v_a(Y)]},$$

so $K^{(a)}(x, y) = K^\star(x, y) + \log \alpha_a$ a.e. The second claim follows by subtraction with $\Delta = \log \alpha_1 - \log \alpha_2$. □

Corollary C.12. *Let s^*_1 and s^*_2 be measurable global minimizers of the symmetric objective $\mathcal{L}_N$ when the InfoNCE sampling model is defined under $P^{(1)}_{XY}$ and $P^{(2)}_{XY}$, respectively (cf. Section C.1). Under Theorem C.11, there exists $\Delta' \in \mathbb{R}$ such that*

$$s^*_1(x, y) = s^*_2(x, y) + \Delta'$$

for $P^\star_X \otimes P^\star_Y$-a.e. (x, y).

Proof. By Theorem C.5 applied to each dataset, $s_a^*(x,y) = K^{(a)}(x,y) + C'_a$ for constants C'_a (a.e.). By Theorem C.11, $K^{(1)}(x,y) = K^{(2)}(x,y) + \Delta$ (a.e.), hence $s_1^* - s_2^*$ is constant. □

Remark C.13 (Implication for contrastive dot-product kernels). If each critic is realized through (f, g) and $(\tilde{f}, \tilde{g})$, then Corollary C.12 implies

$$\langle f(x), g(y)\rangle = \langle \tilde{f}(x), \tilde{g}(y)\rangle + \widetilde{\Delta}$$

for $P_X^\star \otimes P_Y^\star$-a.e. (x,y), for some constant $\widetilde{\Delta} \in \mathbb{R}$. If the temperatures differ, the relation becomes affine in $\langle \tilde{f}(x), \tilde{g}(y)\rangle$.

C.5 Guarantees Under Exact Alignment

C.5.1 Linear Alignment of Independent Contrastive Models

In this section, we prove our first main result showing that matching multimodal kernels on a finite set of anchor points is sufficient to lock the two embedding spaces together up to a linear transformation.

Let $\Omega_X \subseteq \mathcal{X}$ and $\Omega_Y \subseteq \mathcal{Y}$ be subsets (the domain of interest, e.g. a downstream dataset of images and prompts). Let the contrastive model pairs (f, g) and $(\tilde{f}, \tilde{g})$ map inputs to $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and $\mathbb{S}^{\tilde{d}-1} \subset \mathbb{R}^{\tilde{d}}$ respectively, where $d \leq \tilde{d}$, without loss of generality. We fix a set of *image anchors* $\{\bar{x}_j\}_{j=1}^d \subset \Omega_X$ and *text anchors* $\{\bar{y}_i\}_{i=1}^{\tilde{d}} \subset \Omega_Y$ and collect their embeddings into the following matrices:

$$\begin{aligned} G &\coloneqq [g(\bar{y}_1) \ \cdots \ g(\bar{y}_{\tilde{d}})] \in \mathbb{R}^{d\times\tilde{d}}, \\ \tilde{G} &\coloneqq [\tilde{g}(\bar{y}_1) \ \cdots \ \tilde{g}(\bar{y}_{\tilde{d}})] \in \mathbb{R}^{\tilde{d}\times\tilde{d}}, \\ F &\coloneqq [f(\bar{x}_1) \ \cdots \ f(\bar{x}_d)] \in \mathbb{R}^{d\times d}. \end{aligned}$$

Assumption 5.2. The multimodal kernels coincide on the set of anchors:

$$\begin{aligned} \langle f(x), g(\bar{y}_i)\rangle &= \langle \tilde{f}(x), \tilde{g}(\bar{y}_i)\rangle \ \ \forall x \in \Omega_X, \ \forall i \in \{1,\dots,\tilde{d}\}, \\ \langle f(\bar{x}_j), g(y)\rangle &= \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle \ \ \forall y \in \Omega_Y, \ \forall j \in \{1,\dots,d\}. \end{aligned}$$

Theorem 5.3. *(Linear Identifiability, proof in Section C.5.1). Suppose $\tilde{G}$ and F are invertible. Then there exists a linear map A such that $\tilde{f}(x) = Af(x)\ \forall x \in \Omega_X$. Further, if A has full column rank, then for every $y \in \Omega_Y$, $\mathrm{Proj}_{\mathrm{Im}(A)}\, \tilde{g}(y) = A(A^\top A)^{-1} g(y)$. If $\tilde{d} = d$, then $\tilde{g}(y) = A^{-\top} g(y)$.*

Proof. Fix $x \in \Omega_X$ and define

$$k(x) := \big(\langle f(x), g(\bar{y}_i)\rangle\big)_{i=1}^{\tilde{d}} = G^\top f(x), \qquad \tilde{k}(x) := \big(\langle \tilde{f}(x), \tilde{g}(\bar{y}_i)\rangle\big)_{i=1}^{\tilde{d}} = \tilde{G}^\top \tilde{f}(x).$$

By Assumption 5.2, $k(x) = \tilde{k}(x)$, hence $G^\top f(x) = \tilde{G}^\top \tilde{f}(x)$. Since $\tilde{G}$ is invertible,

$$\tilde{f}(x) = \tilde{G}^{-\top} G^\top f(x) = Af(x),$$

for all $x \in \Omega_X$. Now fix $y \in \Omega_Y$. For each anchor $\bar{x}_j$,

$$\langle f(\bar{x}_j), g(y)\rangle = \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle = \langle Af(\bar{x}_j), \tilde{g}(y)\rangle = \langle f(\bar{x}_j), A^\top \tilde{g}(y)\rangle.$$

Thus $\langle f(\bar{x}_j), g(y) - A^\top \tilde{g}(y)\rangle = 0$ for all $j \in \{1,\dots,d\}$. Since $F = [f(\bar{x}_1) \ \cdots \ f(\bar{x}_d)]$ is invertible, $\{f(\bar{x}_j)\}_{j=1}^d$ spans $\mathbb{R}^d$, hence $g(y) = A^\top \tilde{g}(y)$ for all $y \in \Omega_Y$.

Finally, if A has full column rank then $P := A(A^\top A)^{-1} A^\top$ is the orthogonal projector onto $\mathrm{Im}(A)$, and

$$\mathrm{Proj}_{\mathrm{Im}(A)} \tilde{g}(y) = P\tilde{g}(y) = A(A^\top A)^{-1} A^\top \tilde{g}(y) = A(A^\top A)^{-1} g(y).$$

If $\tilde{d} = d$ and A is invertible, then $A(A^\top A)^{-1} = A^{-\top}$, giving $\tilde{g}(y) = A^{-\top} g(y)$. □

C.5.2 Isometric Alignment of Independent Contrastive Models

Theorem 5.3 establishes that the representation is fixed up to a linear transformation A. However, standard contrastive encoders normalize embeddings to the unit hypersphere $\mathbb{S}^{d-1}$, forcing $\|\tilde{f}(x)\|_2 = \|Af(x)\|_2 = 1$ everywhere. This forces A to be an isometry ($A^\top A = I_d$) only if the data is sufficiently diverse to probe the matrix in all directions. We formalize this diversity via the following condition.

Definition 5.4. *($\mathrm{Sym}(d)$-spanning)* A set of vectors $S \subset \mathbb{R}^d$ is $\mathrm{Sym}(d)$-*spanning* if the rank-one matrices $\{xx^\top : x \in S\}$ span the space of symmetric matrices $\mathrm{Sym}(d)$. Equivalently: if $M \in \mathrm{Sym}(d)$ and $x^\top Mx = 0$ for all $x \in S$, then $M = 0$.

Lemma C.14. *Let $A \in \mathbb{R}^{\tilde{d}\times d}$ and let $U \subseteq \mathbb{S}^{d-1} \subset \mathbb{R}^d$. Assume $\|Au\|_2 = 1$ for all $u \in U$. If U contains a $\mathrm{Sym}(d)$-spanning subset, then $A^\top A = I_d$.*

Proof. Let $M := A^\top A - I_d \in \mathrm{Sym}(d)$. For any $u \in \mathbb{S}^{d-1}$,

$$\|Au\|_2^2 = u^\top A^\top Au = u^\top (I_d + M)u = 1 + u^\top Mu.$$

Thus $\|Au\|_2 = 1$ implies $u^\top Mu = 0$ for all $u \in U$. Choose a $\mathrm{Sym}(d)$-spanning subset $S = \{u_i\} \subseteq U$. Then $u_i^\top Mu_i = 0$ for all i implies $M = 0$ by definition of $\mathrm{Sym}(d)$-spanning, hence $A^\top A = I_d$. □

Theorem 5.5. *(Orthogonal Identifiability, proof in Section C.5.2). Assume the conditions of Theorem 5.3 hold. If the set of image embeddings $\{f(x) : x \in \Omega_X\}$ contains a $\mathrm{Sym}(d)$-spanning subset, then the linear map A has orthonormal columns ($A^\top A = I_d$). Consequently, $\tilde{f}(x) = Qf(x) \quad \forall x \in \Omega_X$, where $Q := A$ satisfies $Q^\top Q = I_d$. Furthermore, for the other modality: $\mathrm{Proj}_{\mathrm{Im}(Q)}\, \tilde{g}(y) = Qg(y) \quad \forall y \in \Omega_Y$. If $\tilde{d} = d$, Q is orthogonal i.e. $Q \in O(d)$ and $\tilde{g}(y) = Qg(y)$.*

Proof. By Theorem 5.3, $\tilde{f}(x) = Af(x)$ for all $x \in \Omega_X$. Since $\|f(x)\|_2 = \|\tilde{f}(x)\|_2 = 1$, we have $\|Af(x)\|_2 = 1$ for all $x \in \Omega_X$. Apply Lemma C.14 to conclude $A^\top A = I_d$. Set $Q := A$.

By Theorem 5.3, $g(y) = Q^\top \tilde{g}(y)$ for all $y \in \Omega_Y$. Multiplying by Q gives

$$Qg(y) = QQ^\top \tilde{g}(y) = \mathrm{Proj}_{\mathrm{Im}(Q)}\, \tilde{g}(y),$$

since $QQ^\top$ is the orthogonal projector onto $\mathrm{Im}(Q)$ when $Q^\top Q = I_d$. If $\tilde{d} = d$, then $\mathrm{Im}(Q) = \mathbb{R}^d$ and $\mathrm{Proj}_{\mathrm{Im}(Q)}$ is the identity. □

Remark C.15. Suppose the exact regime of Theorem 5.5 holds: $\tilde{f}(x) = Qf(x)$ for all $x \in \Omega_X$ (and similarly for g), and suppose the means are taken w.r.t. the same distribution on Ω_X: $\mu_f := \mathbb{E}[f(X)]$, $\mu_{\tilde{f}} := \mathbb{E}[\tilde{f}(X)]$. Then $\mu_{\tilde{f}} = Q\mu_f$, hence

$$\tilde{f}(x) - \mu_{\tilde{f}} = Q\big(f(x) - \mu_f\big).$$

Therefore the rigid map $z \mapsto Q(z - \mu_f) + \mu_{\tilde{f}}$ reduces exactly to the pure rotation $z \mapsto Qz$. In this sense, mean-centering is a finite-sample / distribution-mismatch correction that leaves the exact-identifiability statement unchanged.

C.5.3 Isometric Alignment of Independent Contrastive Models When One Modality Lies in Low Dimension

The above analysis assumes that $\mathrm{span}\{f(x) : x \in \Omega_X\} = \mathbb{R}^d$ since we assumed the existence of d anchor vectors from Ω_X that make F invertible. In other works, we assumed the existence of d linearly independent vectors $f(\bar{x}_1), \ldots f(\bar{x}_d)$, that span $\mathbb{R}^d$. We extend this analysis below in Theorem C.16 to cases where these embeddings instead lie in a lower-dimensional subspace.

Let

$$U_X := \mathrm{span}\{f(x) : x \in \Omega_X\} \subseteq \mathbb{R}^d, \qquad \tilde{U}_X := \mathrm{span}\{\tilde{f}(x) : x \in \Omega_X\} \subseteq \mathbb{R}^{\tilde{d}}.$$

where U_X is a subspace of $\mathbb{R}^d$ with rank $r := \dim(U_X)$. Choose anchors $\bar{x}_1, \ldots, \bar{x}_r \in \Omega_X$ such that $\mathrm{span}\{f(\bar{x}_1), \ldots, f(\bar{x}_r)\} = U_X$.

Theorem C.16 (*Orthogonal Identifiability*)**.** *Assume multimodal kernel alignment as in Assumption 5.2 but on r anchors $\bar{x}_1, \dots, \bar{x}_r \in \Omega_X$ instead of d. Assume there exists a matrix $Q \in \mathbb{R}^{\tilde{d} \times d}$ with $Q^\top Q = I_d$ such that $\tilde{f}(x) = Qf(x)$ for all $x \in \Omega_X$. Then $\tilde{U}_X = QU_X$ and for every $y \in \Omega_Y$,*

$$\mathrm{Proj}_{\tilde{U}_X} \tilde{g}(y) = Q \,\mathrm{Proj}_{U_X} g(y).$$

In particular, if $U_X = \mathbb{R}^d$, then $\mathrm{Proj}_{\mathrm{Im}(Q)}\tilde{g}(y) = Qg(y)$ for all $y \in \Omega_Y$, and if additionally $\tilde{d} = d$ then $\tilde{g}(y) = Qg(y)$.

Proof. Since $\tilde{f}(x) = Qf(x)$ and $Q^\top Q = I_d$, for each anchor $\bar{x}_j$ with $j \in \{1, \dots, r\}$ and $y \in \Omega_Y$,

$$\langle f(\bar{x}_j), g(y)\rangle = \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle = \langle Qf(\bar{x}_j), \tilde{g}(y)\rangle = \langle f(\bar{x}_j), Q^\top \tilde{g}(y)\rangle.$$

Hence $\langle f(\bar{x}_j), g(y) - Q^\top \tilde{g}(y)\rangle = 0$ for all anchors $\bar{x}_j$. By linearity, $\langle u, g(y) - Q^\top \tilde{g}(y)\rangle = 0$ for all $u \in U_X$, so $g(y) - Q^\top \tilde{g}(y) \in U_X^\perp$. Equivalently, $\mathrm{Proj}_{U_X} g(y) = \mathrm{Proj}_{U_X}(Q^\top \tilde{g}(y))$. Applying Q and using $\tilde{U}_X = QU_X$ and $\mathrm{Proj}_{\tilde{U}_X} = Q \,\mathrm{Proj}_{U_X} Q^\top$ yields

$$\mathrm{Proj}_{\tilde{U}_X} \tilde{g}(y) = Q \,\mathrm{Proj}_{U_X} Q^\top \tilde{g}(y) = Q \,\mathrm{Proj}_{U_X} g(y).$$

If $U_X = \mathbb{R}^d$, then $\tilde{U}_X = Q\mathbb{R}^d = \mathrm{Im}(Q)$, giving $\mathrm{Proj}_{\mathrm{Im}(Q)}\tilde{g}(y) = Qg(y)$. If moreover $\tilde{d} = d$, then $\mathrm{Im}(Q) = \mathbb{R}^d$ and the projection is the identity. □

C.5.4 Alignment Upto Classification Boundaries

While the preceding theory establishes conditions for exact geometric alignment. In practice however, one often cares about alignment up to concepts or classification that are relevant for the downstream task. We now analyze this regime, where we seek to distinguish a finite family of prompts $\mathcal{Y}_{\mathrm{cls}} = \{y_c\}_{c=1}^K \subseteq \Omega_Y$ even when pointwise alignment is imprecise. Specifically, we care about the within-modality cross-model top-1 retrieval accuracy i.e.

$$\hat{c}(y_c) \;\in\; \arg\max_{c' \in \{1,\dots,K\}} \langle Q\, g(y_c), \tilde{g}(y_{c'})\rangle.$$

Theorem C.16 is stated for all $y \in \Omega_Y$ and its proof is pointwise in y and uses the multimodal kernel equalities only for the pairs $(\bar{x}_j, y)$ with $j \in \{1, \dots, r\}$. Therefore, if one only needs the conclusion for a subset $\mathcal{Y}_0 \subseteq \Omega_Y$ (e.g. $\mathcal{Y}_0 = \mathcal{Y}_{\mathrm{cls}}$), it suffices to assume the anchor kernel equalities only on $\{\bar{x}_1, \dots, \bar{x}_r\} \times \mathcal{Y}_0$. Formally,

Assumption C.17 (Anchor kernel equalities only for class prompts)**.** For the image anchors $\bar{x}_1, \dots, \bar{x}_r$ spanning U_X, assume that for all class prompts $y \in \mathcal{Y}_{\mathrm{cls}}$,

$$\langle f(\bar{x}_j), g(y)\rangle = \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle \qquad \forall j \in \{1, \dots, r\}.$$

Now, for any class prompt $y \in \mathcal{Y}_{\mathrm{cls}}$, decompose the embedding into an identifiable signal $u(y)$ and an unidentifiable residual $w(y)$:

$$g(y) = u(y) + w(y), \;\; u(y) \coloneqq \mathrm{Proj}_{U_X} g(y), \;\; w(y) \in U_X^\perp.$$

Then, from Theorem C.16, define analogously

$$\tilde{g}(y) = Qu(y) + \tilde{w}(y), \qquad \tilde{w}(y) \in (QU_X)^\perp.$$

Definition C.18. Let $\mathcal{Y}_{\mathrm{cls}} = \{y_c\}_{c=1}^K$ Define

$$\gamma := \min_c \Big(\|u(y_c)\|_2^2 - \max_{c' \neq c} \langle u(y_c), u(y_{c'})\rangle \Big),$$

that measures class separability *within the image span* U_X. Define,

$$\eta := \max_{c,c'} \big|\langle Qw(y_c), \tilde{w}(y_{c'})\rangle\big|.$$

that measures the worst-case cross-model interaction of the unidentifiable residuals across models.

Proposition C.19. *Under Assumption C.17, if $\gamma > 2\eta$, then nearest-neighbor class retrieval is correct for every class prompt:*

$$\hat{c}(y_c) = c \qquad \forall c \in \{1, \dots, K\}.$$

Proof. The conclusion of Theorem C.16 under Assumption C.17 implies

$$\tilde{U}_X = QU_X \quad \text{and} \quad \mathrm{Proj}_{\tilde{U}_X}\, \tilde{g}(y) = Q\, \mathrm{Proj}_{U_X}\, g(y) \;\; \forall y \in \mathcal{Y}_{\mathrm{cls}}$$

Fix a query class c and a candidate c'. Using $g(y) = u(y) + w(y)$ and $\tilde{g}(y) = Qu(y) + \tilde{w}(y)$,

$$\langle Qg(y_c), \tilde{g}(y_{c'})\rangle = \langle Q(u(y_c) + w(y_c)), Qu(y_{c'}) + \tilde{w}(y_{c'})\rangle.$$

$Qu(y_c)$ is orthogonal to $\tilde{w}(y_{c'})$ and similarly $\langle Qw(y_c), Qu(y_{c'})\rangle = 0$ since $w(y_c) \in U_X^{\perp}$. Also, $Q^\top Q = I_d$ implies $\langle Qu, Qu'\rangle = \langle u, u'\rangle$. Thus,

$$\langle Qg(y_c), \tilde{g}(y_{c'})\rangle = \langle u(y_c), u(y_{c'})\rangle + \langle Qw(y_c), \tilde{w}(y_{c'})\rangle.$$

By the definition of γ, for any $c' \neq c$ we have $\langle u(y_c), u(y_{c'})\rangle \leq \|u(y_c)\|_2^2 - \gamma$. Therefore, for all $c' \neq c$,

$$\langle Qg(y_c), \tilde{g}(y_c)\rangle - \langle Qg(y_c), \tilde{g}(y_{c'})\rangle \geq (\|u(y_c)\|_2^2 - \eta) - (\|u(y_c)\|_2^2 - \gamma + \eta) = \gamma - 2\eta.$$

If $\gamma > 2\eta$, the right-hand side is strictly positive, so c is the unique maximizer and $\hat{c}(y_c) = c$. □

Proposition C.19 shows that class retrieval depends on a *margin* condition inside U_X, and is stable as long as the residual terms in $U_X^{\perp}$ do not overwhelm that margin.

C.6 Guarantees Under Approximate Alignment

We now relax the anchor equalities in Section C.5 and allow small discrepancies on the anchor pairs. Like before, we assume the contrastive model pairs (f, g) and $(\tilde{f}, \tilde{g})$ to map inputs to $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and $\mathbb{S}^{\tilde{d}-1} \subset \mathbb{R}^{\tilde{d}}$ respectively, where $d \leq \tilde{d}$, without loss of generality.

C.6.1 Approximate Linear Alignment of Independent Contrastive Models

Similar to before, fix $\tilde{d}$ text anchors $\bar{y}_1, \dots, \bar{y}_{\tilde{d}} \in \Omega_Y$ and define

$$G := [g(\bar{y}_1)\ \cdots\ g(\bar{y}_{\tilde{d}})] \in \mathbb{R}^{d\times\tilde{d}}, \qquad \tilde{G} := [\tilde{g}(\bar{y}_1)\ \cdots\ \tilde{g}(\bar{y}_{\tilde{d}})] \in \mathbb{R}^{\tilde{d}\times\tilde{d}}.$$

Assumption C.20 (Approximate multimodal kernel equalities)**.** There exists $\varepsilon \geq 0$ such that for all $x \in \Omega_X$ and all $i \in \{1, \dots, \tilde{d}\}$,

$$\left|\langle f(x), g(\bar{y}_i)\rangle - \langle \tilde{f}(x), \tilde{g}(\bar{y}_i)\rangle\right| \leq \varepsilon.$$

Theorem C.21. *Assume $\tilde{G}$ is invertible with smallest singular value $\sigma_{\min}(\tilde{G}) > 0$. Then under Assumption C.20, there exists a linear map A such that for every $x \in \Omega_X$,*

$$\left\|\tilde{f}(x) - Af(x)\right\|_2 \leq \frac{\sqrt{\tilde{d}}}{\sigma_{\min}(\tilde{G})}\,\varepsilon.$$

Proof. Define $k(x) := G^\top f(x) \in \mathbb{R}^{\tilde{d}}$ and $\tilde{k}(x) := \tilde{G}^\top \tilde{f}(x) \in \mathbb{R}^{\tilde{d}}$. Set $A := \tilde{G}^{-\top} G^\top \in \mathbb{R}^{\tilde{d}\times d}$, then

$$\tilde{f}(x) - Af(x) = \tilde{f}(x) - \tilde{G}^{-\top} G^\top f(x) = \tilde{G}^{-\top}\big(\tilde{G}^\top \tilde{f}(x) - G^\top f(x)\big) = \tilde{G}^{-\top}\big(\tilde{k}(x) - k(x)\big).$$

Therefore

$$\left\|\tilde{f}(x) - Af(x)\right\|_2 \leq \|\tilde{G}^{-\top}\|_2\, \|\tilde{k}(x) - k(x)\|_2 = \frac{1}{\sigma_{\min}(\tilde{G})}\, \|\tilde{k}(x) - k(x)\|_2.$$

By Assumption C.20, each coordinate of $\tilde{k}(x) - k(x)$ has magnitude at most ε, so $\|\tilde{k}(x) - k(x)\|_2 \leq \sqrt{\tilde{d}}\,\varepsilon$. □

C.6.2 Approximate Isometric Alignment of Independent Contrastive Models

Assumption C.22 ($\mathrm{Sym}(d)$-*spanning*)**.** There exist $\bar{x}_1, \dots, \bar{x}_m \in \Omega_X$ such that $S = \{f(\bar{x}_1), \dots, f(\bar{x}_m)\} \subset \mathbb{S}^{d-1}$ is $\mathrm{Sym}(d)$-spanning

Fix $A \in \mathbb{R}^{\tilde{d}\times d}$ and define the uniform deviation

$$\Delta := \sup_{x\in\Omega_X} \|\tilde{f}(x) - Af(x)\|_2.$$

Also define the constant

$$\kappa_S := \sup_{M\in\mathrm{Sym}(d)\setminus\{0\}} \frac{\|M\|_2}{\max_{u\in S}|u^\top M u|} < \infty.$$

Theorem C.23 (Approximate Orthogonal Identifiability)**.** *Assume A has full column rank. Then under Assumption C.22, there exists $Q^\top Q = I_d$, such that for all $x \in \Omega_X$*

$$\|\tilde{f}(x) - Qf(x)\|_2 \le \Delta + \kappa_S\,(2+\Delta)\Delta$$

Proof. Let $M := A^\top A - I_d \in \mathrm{Sym}(d)$. For any $u = f(\bar{x}_i) \in S$, $\|\tilde{f}(\bar{x}_i)\|_2 = 1$ and $\|Au - \tilde{f}(\bar{x}_i)\|_2 \le \Delta$ imply $|\|Au\|_2 - 1| \le \Delta$ and hence $|\|Au\|_2^2 - 1| \le (2+\Delta)\Delta$. But $\|Au\|_2^2 - 1 = u^\top M u$, so $\max_{u\in S}|u^\top M u| \le (2+\Delta)\Delta$. By definition of κ_S, $\|M\|_2 \le \kappa_S \max_{u\in S}|u^\top M u|$, showing that

$$\|A^\top A - I_d\|_2 \le \kappa_S\,(2+\Delta)\Delta.$$

Now, let $Q \in \mathbb{R}^{\tilde{d}\times d}$ denote the orthogonal factor in the polar decomposition of A i.e. $A = Q(A^\top A)^{1/2}$, $Q^\top Q = I_d$. Then $\|A - Q\|_2 = \|(A^\top A)^{1/2} - I_d\|_2 \le \|A^\top A - I_d\|_2$. Finally, $\|\tilde{f}(x) - Qf(x)\|_2 \le \|\tilde{f}(x) - Af(x)\|_2 + \|(A-Q)f(x)\|_2 \le \Delta + \|A-Q\|_2$, proving the claim. □

Now, fix d image anchors $\bar{x}_1, \dots, \bar{x}_d \in \Omega_X$ and define

$$F := [f(\bar{x}_1)\ \cdots\ f(\bar{x}_d)] \in \mathbb{R}^{d\times d}, \qquad \sigma_{\min}(F) > 0.$$

Assumption C.24 (Approximate multimodal kernel equalities)**.** There exists $\varepsilon' \ge 0$ such that for all $y \in \Omega_Y$ and all $j \in \{1, \dots, d\}$,

$$\left|\langle f(\bar{x}_j), g(y)\rangle - \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle\right| \le \varepsilon'.$$

Theorem C.25. *Assume Assumption C.24 and let $Q \in \mathbb{R}^{\tilde{d}\times d}$ satisfy $Q^\top Q = I_d$ and*

$$\max_{j\in\{1,\dots,d\}} \left\|\tilde{f}(\bar{x}_j) - Qf(\bar{x}_j)\right\|_2 \le \delta_f.$$

Then for every $y \in \Omega_Y$,

$$\left\|\mathrm{Proj}_{\mathrm{Im}(Q)}\tilde{g}(y) - Qg(y)\right\|_2 = \|Q^\top \tilde{g}(y) - g(y)\|_2 \le \frac{\sqrt{d}}{\sigma_{\min}(F)}\,(\varepsilon' + \delta_f).$$

In particular, if $\tilde{d} = d$ then $\mathrm{Im}(Q) = \mathbb{R}^d$ and the projection is redundant, giving a bound on $\|\tilde{g}(y) - Qg(y)\|_2$.

Proof. Fix $y \in \Omega_Y$ and set $v := g(y) - Q^\top \tilde{g}(y) \in \mathbb{R}^d$. For each anchor $\bar{x}_j$,

$$\langle f(\bar{x}_j), v\rangle = \langle f(\bar{x}_j), g(y)\rangle - \langle Qf(\bar{x}_j), \tilde{g}(y)\rangle.$$

Thus

$$|\langle f(\bar{x}_j), v\rangle| \le |\langle f(\bar{x}_j), g(y)\rangle - \langle \tilde{f}(\bar{x}_j), \tilde{g}(y)\rangle| + |\langle \tilde{f}(\bar{x}_j) - Qf(\bar{x}_j), \tilde{g}(y)\rangle| \le \varepsilon' + \delta_f,$$

since $\|\tilde{g}(y)\|_2 = 1$. Hence $\|F^\top v\|_2 \le \sqrt{d}(\varepsilon' + \delta_f)$ and, since F is invertible,

$$\|v\|_2 \le \|F^{-\top}\|_2 \|F^\top v\|_2 = \frac{1}{\sigma_{\min}(F)}\|F^\top v\|_2 \le \frac{\sqrt{d}}{\sigma_{\min}(F)}(\varepsilon' + \delta_f).$$

Finally, using $Q^\top Q = I_d$,

$$\|\mathrm{Proj}_{\mathrm{Im}(Q)}\tilde{g}(y) - Qg(y)\|_2 = \|Q(Q^\top \tilde{g}(y) - g(y))\|_2 = \|v\|_2.$$

□

Remark C.26. Assumption C.20 and Assumption C.24 are ε-relaxations of Assumption 5.2. Theorem C.21 and Theorem C.25 are the perturbation analogues of Theorem 5.3 and Theorem 5.5 respectively.

C.6.3 Finite Anchor-Anchor Alignment

A purely finite version of this result also follows from a short reduction to our existing results.

Proposition C.27 (Finite anchors under representation-space coverage)**.** *Let finite anchor sets*

$$\{\bar{x}_j\}_{j=1}^{N_X}$$

and

$$\{\bar{y}_i\}_{i=1}^{N_Y}$$

contain the basis anchors used in Assumption 5.2, and define

$$\rho_X := \sup_{x\in\Omega_X} \min_j \Big(\|f(x) - f(\bar{x}_j)\|_2 + \|\widetilde{f}(x) - \widetilde{f}(\bar{x}_j)\|_2\Big),$$

with ρ_Y defined analogously. Suppose that the kernel discrepancy is at most ε on the finite set

$$\{\bar{x}_1, \dots, \bar{x}_{N_X}\} \times \{\bar{y}_1, \dots, \bar{y}_{\tilde{d}}\} \;\cup\; \{\bar{x}_1, \dots, \bar{x}_d\} \times \{\bar{y}_1, \dots, \bar{y}_{N_Y}\}.$$

Then the approximate multimodal kernel agreement assumptions Assumptions C.20 and C.24 on kernel sections *hold with errors $\varepsilon + \rho_X$ and $\varepsilon + \rho_Y$, respectively.*

Proof. Indeed, for arbitrary x, choose an anchor $\bar{x}_j$ attaining the definition of ρ_X. For every basis text anchor $\bar{y}_i$, unit normalization and the triangle inequality give

$$\Big|\langle f(x), g(\bar{y}_i)\rangle - \langle \widetilde{f}(x), \widetilde{g}(\bar{y}_i)\rangle\Big| \le \varepsilon + \|f(x) - f(\bar{x}_j)\|_2 + \|\widetilde{f}(x) - \widetilde{f}(\bar{x}_j)\|_2 \le \varepsilon + \rho_X.$$

The text-side statement is identical. Thus Theorems C.21, C.23 and C.25 immediately yield a single global approximate orthogonal map aligning both modalities, with bounds already established in the paper. □

Finite kernel evaluations do not constrain unseen representations, making a coverage condition necessary. Our condition is purely representation-space based, requiring no input-space metric or smoothness.

C.6.4 Approximate Isometric Alignment of Independent Contrastive Models When One Modality Lies in Low Dimension

Similar to Theorem C.16, we extend our analysis below to cases where the image embeddings lie in a lower-dimensional subspace. Let $U_X := \mathrm{span}\{f(x) : x \in \Omega_X\} \subseteq \mathbb{R}^d$ where U_X is a subspace of $\mathbb{R}^d$ of $\dim(U_X) = r$. Choose anchors $\bar{x}_1, \dots, \bar{x}_r \in \Omega_X$ such that $\mathrm{span}\{f(\bar{x}_1), \dots, f(\bar{x}_r)\} = U_X$. Accordingly, denote

$$F := [f(\bar{x}_1)\ \cdots\ f(\bar{x}_r)] \in \mathbb{R}^{d\times r}, \qquad \sigma_{\min}(F) > 0.$$

Assumption C.28. Assume there exists a matrix $Q \in \mathbb{R}^{\tilde{d}\times d}$ with $Q^\top Q = I_d$ and constants $\delta_f, \varepsilon \ge 0$ such that

$$\sup_{x\in\Omega_X} \|\tilde{f}(x) - Qf(x)\|_2 \le \delta_f, \qquad \sup_{x\in\Omega_X,\, y\in\Omega_Y} \big|\langle f(x), g(y)\rangle - \langle \tilde{f}(x), \tilde{g}(y)\rangle\big| \le \varepsilon.$$

Theorem C.29. *Under Assumption C.28, for every $y \in \Omega_Y$,*

$$\Big\|\mathrm{Proj}_{U_X}\big(g(y) - Q^\top \tilde{g}(y)\big)\Big\|_2 \le \rho, \qquad \rho := \frac{\sqrt{r}}{\sigma_{\min}(F)}\,(\varepsilon + \delta_f).$$

Equivalently,

$$\Big\|\mathrm{Proj}_{QU_X}\tilde{g}(y) - Q\,\mathrm{Proj}_{U_X} g(y)\Big\|_2 \le \rho.$$

Proof. Fix $y \in \Omega_Y$ and define $h := g(y) - Q^\top \tilde{g}(y)$. For any $x \in \Omega_X$,

$$\begin{aligned}\big|\langle f(x), h\rangle\big| &= \big|\langle f(x), g(y)\rangle - \langle Qf(x), \tilde{g}(y)\rangle\big| \\ &\le \big|\langle f(x), g(y)\rangle - \langle \tilde{f}(x), \tilde{g}(y)\rangle\big| + \big|\langle \tilde{f}(x) - Qf(x), \tilde{g}(y)\rangle\big| \le \varepsilon + \delta_f,\end{aligned}$$

since $\|\tilde{g}(y)\|_2 = 1$. Let $b \in \mathbb{R}^r$ have coordinates $b_j := \langle f(\bar{x}_j), h\rangle$. Then $\|b\|_2 \le \sqrt{r}(\varepsilon + \delta_f)$. Moreover, $F^\top \mathrm{Proj}_{U_X} h = F^\top h = b$, and since F has full column rank,

$$\|\mathrm{Proj}_{U_X} h\|_2 \le \|(F^\top)^\dagger\|_2 \, \|b\|_2 = \frac{1}{\sigma_{\min}(F)} \, \|b\|_2 \le \frac{\sqrt{r}}{\sigma_{\min}(F)}(\varepsilon + \delta_f).$$

The equivalent bound in $\tilde{U}_X$ follows by left-multiplying by Q and using $\mathrm{Proj}_{QU_X} = Q \, \mathrm{Proj}_{U_X} \, Q^\top$ (which holds whenever $Q^\top Q = I_d$).

□

D Supplementary Experimental Details and Assets Disclosure

D.1 Assets

We do not introduce new data in the course of this work. Instead, we use publicly available, widely used image datasets for the purposes of benchmarking and comparison.

D.2 Hardware and Setup

Each experiment was conducted on 1 NVIDIA Tesla V100 GPUs, each with 32GB of accelerator RAM. The CPUs used were Intel Xeon E5-2698 v4 processors with 20 cores and 384GB of RAM. All experiments were implemented using the PyTorch deep learning framework. We provide our experimental code for reproducibility in the supplementary material.

D.3 Datasets

D.3.1 Benchmarks

We evaluate cross-model alignment on eight benchmarks. For image-text classification, we use Oxford Pets (Parkhi et al., 2012), CIFAR-100 (Krizhevsky et al., 2009), Caltech-101 (Fei-Fei et al., 2004), STL10 (Coates et al., 2011), and DTD (Cimpoi et al., 2014). For audio-text classification, we use ESC-50 (Piczak, 2015). For fine-grained image-text retrieval, we use Flickr30k (Plummer et al., 2015) and MSCOCO (Lin et al., 2014). Dataset statistics and split details are provided in Table 1.

Table 1: Detailed statistics of the 8 datasets used in our experiments.

Dataset	Classes	Train	Val	Test
Caltech-101 (Fei-Fei et al., 2004)	100	4,128	1,649	2,465
Oxford Pets (Parkhi et al., 2012)	37	2,944	736	3,669
CIFAR100 (Krizhevsky et al., 2009)	100	50,000	-	10,000
STL10 (Coates et al., 2011)	10	50,000	-	8,000
DTD (Cimpoi et al., 2014)	47	2,820	1,128	1,692
ESC-50 (Piczak, 2015)	50	1,600	-	400
Flickr30k (Plummer et al., 2015)	-	29,000	1,000	1,000
MS-COCO (Lin et al., 2014)	-	113,287	5,000	5,000

D.3.2 Constructing Text Templates

We construct class-level text prototypes for all classification datasets by instantiating natural-language templates with each class name and embedding the resulting strings using each model's text encoder. We exclude Flickr30k from this prototype construction because it is a fine-grained image-text retrieval benchmark with human-written captions rather than class labels; converting it to class-name prompts would discard the instance-specific caption information that defines the task. All templates are taken directly from the official CLIP repository and prior CLIP evaluation code, ensuring that our protocol matches standard zero-shot classification practice. Table 2 summarizes the exact prompts used for

each dataset. We embed all prompts for a class, aggregate them into a single prototype by averaging, and then ℓ_2-normalizing. Concretely, for templates $\{\tau_k(c)\}_{k=1}^{K}$ for a class c,

$$p_c = \text{norm}\Big(\frac{1}{K}\sum_k g(\tau_k(c))\Big) \tag{1}$$

When the orthogonal map Q is fit on images and evaluated on text, we use these averaged class prototypes p_c for text-side evaluation. In contrast, when fitting Q on text, averaging would collapse each class to a single vector and drastically reduce supervision; therefore, we fit Q on individual prompt embeddings $g(\tau_k(c))_{k=1}^{K}$, treating each template instantiation as a separate training sample.

Dataset	Text Templates (from official CLIP repo)
Oxford Pets	`a photo of a {class}, a type of pet.`

CIFAR-100	`a photo of a {class}.`
	`a blurry photo of a {class}.`
	`a black and white photo of a {class}.`
	`a low contrast photo of a {class}.`
	`a high contrast photo of a {class}.`
	`a bad photo of a {class}.`
	`a good photo of a {class}.`
	`a photo of a small {class}.`
	`a photo of a big {class}.`
	`a photo of the {class}.`
	`a blurry photo of the {class}.`
	`a black and white photo of the {class}.`
	`a low contrast photo of the {class}.`
	`a high contrast photo of the {class}.`
	`a bad photo of the {class}.`
	`a good photo of the {class}.`
	`a photo of the small {class}.`
	`a photo of the big {class}.`

Caltech-101	`a photo of a {class}.`
	`a painting of a {class}.`
	`a plastic {class}.`
	`a sculpture of a {class}.`
	`a sketch of a {class}.`
	`a tattoo of a {class}.`
	`a toy {class}.`
	`a rendition of a {class}.`
	`a embroidered {class}.`
	`a cartoon {class}.`
	`a {class} in a video game.`
	`a plushie {class}.`
	`a origami {class}.`
	`art of a {class}.`
	`graffiti of a {class}.`
	`a drawing of a {class}.`
	`a doodle of a {class}.`
	`a photo of the {class}.`
	`a painting of the {class}.`
	`the plastic {class}.`
	`a sculpture of the {class}.`
	`a sketch of the {class}.`
	`a tattoo of the {class}.`
	`the toy {class}.`
	`a rendition of the {class}.`
	`the embroidered {class}.`
	`the cartoon {class}.`
	`the {class} in a video game.`
	`the plushie {class}.`
	`the origami {class}.`
	`art of the {class}.`
	`graffiti of the {class}.`
	`a drawing of the {class}.`
	`a doodle of the {class}.`

STL-10	`a photo of a {class}.` `a photo of the {class}.`

DTD	`a photo of a {class} texture.`
	`a photo of a {class} pattern.`
	`a photo of a {class} thing.`
	`a photo of a {class} object.`
	`a photo of the {class} texture.`
	`a photo of the {class} pattern.`
	`a photo of the {class} thing.`
	`a photo of the {class} object.`

Table 2: Text templates used to construct class-level text prototypes for each dataset. All templates are taken from the official CLIP repository. For all reported results, we use a single template per class for evaluation

D.4 Training and Evaluation Protocol

D.4.1 Learning An Orthogonal Transformation

We consider two independently trained multimodal contrastive models with image encoders $f, \tilde{f}$ and text encoders $g, \tilde{g}$. Our goal is to learn a single orthogonal map $\mathcal{R} \in O(d)$ that aligns representations across models. All experiments use fixed train/validation/test splits with a fixed random seed. We report the mean and standard deviation averaged over three random seeds. We fit $\mathcal{R}$ using paired

embeddings from a single modality (images or text). Let $\{x_i\}_{i=1}^n$ denote paired inputs in the chosen modality, and define

$$X = \begin{bmatrix} f(x_1)^\top \\ \vdots \\ f(x_n)^\top \end{bmatrix}, \qquad \tilde{X} = \begin{bmatrix} \tilde{f}(x_1)^\top \\ \vdots \\ \tilde{f}(x_n)^\top \end{bmatrix},$$

with $f, \tilde{f}$ replaced by $g, \tilde{g}$ when fitting on text. For numerical stability, we center the embeddings only during estimation i.e. $X \to X - \mu_X$, $\tilde{X} \to \tilde{X} - \mu_{\tilde{X}}$, and compute the closed form Orthogonal Procrustes solution

$$\mathcal{Q} \;=\; \arg\min_{Q \in O(d)} \|XQ - \tilde{X}\|_F \;=\; UV^\top,$$

where $U\Sigma V^\top = \mathrm{SVD}(X^\top \tilde{X})$. At evaluation, we deploy $\mathcal{Q}$ as a pure orthogonal transformation on the centered point clouds,

$$z \;\mapsto\; Q\big(z - \mu_A^{(\cdot)}\big) + \mu_B^{(\cdot)},$$

where $\mu_A^{(\cdot)}$ and $\mu_B^{(\cdot)}$ are modality-specific means of the source and target embeddings (image or text, as appropriate). Centering isolates the rotational relationship by removing this offset while preserving the orthogonal correspondence in the centered space. Note that, we also ablate using pure orthogonal transformation on the raw embedding point clouds and find that mean centering mainly improves *pointwise* cosine agreement, while class-level retrieval and decision geometry remain essentially unchanged even without mean-centering (see Section E.14).

For more details, refer to the pseudocode in Algorithm 1.

Algorithm 1 Pseudocode for fitting an orthogonal cross-model map $\mathcal{R}$ (Procrustes)

```
# f, g: source image/text encoders with outputs of unit-norm
# f_tilde, g_tilde: target image/text encoders with outputs of unit-norm
# pairs: paired samples from ONE modality (images x_i or texts y_i)
# fit_modality: "image" or "text"
# norm(v): l2-normalize v

def fit_procrustes(pairs, fit_modality = 'image'):
   X, X_tilde = [], []
   for s in pairs:
      if fit_modality == "image":
         X.append(f(s))
         X_tilde.append(f_tilde(s))
      else:
         X.append(g(s))
         X_tilde.append(g_tilde(s))

   X = stack_rows(X) # (n, d)
   X_tilde = stack_rows(X_tilde) # (n, d)

   # center for numerical stability during estimation
   Xc = X - mean_row(X)
   Xtc = X_tilde - mean_row(X_tilde)

   # orthogonal Procrustes
   M = transpose(Xc) @ Xtc
   U, S, Vt = svd(M)
   R = U @ Vt

# deployment (used for BOTH modalities at evaluation time)
def deploy(z, R, mu, mu_tilde):
   return R @ (z-mu) + mu_tilde

R = fit_procrustes(pairs, fit_modality)
```

D.4.2 Performance Metrics at Evaluation

We evaluate alignment at two granularities: the instance level and the class level. For the former, we report paired-image cosine $\langle \mathcal{R}f(x), \tilde{f}(x) \rangle$ and paired-text cosine $\langle \mathcal{R}g(y), \tilde{g}(y) \rangle$ and the correspondingly the Euclidean distance between image embeddings ($Rf(x)$ and $\tilde{f}(x)$) and text embeddings ($Rg(y)$ and $\tilde{g}(y)$), measuring how well a single orthogonal map $\mathcal{R}$ matches corresponding embeddings across models at the instance level. For the class level alignment, we report:

- **Intra-modal alignment across models.** We measure class-level image–image and text–text top-1 retrieval across models. For image–image, each query is an aligned image embedding $\mathcal{R}f(x)$, and we predict its class by nearest-neighbor search over the target image embedding set $\{\tilde{f}(x')\}_{x'}$, counting a hit if the retrieved neighbor shares the same class label. For text-text, each query is an aligned class text embedding $\mathcal{R}g(y_c)$, and we retrieve within the target class text embedding set $\{\tilde{g}(y_{c'})\}_{c'}$, counting a hit if the retrieved class equals c.
- **Cross-modal transfer across models (task accuracy).** We evaluate zero-shot classification under three retrieval settings, each probing a different notion of transfer: (a) aligned images with target text $\hat{c} = \arg\max_c \langle \mathcal{R}f(x), \tilde{g}(y_c) \rangle$ measures whether image–text semantics are preserved when only the image space is aligned; (b) target images with aligned text $\hat{c} = \arg\max_c \langle \tilde{f}(x), \mathcal{R}g(y_c) \rangle$ measures whether text semantics transfer under alignment without modifying the image space; (c) aligned images with aligned text $\hat{c} = \arg\max_c \langle \mathcal{R}f(x), \mathcal{R}g(y_c) \rangle$ measures whether a single map $\mathcal{R}$ induces a coherent shared space across both modalities.

For Flickr30k, which is a fine-grained image-text retrieval benchmark rather than a classification dataset, we replace class-level accuracy with cross-modal retrieval. We evaluate image-to-text and text-to-image retrieval after alignment, reporting Recall@K between mapped source embeddings and target-model embeddings. This tests whether the same orthogonal alignment preserves instance-specific image-caption semantics, rather than only class-level structure.

In all, we report five metrics: (1) paired-instance cosine (or euclidean distance), either image $\langle \mathcal{R}f(x), , \tilde{f}(x) \rangle$ or text $\langle \mathcal{R}g(y), , \tilde{g}(y) \rangle$ depending on the transfer direction; (2) class-level image-image top-1 retrieval or text-text top-1 retrieval; (3) zero-shot accuracy with aligned images and target text, $\hat{c} = \arg\max_c \langle \mathcal{R}f(x), , \tilde{g}(y_c) \rangle$; (4) zero-shot accuracy with target images and aligned text, $\hat{c} = \arg\max_c \langle \tilde{f}(x), , \mathcal{R}g(y_c) \rangle$; and (5) zero-shot accuracy with aligned images and aligned text, $\hat{c} = \arg\max_c \langle \mathcal{R}f(x), , \mathcal{R}g(y_c) \rangle$. Throughout the evaluation, every inner product $\langle \cdot, \cdot \rangle$ is measured through *cosine similarity*, i.e., a dot product between ℓ_2-normalized embeddings.

E Additional Experiments

E.1 Modality Gap in Contrastive Models

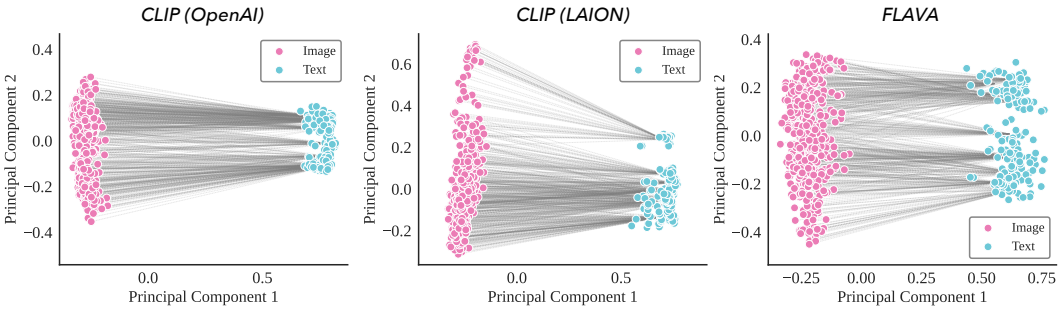

Figure 9: PCA visualization in 2D of generated embeddings of CIFAR-100 from pre-trained models showing pervasive modality gap in multi-modal contrastive representation learning whereby image and text embeddings are located in two completely separate regions of the embedding space.

Figure 9 visualizes the modality gap in contrastive models via a 2D PCA projection of image and text embeddings. Across CLIP (OpenAI), CLIP (LAION), and FLAVA, image and text representations

form two well-separated clusters, occupying distinct regions of the embedding space rather than a unified manifold. Gray lines connect matched image-text pairs, highlighting that although the pairs are semantically aligned, their embeddings remain geometrically distant.

E.2 Visualization of Multimodal Kernels

In Figure 10, we visualize the PCA projection of the image/text geometry across models. We observe that while the absolute coordinates of the embedding cones shift arbitrarily between models, the angular arrangement of the texts with respect to the images remains consistent. Mathematically, this means that the *multimodal kernels* are approximately preserved across models: $\langle f, g \rangle \approx \langle \tilde{f}, \tilde{g} \rangle$. This observation can be viewed as a multimodal analogue of the Platonic Representation Hypothesis (Huh et al., 2024) that posits that models converge to similar *unimodal* kernels ($\langle f, f \rangle \approx \langle \tilde{f}, \tilde{f} \rangle$).

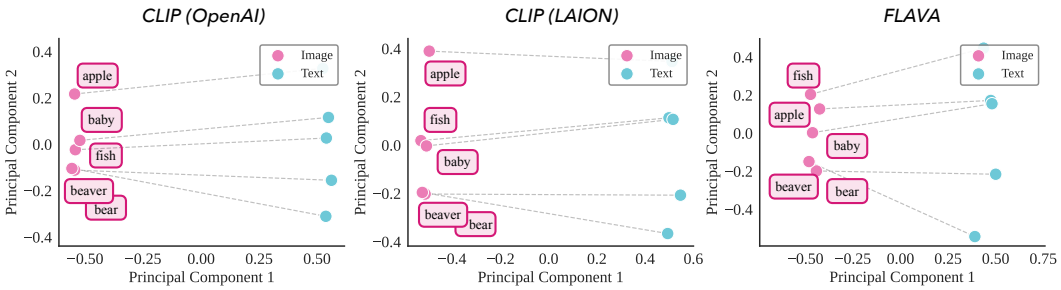

Figure 10: Across CLIP variants, the multimodal kernel $\langle f(x), g(y) \rangle$ (relative angles between image and text embeddings) is strongly preserved (dashed lines), unlike the unimodal kernel $\langle f(x), \tilde{f}(x') \rangle$;

E.3 Independently Trained Image-Text Contrastive Models Differ by an Orthogonal Map That Is Shared Across Modalities

In this section, we report complete results across 5 benchmarks—Oxford Pets (Figure 11), Caltech-101 (Figure 12), CIFAR-100 (Figure 13), STL10 (Figure 14) and DTD (Figure 15), covering all five evaluation metrics used throughout the paper and explained in Section D.4.2. Two consistent conclusions emerge across datasets and model pairs. First, a single orthogonal map $\mathcal{Q}$ accurately aligns image embeddings between independently trained contrastive models, yielding near-oracle image-image agreement. Second, the *same* map transfers across modalities: applying $\mathcal{Q}$ learned from images substantially improves text-text pointwise alignment, cross-model retrieval, and zero-shot classification, without degrading image performance. Together, these results confirm that independently trained contrastive models are related by a shared, modality-invariant orthogonal reparameterization.

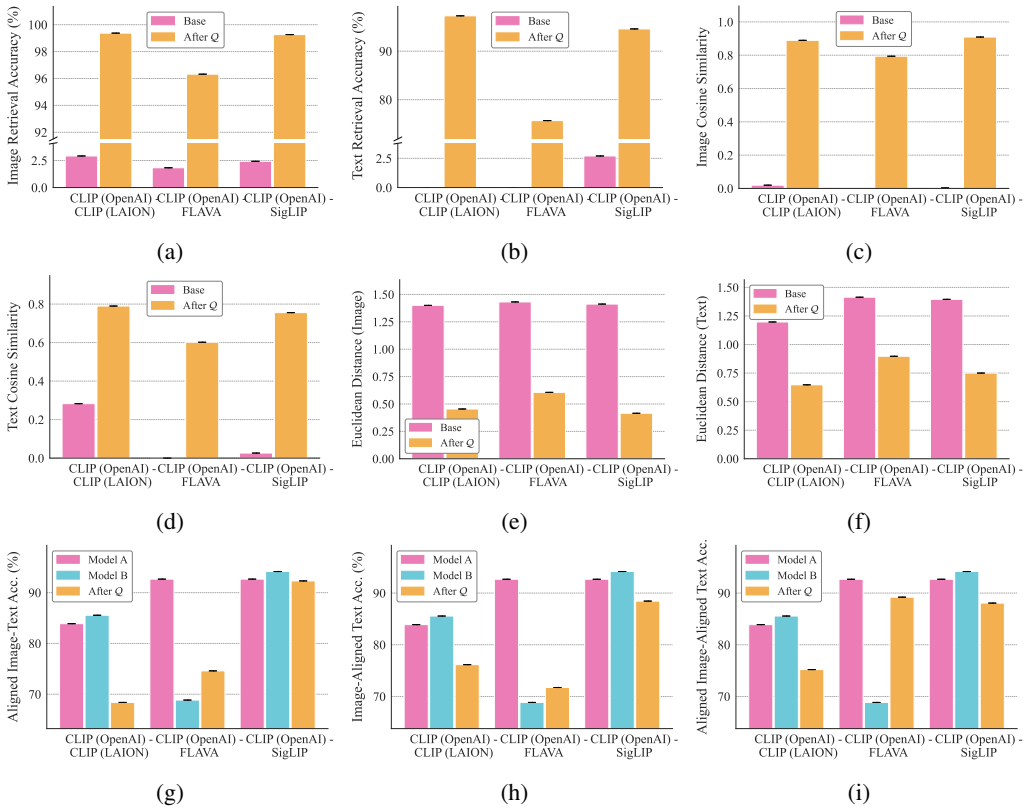

Figure 11: *Cross-model alignment on Oxford Pets across independently trained contrastive models before and after fitting a single orthogonal map $\mathcal{R}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Euclidean distance (without normalization) between image embeddings. (f) Euclidean distance (without normalization) between text embeddings. (g) Image–text retrieval using aligned images from model A and text from model B. (h) Image–text retrieval using images from model B and aligned text from model A. (i) Image–text retrieval using aligned images and aligned text from model A. $\mathcal{Q}$ aligns images across models with a single orthogonal map, and the same $\mathcal{Q}$ learned only from image embeddings transfers to text, boosting text-text retrieval from near-chance to near-oracle, all while preserving strong image classification accuracy.

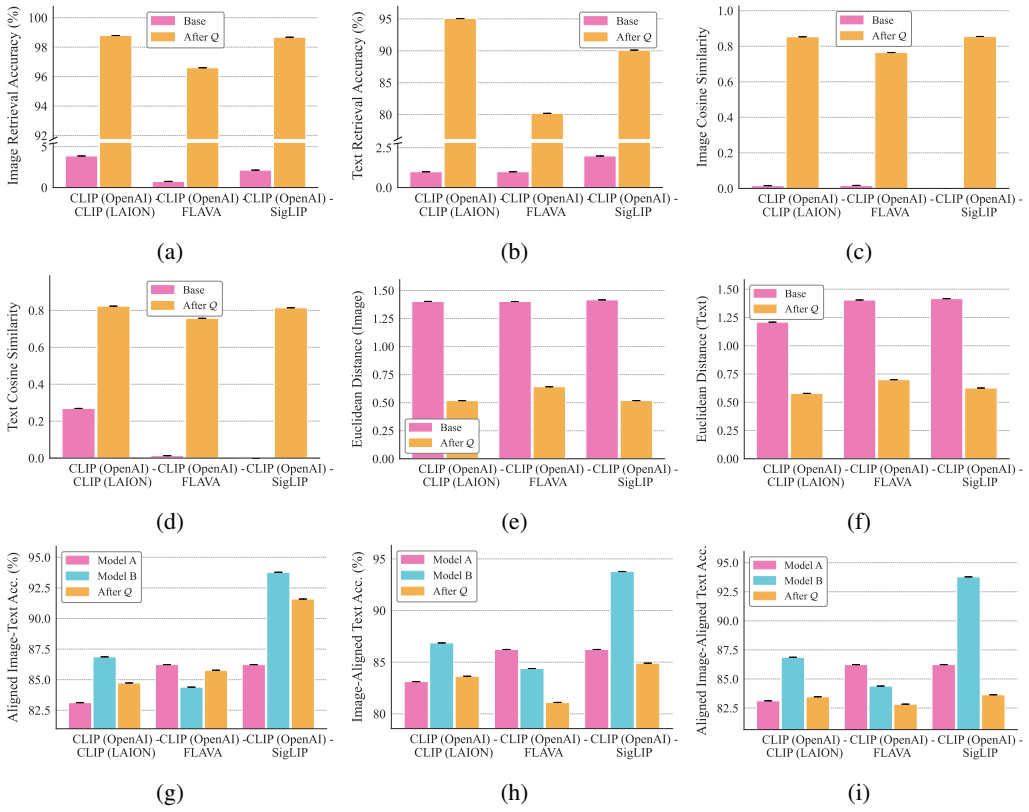

Figure 12: *Cross-model alignment on Caltech-101 across independently trained contrastive models before and after fitting a single orthogonal map $\mathcal{R}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Euclidean distance (without normalization) between image embeddings. (f) Euclidean distance (without normalization) between text embeddings. (g) Image–text retrieval using aligned images from model A and text from model B. (h) Image–text retrieval using images from model B and aligned text from model A. (i) Image–text retrieval using aligned images and aligned text from model A. $\mathcal{Q}$ aligns images across models with a single orthogonal map, and the same $\mathcal{Q}$ learned only from image embeddings transfers to text, boosting text-text retrieval from near-chance to near-oracle, all while preserving strong image classification accuracy.

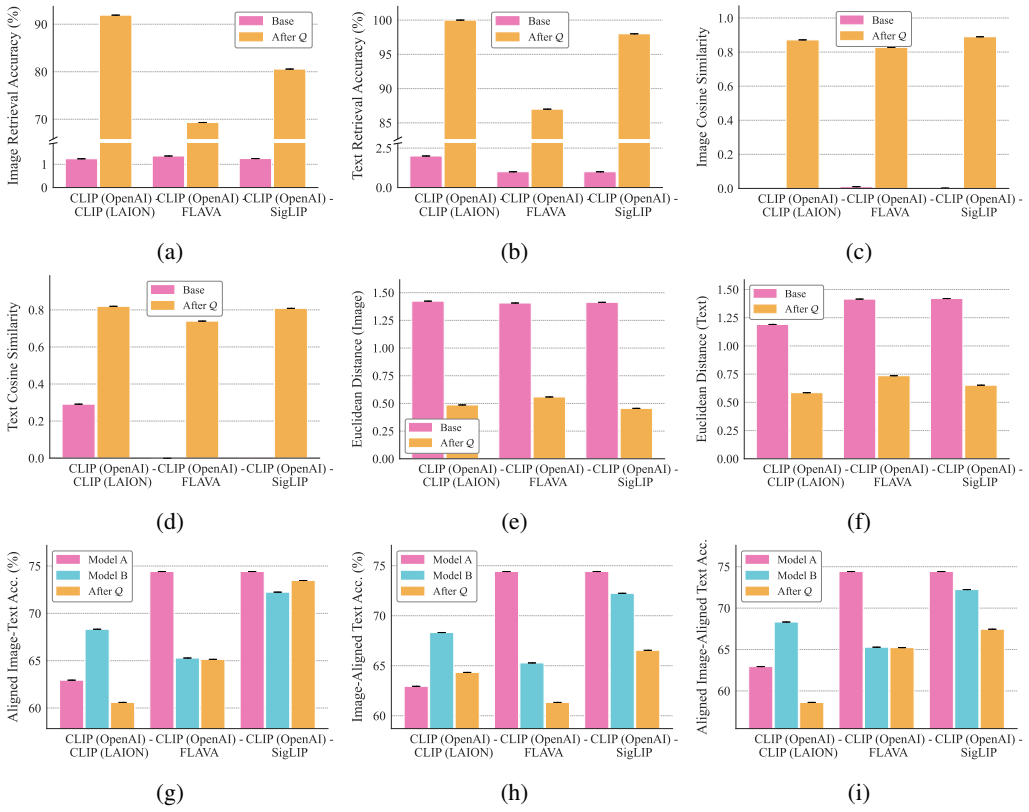

Figure 13: *Cross-model alignment on CIFAR-100 across independently trained contrastive models before and after fitting a single orthogonal map $\mathcal{R}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Euclidean distance (without normalization) between image embeddings. (f) Euclidean distance (without normalization) between text embeddings. (g) Image–text retrieval using aligned images from model A and text from model B. (h) Image–text retrieval using images from model B and aligned text from model A. (i) Image–text retrieval using aligned images and aligned text from model A. $\mathcal{Q}$ aligns images across models with a single orthogonal map, and the same $\mathcal{Q}$ learned only from image embeddings transfers to text, boosting text-text retrieval from near-chance to near-oracle, all while preserving strong image classification accuracy.

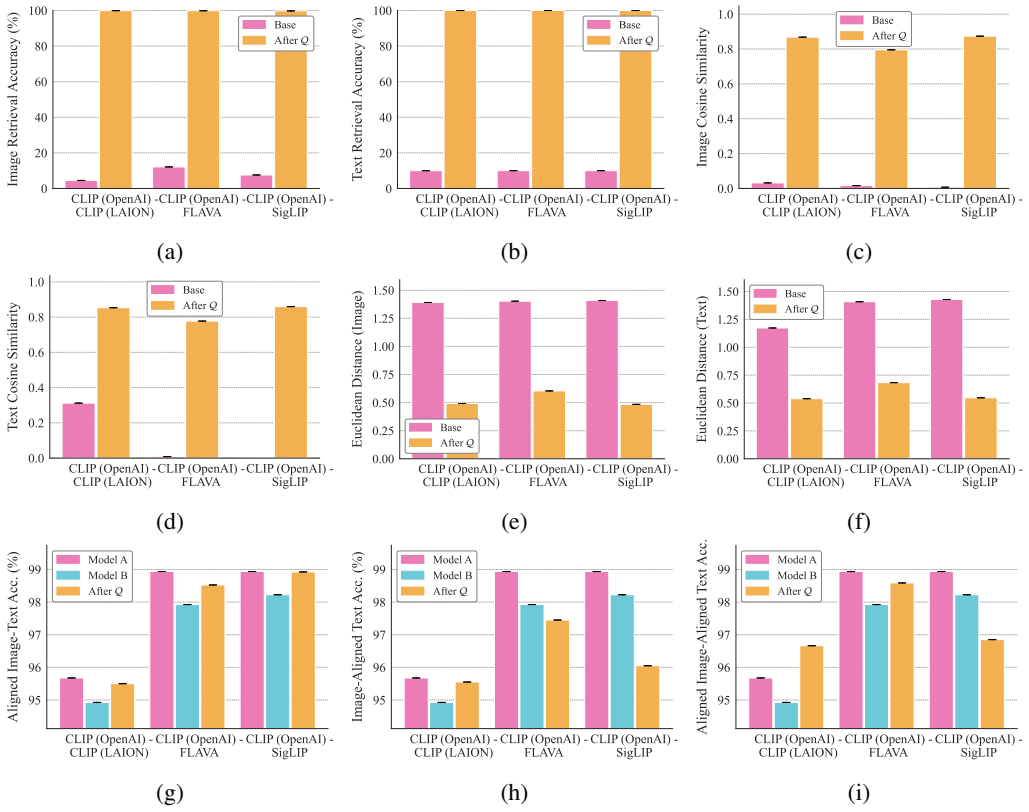

Figure 14: *Cross-model alignment on STL10 across independently trained contrastive models before and after fitting a single orthogonal map $\mathcal{R}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Euclidean distance (without normalization) between image embeddings. (f) Euclidean distance (without normalization) between text embeddings. (g) Image–text retrieval using aligned images from model A and text from model B. (h) Image–text retrieval using images from model B and aligned text from model A. (i) Image–text retrieval using aligned images and aligned text from model A. $\mathcal{Q}$ aligns images across models with a single orthogonal map, and the same $\mathcal{Q}$ learned only from image embeddings transfers to text, boosting text-text retrieval from near-chance to near-oracle, all while preserving strong image classification accuracy.

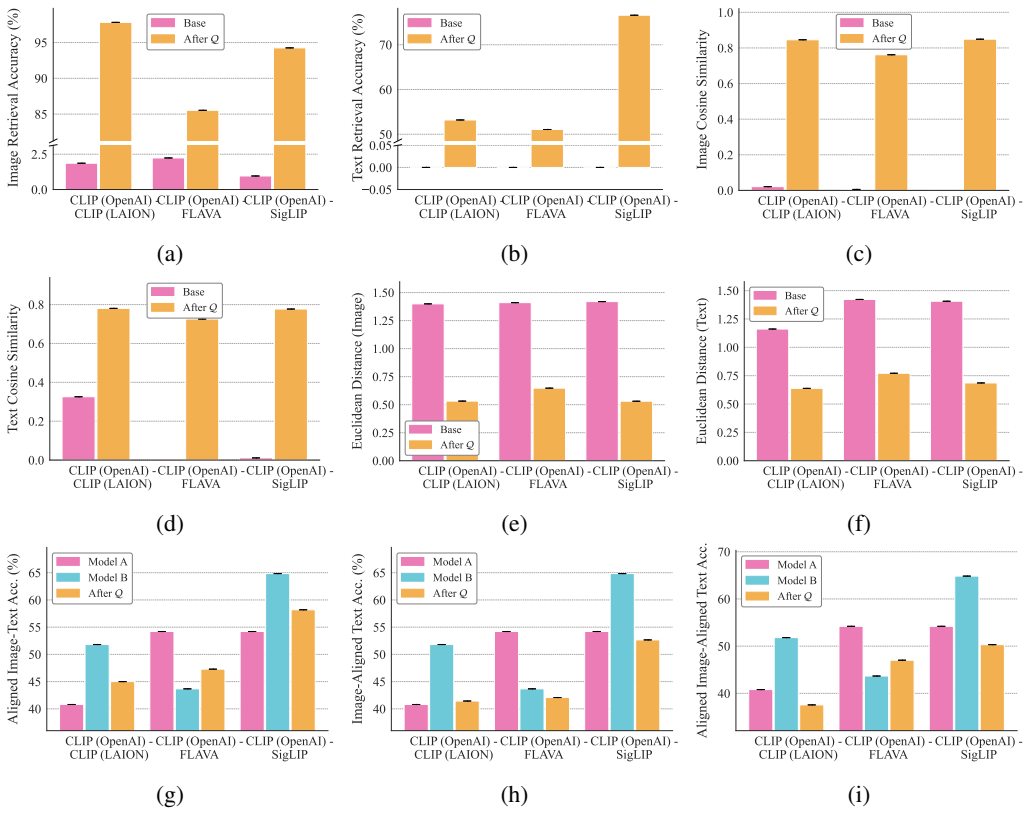

Figure 15: *Cross-model alignment on DTD across independently trained contrastive models before and after fitting a single orthogonal map $\mathcal{R}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Euclidean distance (without normalization) between image embeddings. (f) Euclidean distance (without normalization) between text embeddings. (g) Image–text retrieval using aligned images from model A and text from model B. (h) Image–text retrieval using images from model B and aligned text from model A. (i) Image–text retrieval using aligned images and aligned text from model A. $\mathcal{Q}$ aligns images across models with a single orthogonal map, and the same $\mathcal{Q}$ learned only from image embeddings transfers to text, boosting text-text retrieval from near-chance to near-oracle, all while preserving strong image classification accuracy.

E.4 Extension to Audio Modality

We extend our framework beyond vision–language by aligning two independently trained CLAP models using the ESC-50 dataset. We consider two settings: (i) LAION-CLAP (HTSAT+RoBERTa) to Microsoft-CLAP (HTSAT+GPT-2), which have *different embedding dimensions*, and (ii) LAION-630K-best to LAION-630K-AudioSet-best, two checkpoints with the same architecture but different training data. As shown in Figure 16 and Figure 17, across both cases, Q is fit only on audio embeddings. As shown below, Q significantly improves pointwise cosine similarity on both audio and text modalities, and closely recovers the target model's native zero-shot accuracy, confirming that our findings generalize to audio modality.

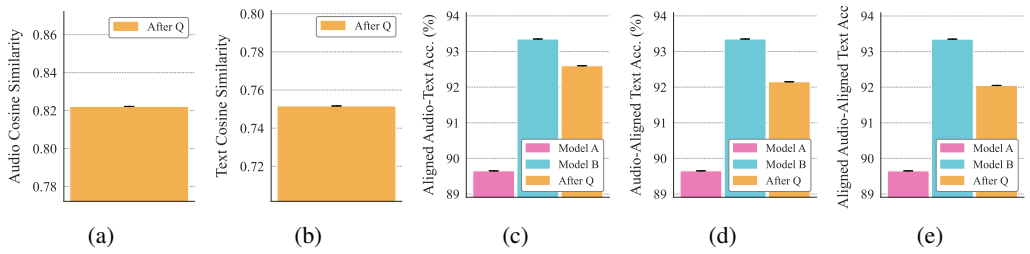

Figure 16: *Cross-model alignment on ESC-50 from LAION-CLAP (HTSAT+RoBERTa) to Microsoft-CLAP (HTSAT+GPT-2), before and after fitting a single orthogonal map $\mathcal{Q}$.*

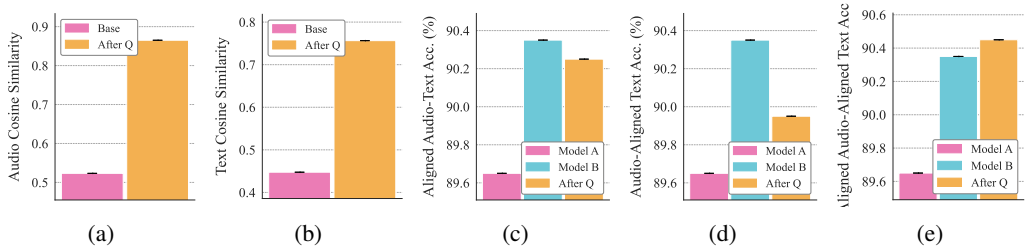

Figure 17: *Cross-model alignment on ESC-50 from LAION-630K-best to LAION-630K-AudioSet-best, before and after fitting a single orthogonal map $\mathcal{Q}$.*

E.5 Matched and Mismatched Cross-Model Similarity

To test whether Q aligns individual examples rather than just the embedding means, we compare matched and mismatched similarities. Here, $f, \tilde{f}$ are the image encoders and $g, \tilde{g}$ are the text encoders of the two models. For each modality, c^+ is the mean cosine similarity between matched inputs and c^- is the mean over mismatched inputs. We further report the separation statistic

$$z = \frac{c^+ - c^-}{\mathrm{std}(c^-)},$$

which normalizes the gap between matched and mismatched similarities by the variability of mismatched pairs. As shown in Table 3, the raw embeddings exhibit essentially no separation between matched and mismatched pairs ($z \approx 0$). After applying Q, matched similarities become substantially higher than mismatched similarities ($z_I = 4.52$, $z_T = 4.13$), while preserving strong native compatibility.

Table 3: Matched and mismatched cross-model similarities on Caltech-101 for CLIP ViT-B/32 (OpenAI) and CLIP ViT-B/32 (LAION-400M).

Method	$\langle f, g \rangle$	$\langle \tilde{f}, \tilde{g} \rangle$	$\langle Qf, \tilde{g} \rangle$	$\langle \tilde{f}, Qg \rangle$	$\langle Qf, Qg \rangle$	c_I^+	c_I^-	c_T^+	c_T^-	z_I	z_T
Baseline	82.4	87.0	–	–	–	0.015	0.015	0.269	0.270	0.00	-0.04
Q	82.4	87.0	84.9	82.6	82.2	0.853	0.421	0.822	0.640	4.52	4.13

E.6 Extension to Fine-Grained Retrieval

We further evaluate instance-level image-text retrieval on Flickr30k and MS-COCO, providing a stronger test of fine-grained cross-modal alignment than classification. On Flickr30k, we embed images with Model A (CLIP ViT-B/32, OpenAI), map them via Q into Model B's space (CLIP ViT-B/32, LAION-400M), and retrieve against Model B's text embeddings. As shown in Figure 18, after applying Q, image cosine similarity increases from 0.01 to 0.84, and image-text retrieval R@10 reaches 97.10, compared with Model B's native 96.70, despite Q being fit solely on images.

We additionally evaluate at larger scale on MS-COCO using CLIP ViT-B/32 (OpenAI) and CLIP ViT-B/16 (OpenAI). Here, $f, \tilde{f}$ are the image encoders and $g, \tilde{g}$ are the text encoders of the two models. As shown in Table 4, the map estimated from one modality transfers to the unseen modality and remains strongly compatible with the native target embeddings: $\langle Qf, \tilde{g} \rangle$ achieves image-to-text/text-to-image Recall@1 of 48.0/27.8, compared with the native target model's 51.8/30.3, while $\langle \tilde{f}, Qg \rangle$ achieves 43.8/28.6.

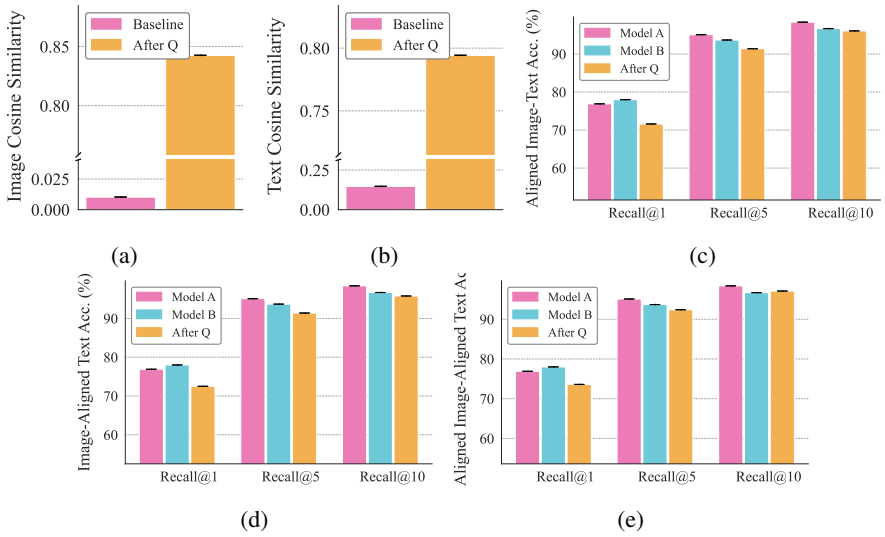

Figure 18: *Cross-model alignment from CLIP ViT-B/32 (OpenAI) to CLIP ViT-B/32 (LAION-400M) on Flickr30k*

Table 4: Fine-grained retrieval and cross-model alignment on MS-COCO. Retrieval entries report image-to-text/text-to-image Recall@1 (%); $c^{\pm}$ and z report matched-mismatched cross-model similarity statistics as detailed in Section E.5.

Method	$\langle f, g \rangle$	$\langle \tilde{f}, \tilde{g} \rangle$	$\langle Qf, \tilde{g} \rangle$	$\langle \tilde{f}, Qg \rangle$	$\langle Qf, Qg \rangle$	c_I^+	c_I^-	c_T^+	c_T^-	z_I	z_T
Baseline	50.2/30.6	51.8/30.3	–	–	–	0.009	0.008	-0.001	0.002	0.03	-0.08
After Q	50.2/30.6	51.8/30.3	48.0/27.8	43.8/28.6	45.9/29.5	0.902	0.539	0.885	0.549	4.56	4.43

E.7 Only a Few Data Points Are Needed to Learn the Orthogonal Map

In Section C.5.2, we proved that if the multimodal kernels induced by two contrastive models agree on a sufficiently rich but *finite* anchor set, then a single global orthogonal map aligns their representations across both modalities. We empirically validate this prediction by fitting $\mathcal{Q}$ using paired images from only N classes and evaluating transfer on the remaining unseen classes. We report this analysis across additional model pairs and datasets. The results for CLIP (OpenAI) aligned to CLIP (OpenAI), CLIP (LAION), FLAVA, and SigLIP on Caltech-101 are shown in Figures 19 to 22. For CIFAR100 and Oxford Pets, the analogous results are reports in Figures 23 to 26 and Figures 27 to 30.

Across all settings, the same trend holds: performance on both seen and unseen classes improves quickly and then saturates at a small fraction of the data, after which additional anchors provide little benefit. Thus, practitioners can recover near-full cross-model transfer by fitting $\mathcal{Q}$ on a lightweight image-only calibration set, rather than curating large-scale cross-model supervision.

Below, we quantify N^*, the fewest anchor classes at which unseen-class accuracy reaches within 5% of its maximum. Note that for datasets with many classes (CIFAR-100, Caltech-101), our current sweep uses a granularity of 20 classes. As shown in Table 5, across all 12 configurations, the median N^*_{95} is 25 classes, and 11 out of 12 settings reach 95% of peak with $\leq$40 classes.

Table 5: Number of anchors required to reach 95% of peak alignment performance across datasets and model pairs.

Dataset (N_{total})	Model Pair	N^*_{95}	$N^*_{95}/N_{\text{total}}$
Oxford Pets (37)	CLIP B/32 to CLIP B/32 (LAION)	25	68%
	CLIP L/14 to FLAVA	15	41%
	CLIP L/14 to SigLIP	15	41%
	CLIP B/32 to CLIP B/16	20	54%
Caltech-101 (101)	CLIP B/32 to CLIP B/32 (LAION)	40	40%
	CLIP L/14 to FLAVA	25	25%
	CLIP L/14 to SigLIP	40	40%
	CLIP B/32 to CLIP B/16	40	40%
CIFAR-100 (100)	CLIP B/32 to CLIP B/32 (LAION)	5	5%
	CLIP L/14 to FLAVA	60	60%
	CLIP L/14 to SigLIP	5	5%
	CLIP B/32 to CLIP B/16	40	40%

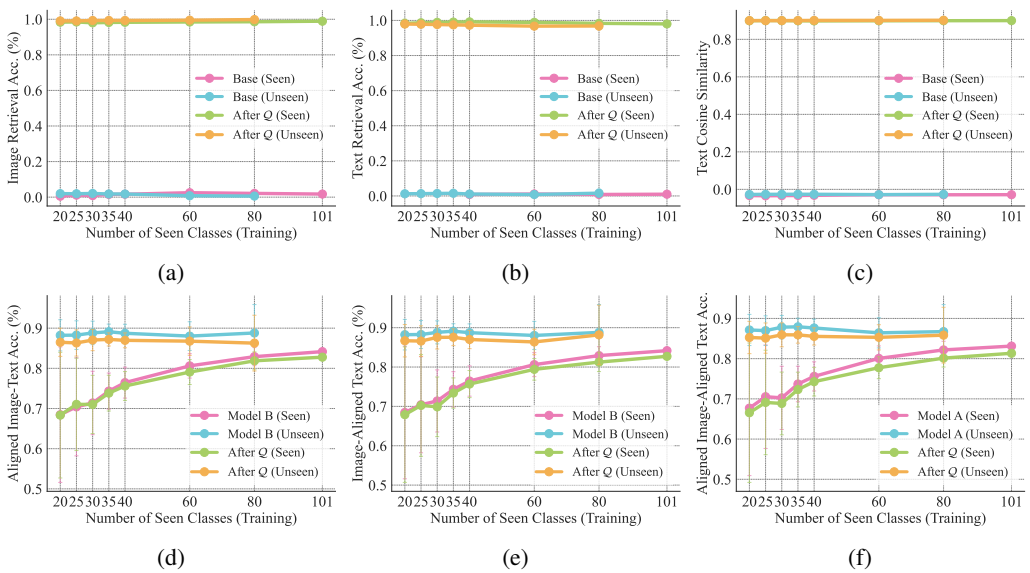

Figure 19: *Generalization of orthogonal alignment under limited supervision (Caltech-101; CLIP ViT-B/32 to CLIP ViT-B/16, OpenAI).* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

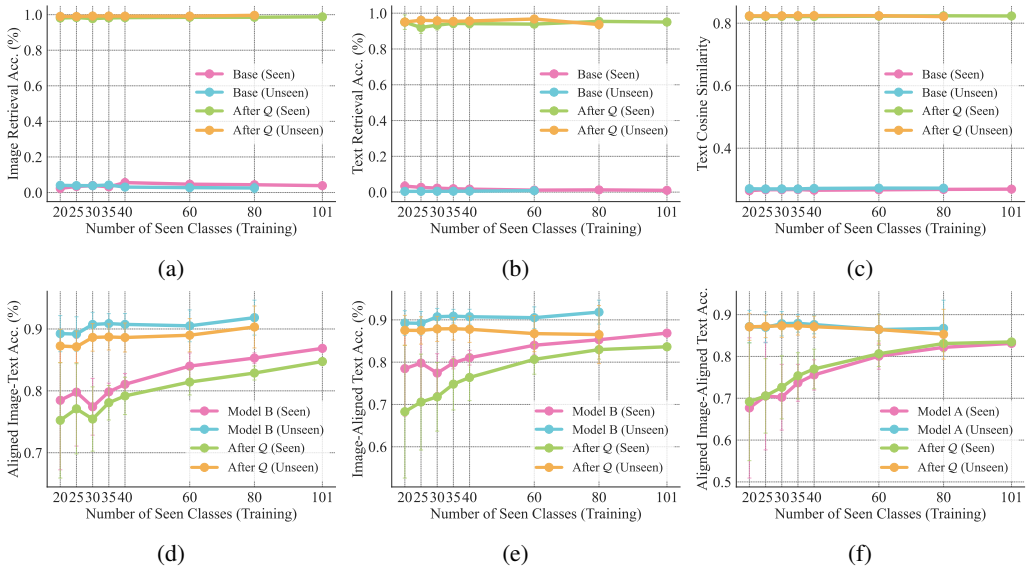

Figure 20: *Generalization of orthogonal alignment under limited supervision (Caltech-101; CLIP ViT-B/32 (OpenAI) to CLIP ViT-B/32 (LAION).* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

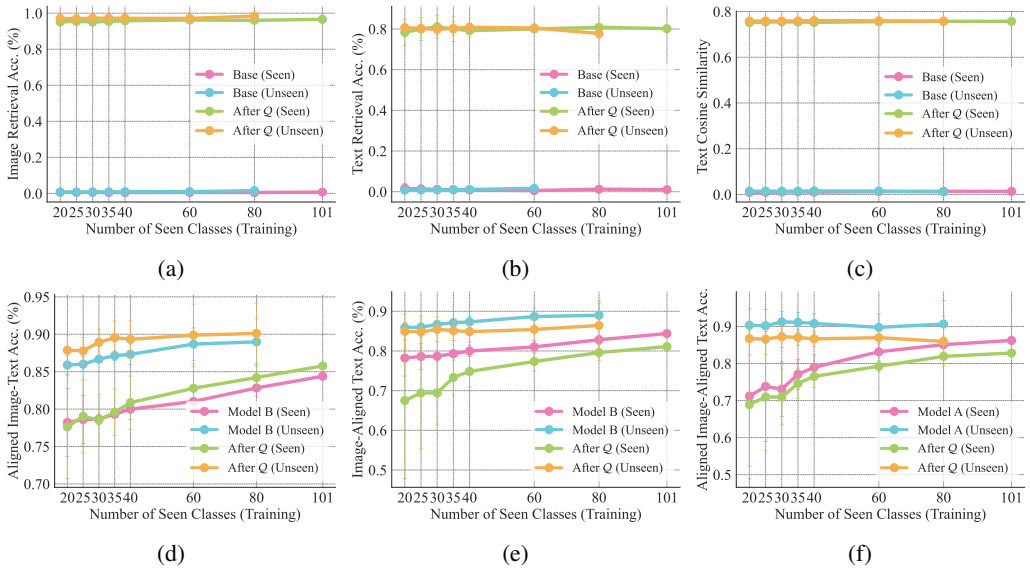

Figure 21: *Generalization of orthogonal alignment under limited supervision (Caltech-101; CLIP ViT-L/14 (OpenAI) to FLAVA* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

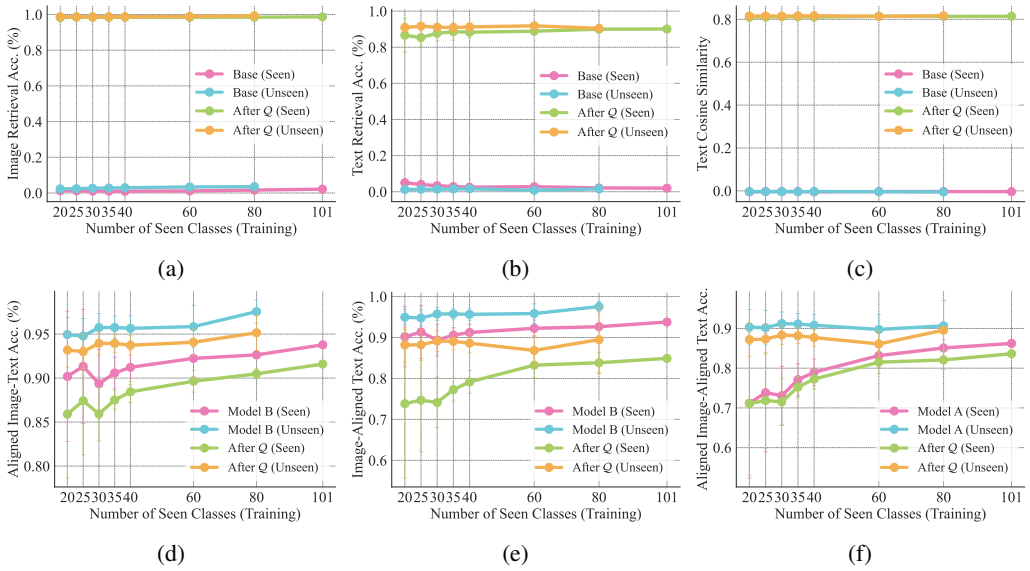

Figure 22: *Generalization of orthogonal alignment under limited supervision (Caltech-101; CLIP ViT-L/14 (OpenAI) to SigLIP.* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

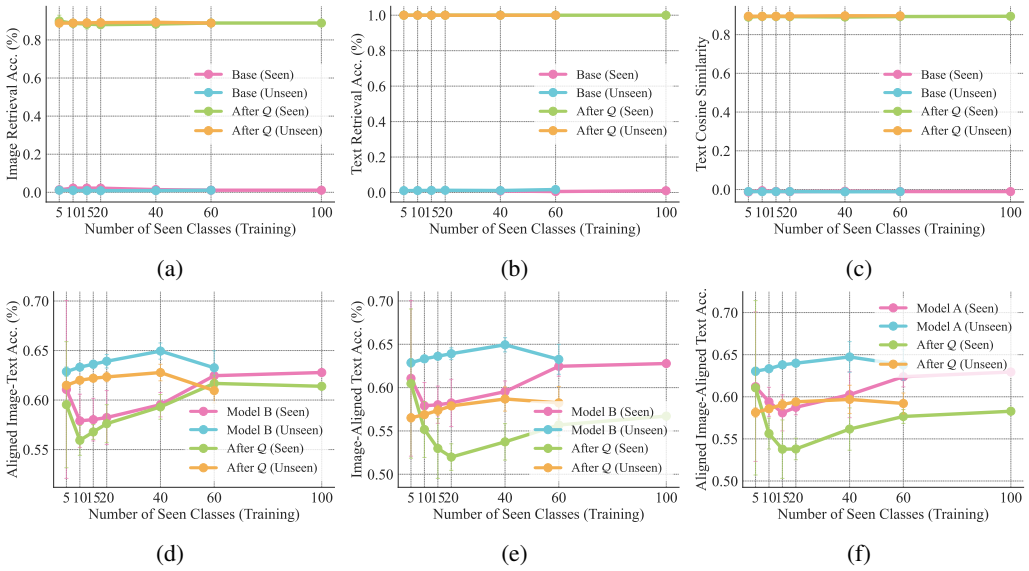

Figure 23: *Generalization of orthogonal alignment under limited supervision (CIFAR-100; CLIP ViT-B/32 to CLIP ViT-B/16, OpenAI).* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

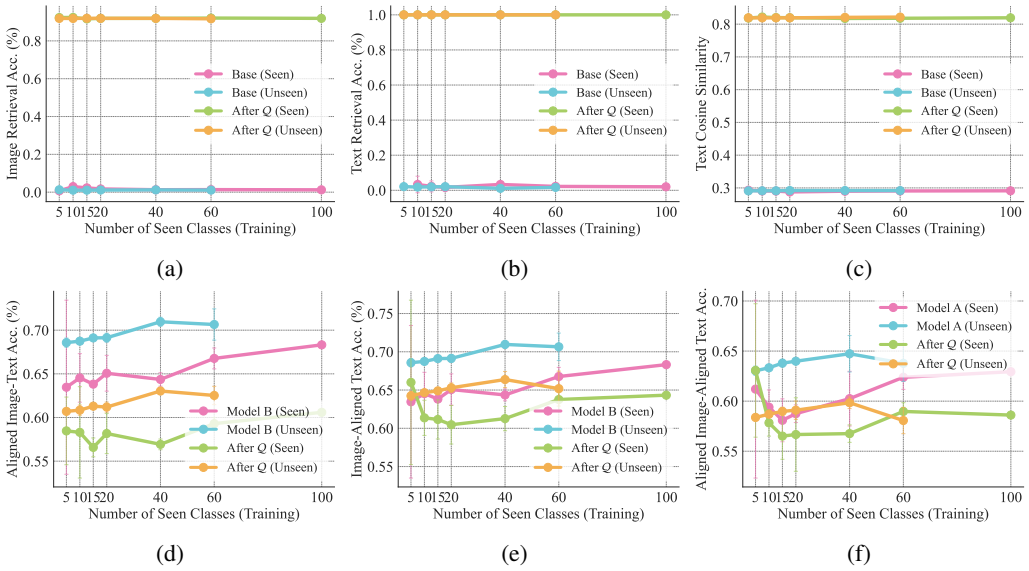

Figure 24: *Generalization of orthogonal alignment under limited supervision (CIFAR100; CLIP ViT-B/32 (OpenAI) to CLIP ViT-B/32 (LAION).* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

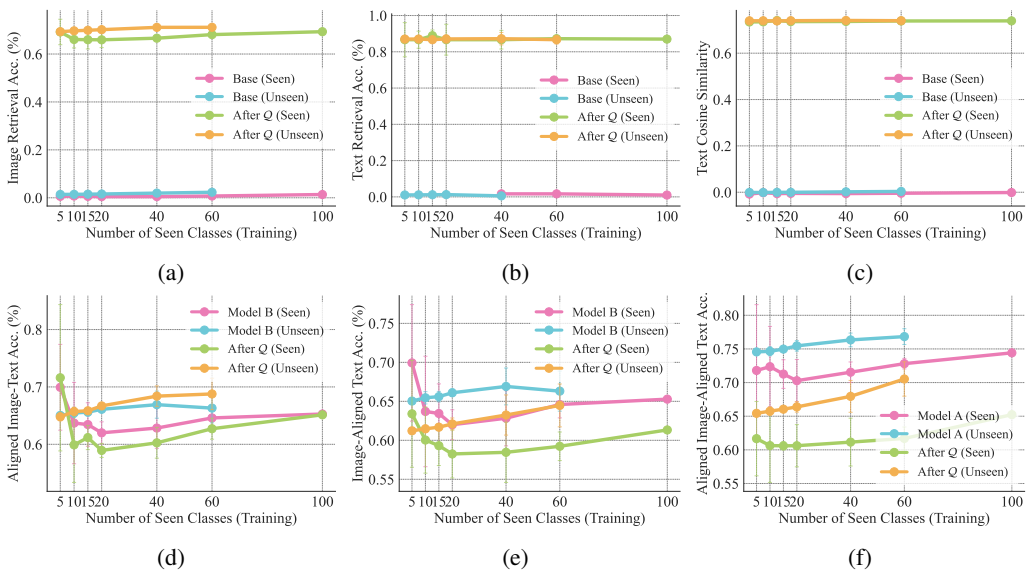

Figure 25: *Generalization of orthogonal alignment under limited supervision (CIFAR100; CLIP ViT-L/14 (OpenAI) to FLAVA* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

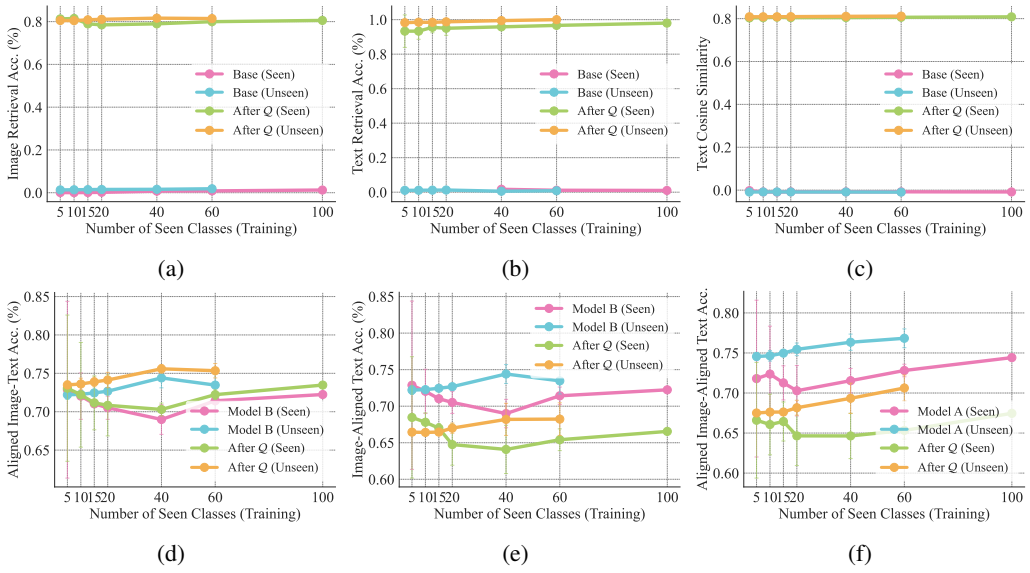

Figure 26: *Generalization of orthogonal alignment under limited supervision (CIFAR100; CLIP ViT-L/14 (OpenAI) to SigLIP.* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

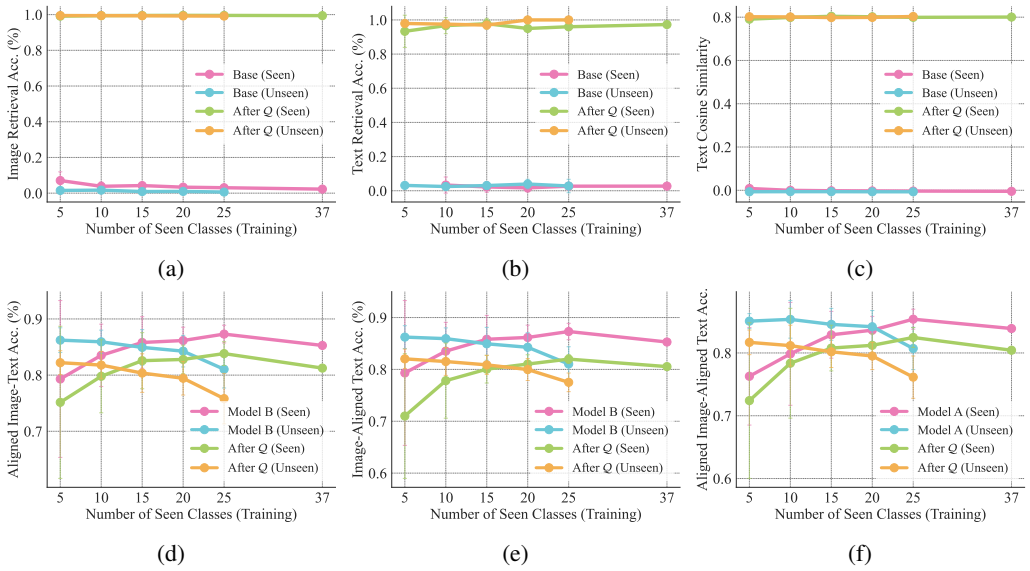

Figure 27: *Generalization of orthogonal alignment under limited supervision (Oxford Pets; CLIP ViT-B/32 to CLIP ViT-B/16, OpenAI).* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

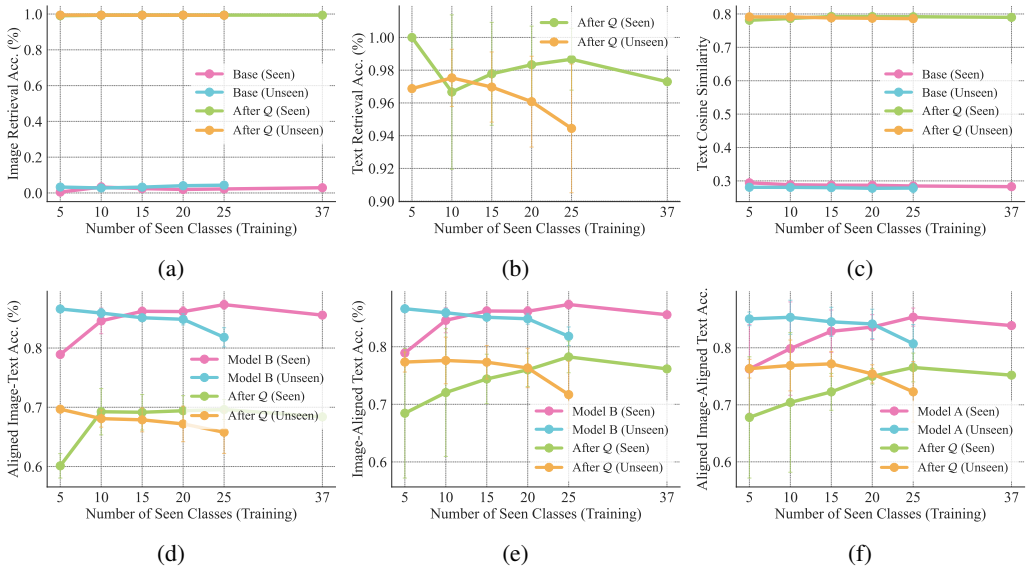

Figure 28: *Generalization of orthogonal alignment under limited supervision (Oxford Pets; CLIP ViT-B/32 (OpenAI) to CLIP ViT-B/32 (LAION).* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

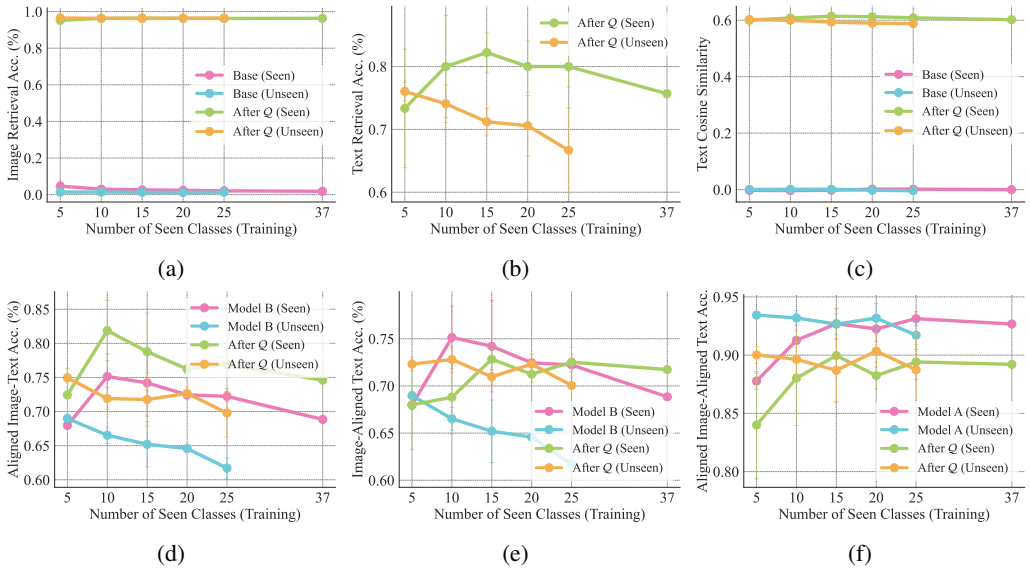

Figure 29: *Generalization of orthogonal alignment under limited supervision (Oxford Pets; CLIP ViT-L/14 (OpenAI) to FLAVA* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

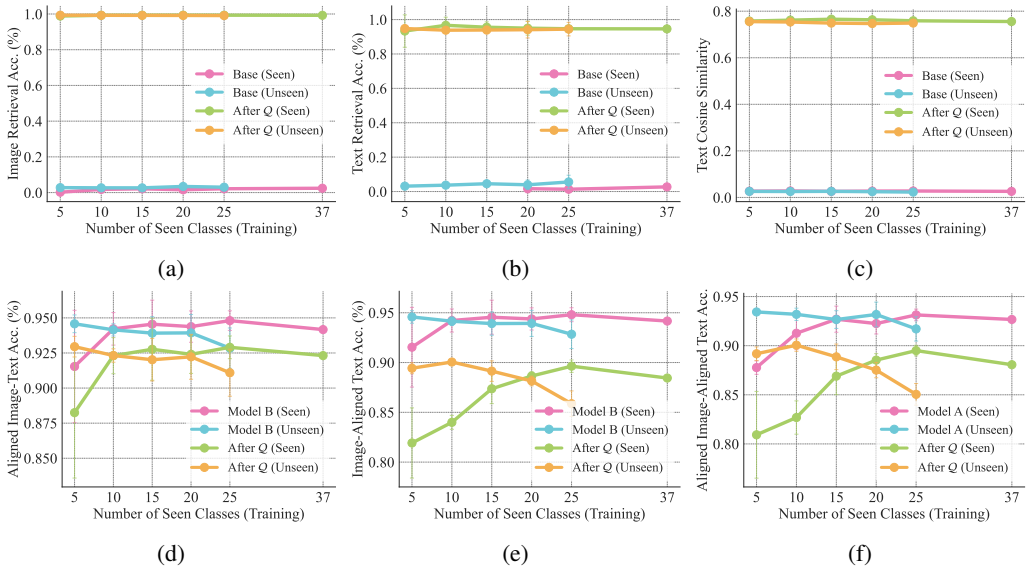

Figure 30: *Generalization of orthogonal alignment under limited supervision (Oxford Pets; CLIP ViT-L/14 (OpenAI) to SigLIP.* (a) Image-image class retrieval, (b) text-text class retrieval, and (c) mean text-text cosine similarity, each reported on seen and unseen classes. (d-f) Downstream transfer: (d) aligned images from model A with text from model B, (e) images from model B with aligned text from model A, and (f) aligned images with aligned text from model A. $\mathcal{R}$ learned from few classes transfers across modalities and generalizes to unseen classes, achieving near-oracle cross-model retrieval and classification.

E.8 Sensitivity To The Choice And Diversity Of Anchors

In Figure 6(a), we randomly sample classes and then randomly sample data points from them, repeating across 5 independent draws. The standard deviation bars in the plot reflect sensitivity to both class choice and sample choice. When the number of anchors is small (e.g., ≤ 10 classes), the variance can be as high as $\pm \approx 5\%$ in performance, but it shrinks rapidly as more classes are added and becomes negligible past 10–15 classes. To further probe this, we report below in Figure 31, an ablation in which anchors are drawn from a single semantic cluster, deliberately violating the Sym(d)-spanning condition. We fix a budget of $N = 93$ anchor images since that is the minimum class size in Oxford Pets and either (a) draw all N from a single class (*one cluster*), or (b) spread them across 22 classes (out of 37 classes) with ≈ 4 samples each (*diverse clusters*). As shown below, diverse anchors dramatically outperform single-cluster anchors on retrieval and cross-modal metrics, confirming that semantic spread is crucial for alignment quality.

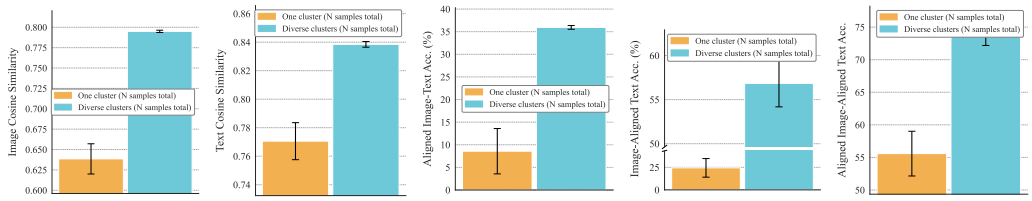

Figure 31: *Cross-model alignment on Oxford-Pets from CLIP ViT-B/32 (OpenAI) to CLIP ViT-B/32 when $\mathcal{Q}$ is fit on (i) N samples from a single class versus (ii) N samples spread across class clusters. Total sample budget is identical in both settings; only semantic diversity differs.*

E.9 The Learned Orthogonal Map Generalizes Broadly

The previous experiment shows that $\mathcal{Q}$ is identifiable from a small fraction of the data within a dataset. We next ask whether the same $\mathcal{Q}$ depends on the downstream distribution used to estimate it. We fit $\mathcal{Q}$ using paired images from one dataset (Oxford Pets) and evaluate on a different dataset (Caltech-101). Results for transfer from CLIP (OpenAI) to CLIP (OpenAI), CLIP (LAION), FLAVA, and SigLIP are shown in Figures 32 to 35 respectively.

Across all figures, the orthogonal map learned on images from one dataset transfers to the text from another dataset, as shown by strong text-text cosine similarity (subplot c). Further, downstream image-text retrieval metrics (subplots d,e,f) remain strong under transfer—often closely matching or even exceeding an in-domain fit—indicating that Q generalizes beyond the calibration dataset.

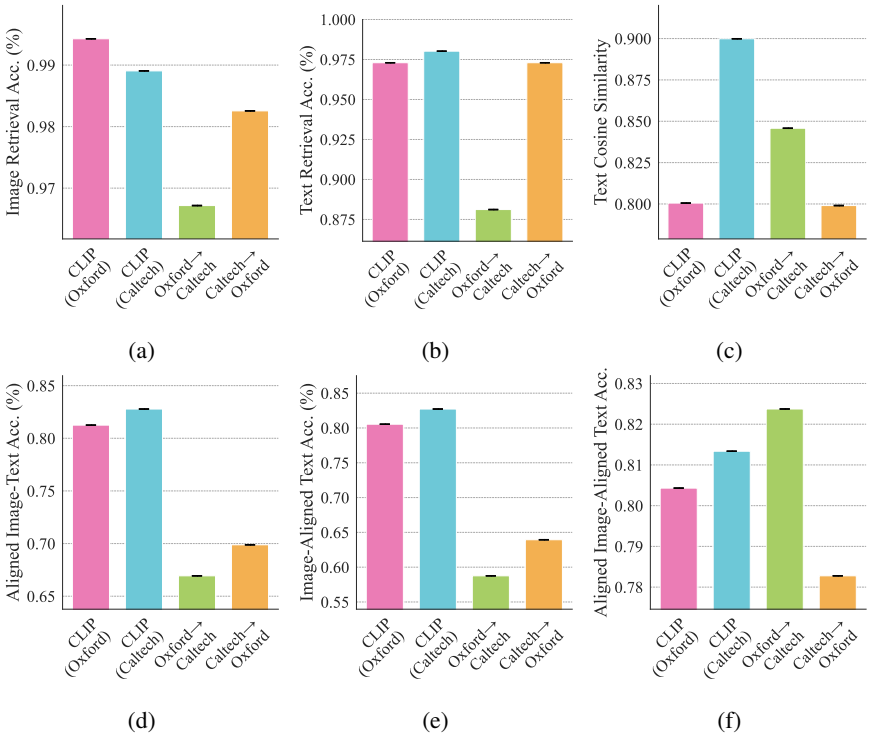

Figure 32: *Transfer of $\mathcal{Q}$ learnt on Oxford Pets to Caltech-101 and vice-versa across CLIP ViT-B/32 and ViT-B/16 (OpenAI).* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A.

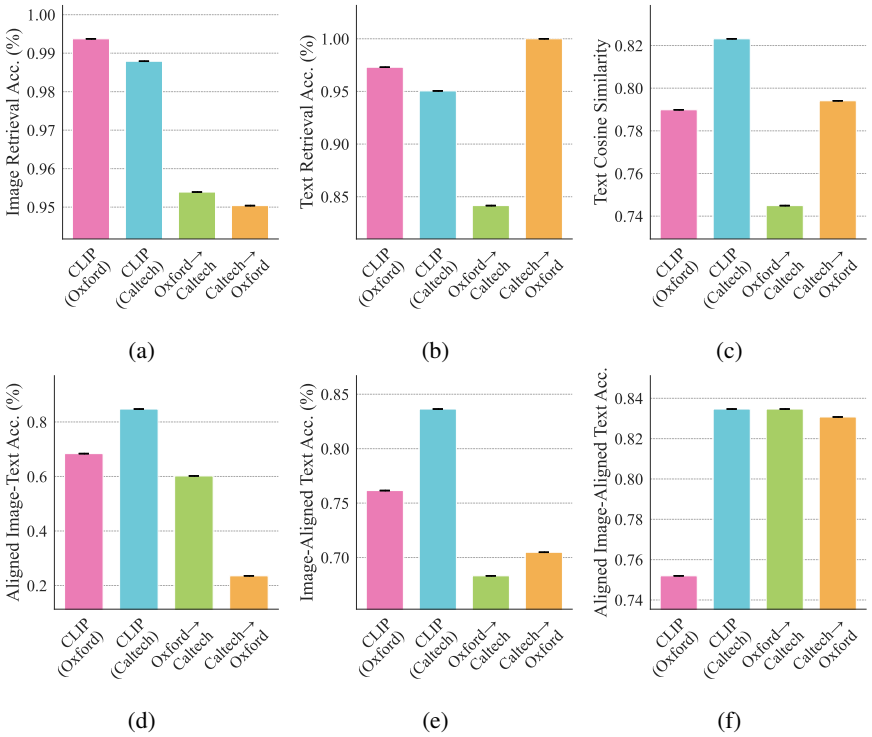

Figure 33: *Transfer of $\mathcal{Q}$ learnt on Oxford Pets to Caltech-101 and vice-versa across CLIP ViT-B/32 (OpenAI) and ViT-B/32 (LAION)* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A.

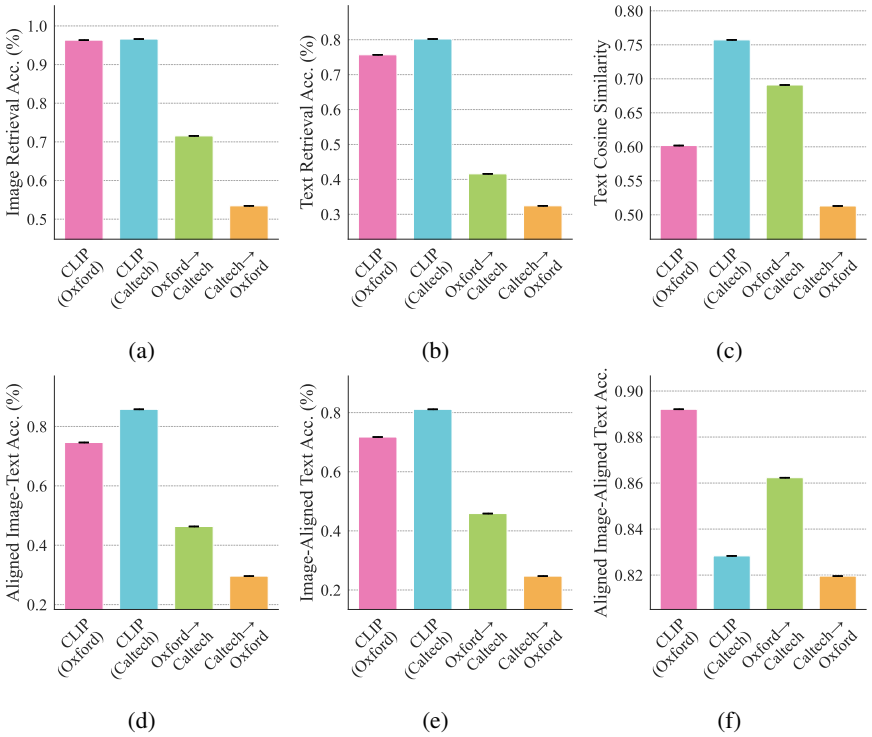

Figure 34: *Transfer of Q learnt on Oxford Pets to Caltech-101 and vice-versa across CLIP ViT-L/14 (OpenAI) and FLAVA* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A.

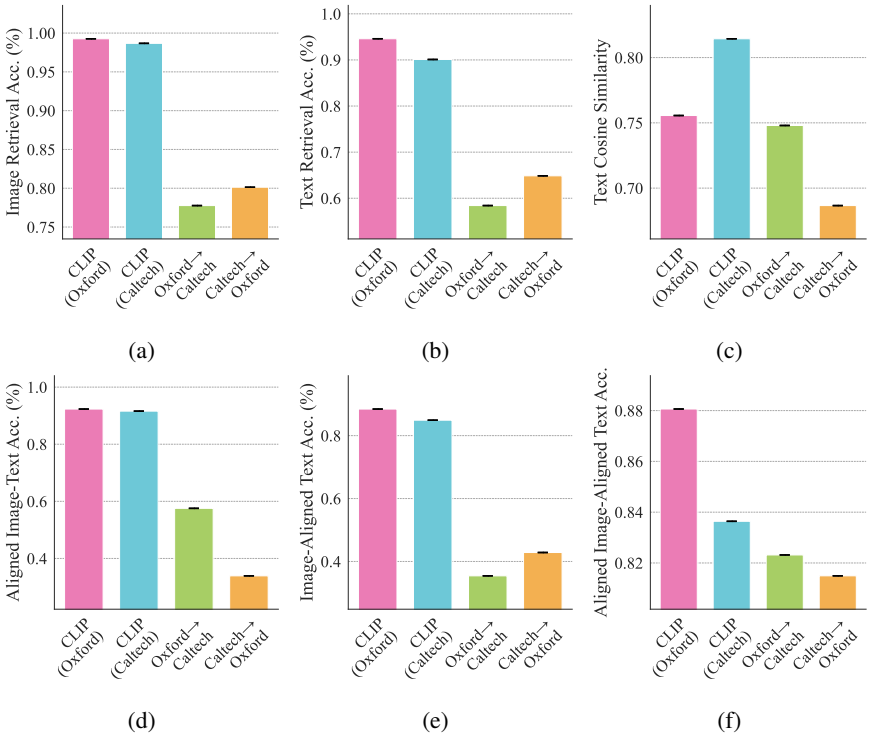

Figure 35: *Transfer of $\mathcal{Q}$ learnt on Oxford Pets to Caltech-101 and vice-versa across CLIP ViT-L/14 (OpenAI) and SigLIP.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A.

E.10 Learning the Orthogonal Map from Text Instead of Images Transfers to Images

So far, we fit $\mathcal{Q}$ using paired images and evaluated whether it transfers to the text modality. We now ask the converse: can we fit $\mathcal{Q}$ using only text and recover the same cross-model transform that governs images? Figures 36 to 38, shows these results for Caltech-101, CIFAR-100 and Oxford Pets, respectively. As observed from these figures, across datasets and model pairs, this text-trained map substantially improves downstream task accuracy after $\mathcal{Q}$ (image-text retrieval), recovering a large fraction of the stronger model's performance. However, compared to image-trained maps, text-trained maps yield weaker visual transfer (image-image retrieval and paired-image cosine), indicating that dense image supervision better constrains the shared orthogonal map than class-level text anchors.

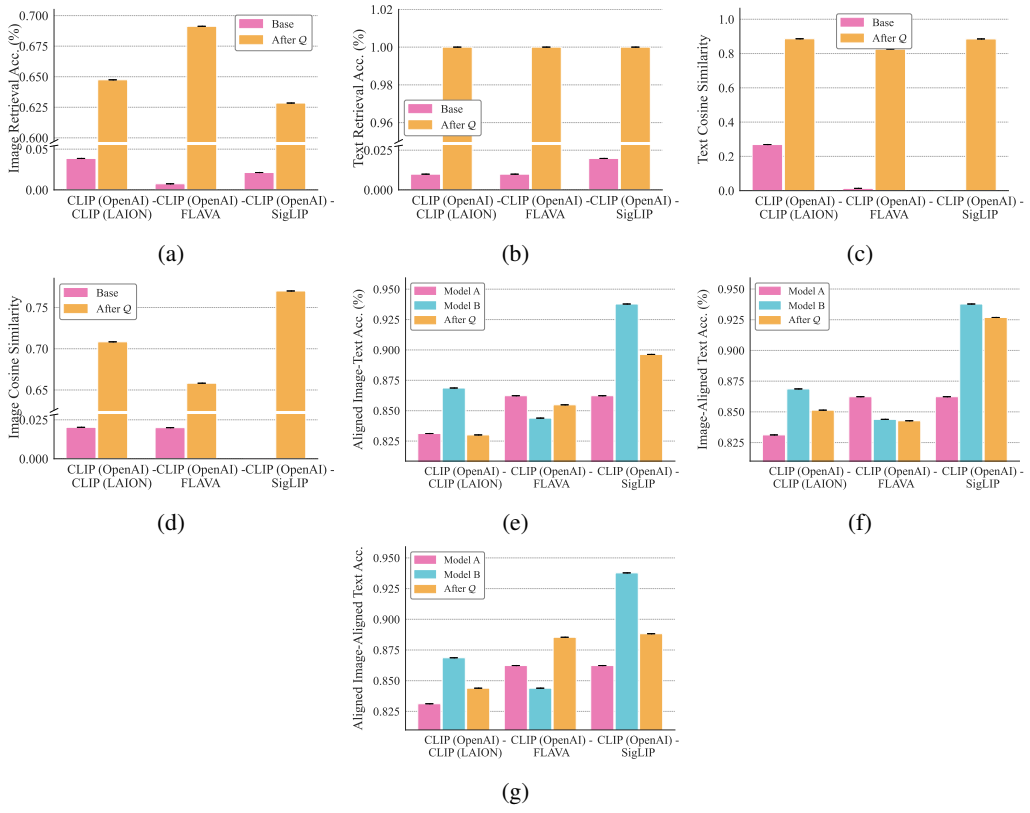

Figure 36: *Cross-model alignment on Caltech-101 learned from text prototypes instead of images, before and after fitting a single orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text-text cosine similarity. (d) Mean image-image cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

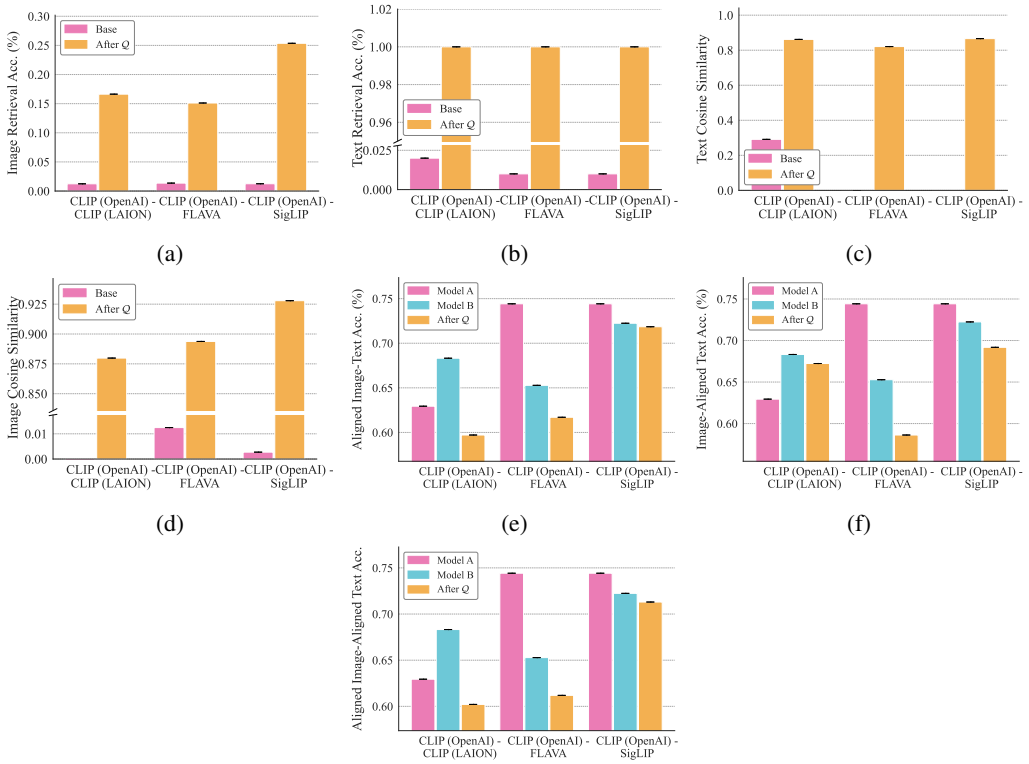

Figure 37: *Cross-model alignment on CIFAR-100 learned from text prototypes alone before and after fitting a single orthogonal map Q.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text-text cosine similarity. (d) Mean image-image cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

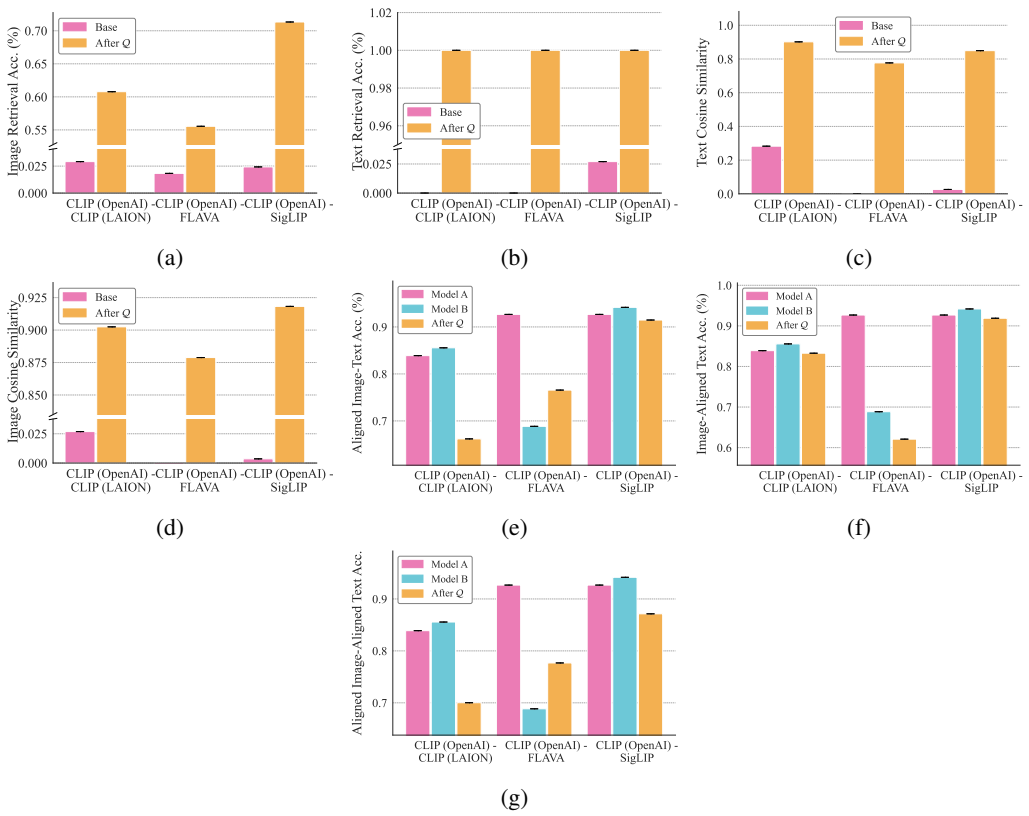

Figure 38: *Cross-model alignment on Oxford Pets learned from text prototypes alone before and after fitting a single orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval (c) Mean text-text cosine similarity. (d) Mean image-image cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

E.11 Cycle Consistency and Consistency Under Composition

Our theory and experiment so far posit that different contrastive models represent the same underlying semantic geometry, differing only by a global orthogonal reparameterization. This would imply two forms of consistency: (1) aligning model (f, g) to $(\tilde{f}, \tilde{g})$ and then $(\tilde{f}, \tilde{g})$ to $(\hat{f}, \hat{g})$ should align (f, g) to $(\hat{f}, \hat{g})$ (composition), and aligning (f, g) to $(\tilde{f}, \tilde{g})$ should be undone by aligning $(\tilde{f}, \tilde{g})$ back to (f, g) (cycle). We therefore test transitivity and cycle consistency to ensure the learned maps reflect a shared geometry rather than pair-specific fitting.

In this section, we present complete results for cycle consistency and composition on Caltech-101 in Figure 39 and Figure 40 respectively, on CIFAR100 in Figure 41 and Figure 42 respectively, and on Oxford Pets in Figure 43 and Figure 44 respectively.

Across all composition figures i.e. Figures 40, 42 and 44, we compare composing two learned isometries, $\mathcal{Q}_{BC} \circ \mathcal{Q}_{AB}$, with directly learning $\mathcal{Q}_{AC}$ between the endpoint models. Across all model triplets, the composed map almost matches the direct map. Strikingly, in some cases, the composed map outperforms the direct map, showing that the intermediate model's stronger geometry propagates through the chain. Thus, isometries learned on disjoint model pairs compose reliably and can align a model pair that has never been calibrated together.

Across all cycle-consistency figures i.e. Figures 39, 41 and 43, we compare the downstream performance of Model A to a round trip that applies the forward map followed by the reverse map (the composition labeled $\mathcal{Q}_{BA} \circ \mathcal{Q}_{AB}$). Since each map is estimated via closed-form Procrustes, $\mathcal{Q}_{BA}$ is the inverse of $\mathcal{Q}_{AB}$ by construction; we therefore fit them on independent 95% random subsets of the training pairs to test under finite-sample estimation. Across all pairs and datasets, image-text classification accuracies (across different aligned pairs) nearly matches the original model, indicating that the learned forward and reverse isometries behave as near-inverses and that the induced coordinate change is stable.

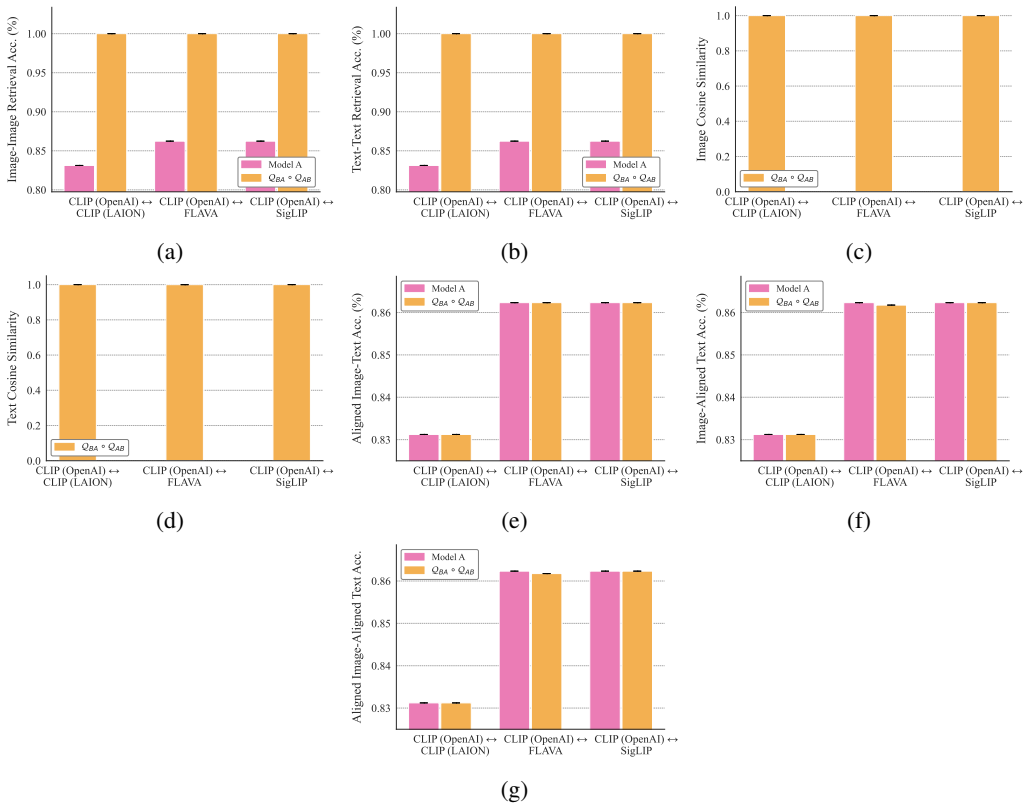

Figure 39: *Cycle consistency on Caltech-101.* (a) Image–image class retrieval. (b) text–text class retrieval (c) Mean image-image cosine similarity. (c) Mean text-text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

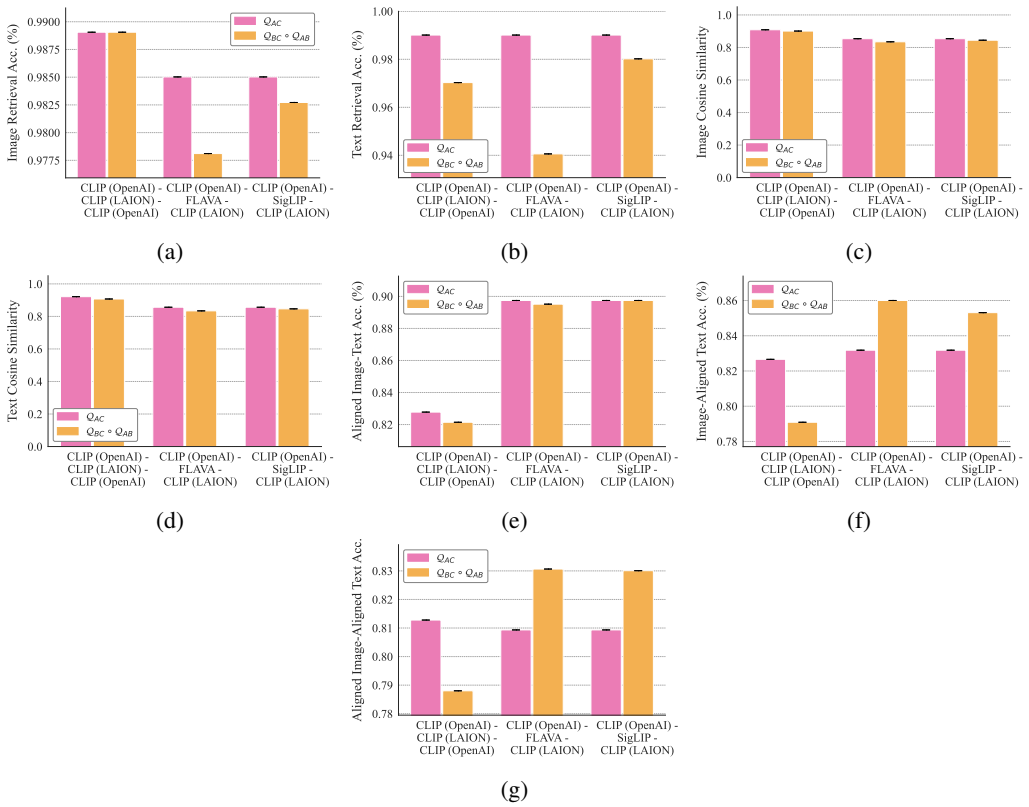

Figure 40: *Compositionality on Caltech-101 by comparing $Q_{BC} \circ Q_{AB}$ with Q_{AC}.* (a) Image–image class retrieval. (b) text–text class retrieval (c) Mean image-image cosine similarity. (c) Mean text-text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

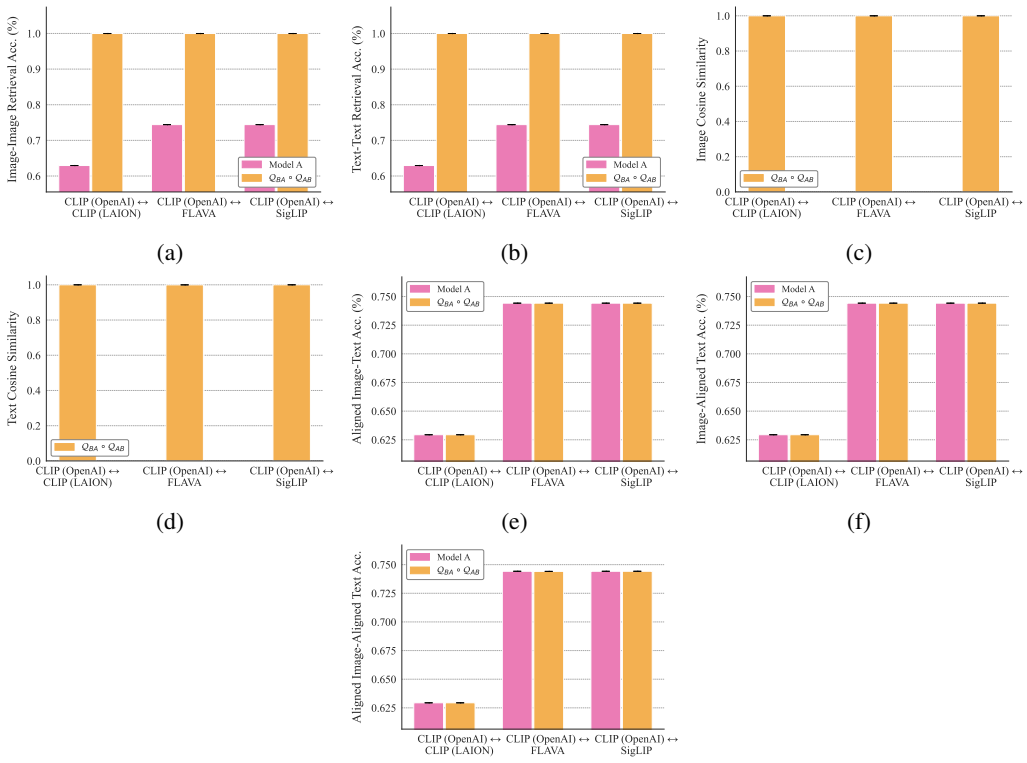

Figure 41: *Cycle consistency on CIFAR-100.* (a) Image–image class retrieval. (b) text–text class retrieval (c) Mean image-image cosine similarity. (c) Mean text-text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

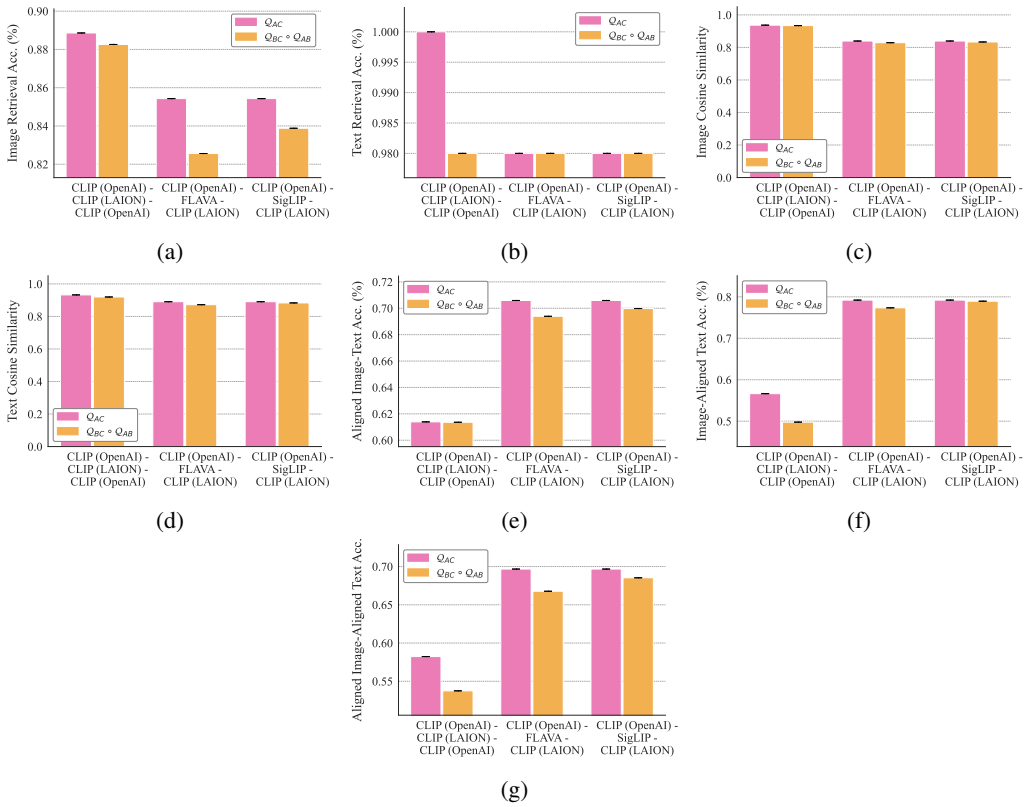

Figure 42: *Compositionality on CIFAR-100 by comparing $Q_{BC} \circ Q_{AB}$ with Q_{AC}.* (a) Image–image class retrieval. (b) text–text class retrieval (c) Mean image-image cosine similarity. (c) Mean text-text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

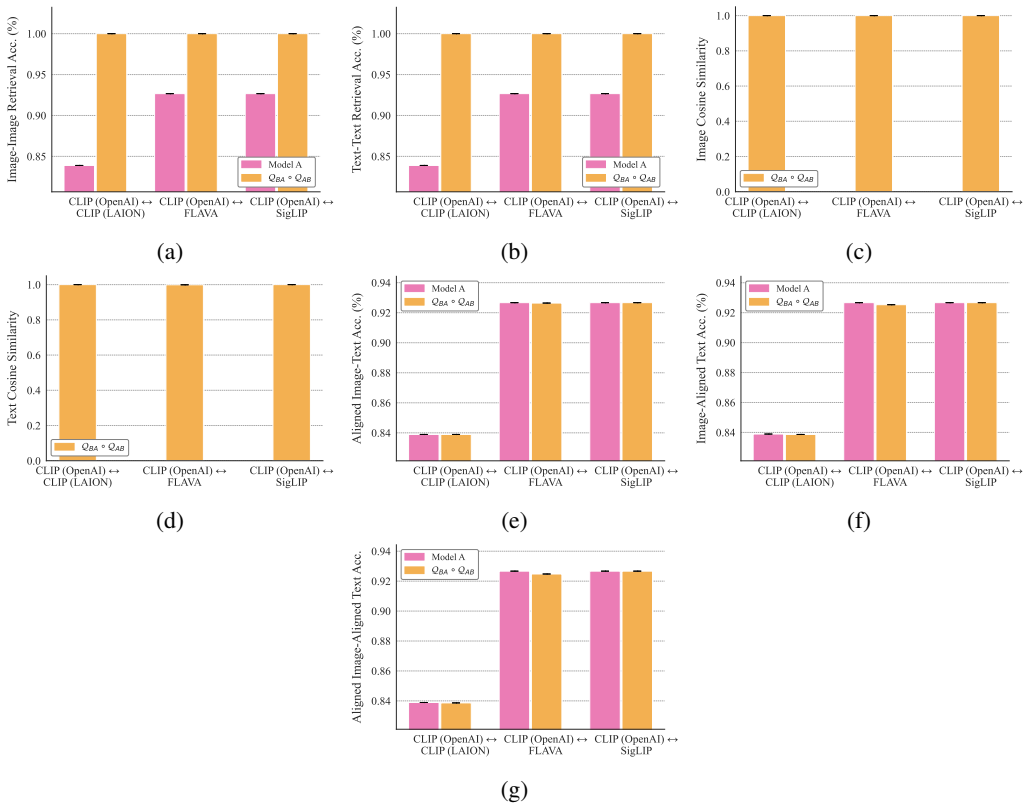

Figure 43: *Cycle consistency on Oxford Pets.* (a) Image–image class retrieval. (b) text–text class retrieval (c) Mean image-image cosine similarity. (c) Mean text-text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

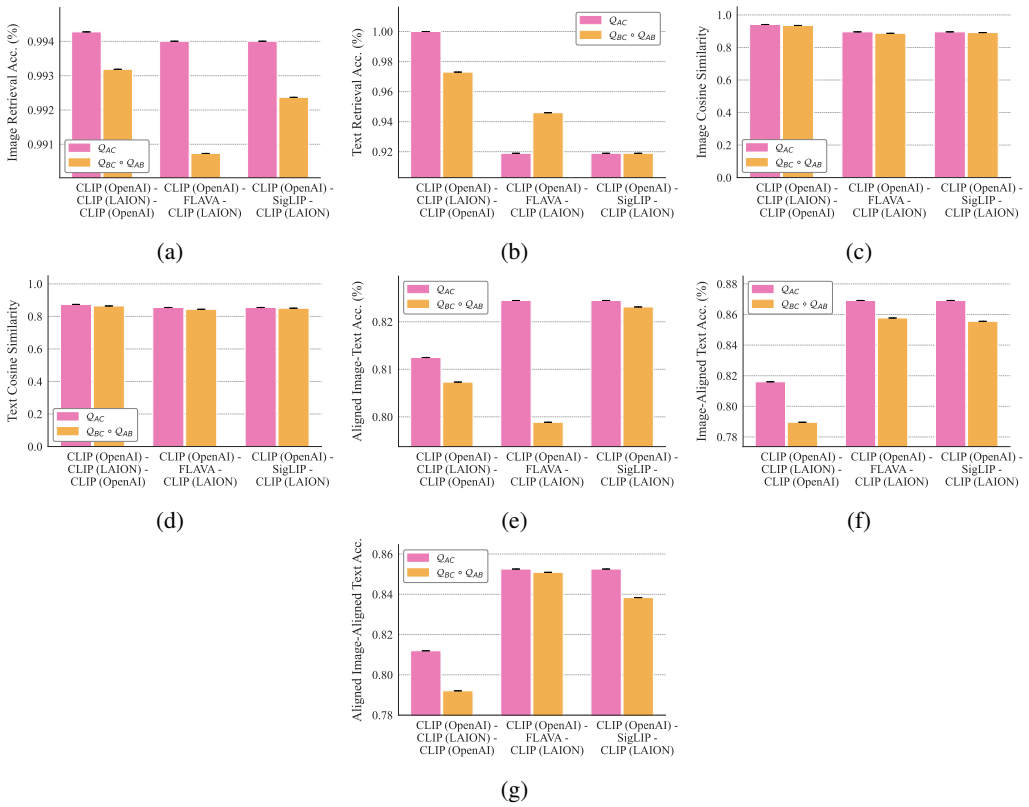

Figure 44: *Compositionality on Oxford Pets by comparing $Q_{BC} \circ Q_{AB}$ with Q_{AC}.* (a) Image–image class retrieval. (b) text–text class retrieval (c) Mean image-image cosine similarity. (c) Mean text-text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A.

E.12 Alignment Across Embedding Dimensions.

All results so far use matched embedding sizes, where an orthogonal map is well-defined. We next align CLIP ViT-B/32 (d=512) with CLIP ViT-L/14 (d=768) using a rectangular $\mathcal{Q}$ projected onto the Stiefel manifold, enforcing $\mathcal{Q}^\top \mathcal{Q} \approx I$. In Figures 45 to 47, we report performance before and after learning $\mathcal{Q}$ on image embeddings, and evaluate how effectively the learned map transfers to texts. Across all figures, image-text accuracy remains high in the aligned space, whether using aligned images with native text, native images with aligned text, or aligning both. Text-text pointwise cosine and retrieval also remain strong, indicating that near-isometric maps preserve task-relevant geometry while enabling reliable transfer to the other modality (text in this case).

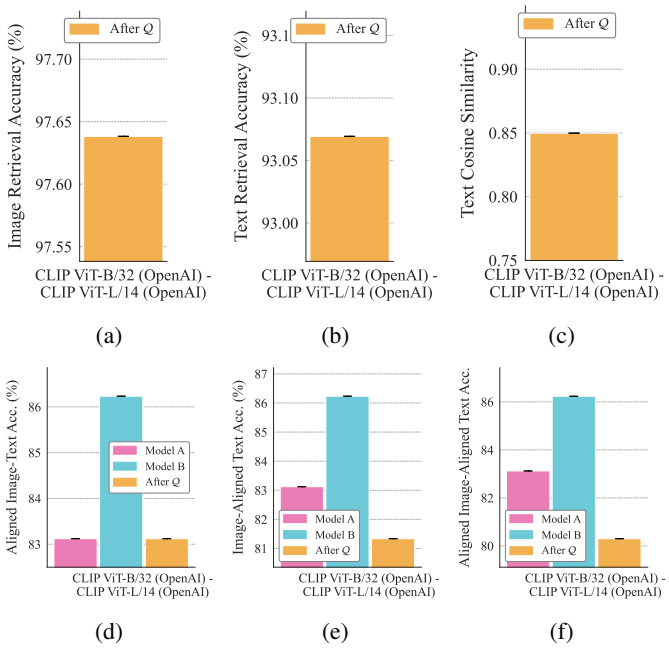

Figure 45: *Cross-model alignment across differing embedding dimensions (Caltech-101; CLIP ViT-B/32 to CLIP ViT-L/14, OpenAI) before and after fitting an orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A. Enforcing semi-orthogonality via Stiefel manifold projection allows the 512-dimensional source manifold to be isometrically embedded into the 768-dimensional target space, maintaining high cross-modal retrieval accuracy.

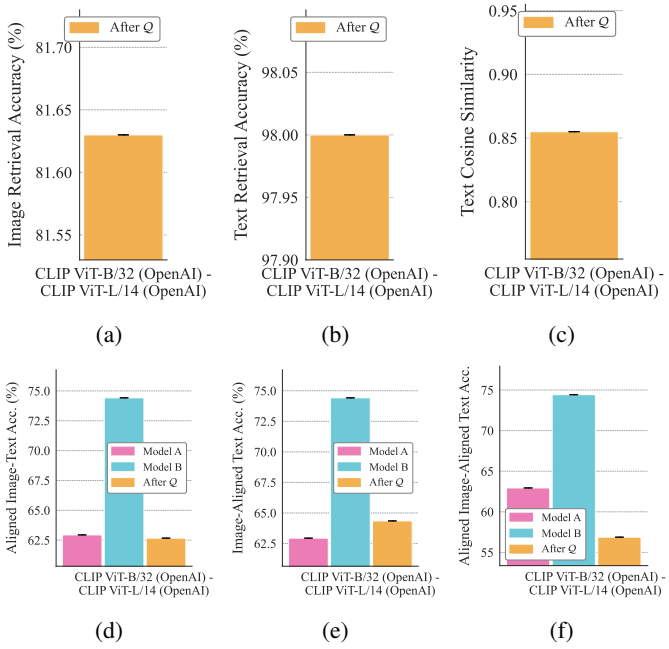

Figure 46: *Cross-model alignment across differing embedding dimensions (CIFAR-100; CLIP ViT-B/32 to CLIP ViT-L/14, OpenAI) before and after fitting an orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A. Enforcing semi-orthogonality via Stiefel manifold projection allows the 512-dimensional source manifold to be isometrically embedded into the 768-dimensional target space, maintaining high cross-modal retrieval accuracy.

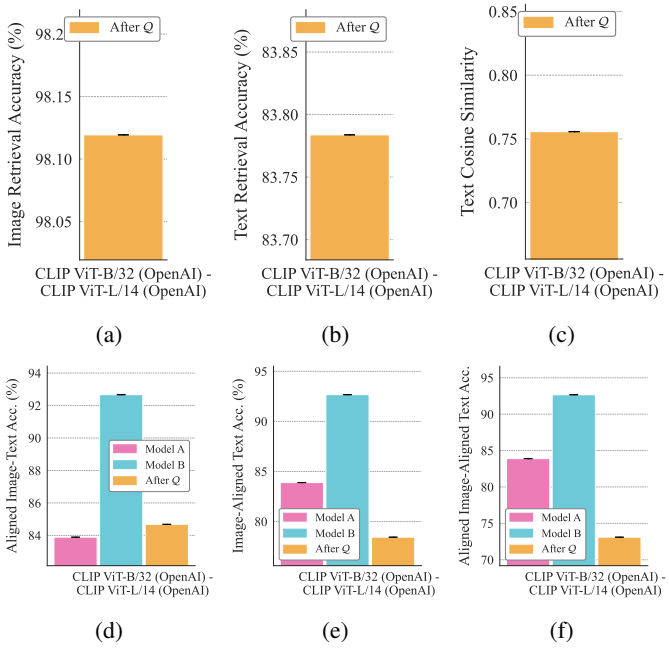

Figure 47: *Cross-model alignment across differing embedding dimensions (Oxford Pets; CLIP ViT-B/32 to CLIP ViT-L/14, OpenAI) before and after fitting an orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A. Enforcing semi-orthogonality via Stiefel manifold projection allows the 512-dimensional source manifold to be isometrically embedded into the 768-dimensional target space, maintaining high cross-modal retrieval accuracy.

E.13 Evaluating Alternative Alignment Maps Than The Orthogonal Mapping

In this section, we ablate the design choices for the alignment function and test three alignment strategies of increasing expressiveness: (i) an orthogonal Procrustes map Q, (ii) an unconstrained linear map, and (iii) a small MLP with a residual connection. All methods are fit on the full training split of each dataset (up to 50k samples on CIFAR-100). The linear and MLP baselines are trained to convergence under a hyperparameter grid search, ensuring their weaker transfer is therefore not due to under-fitting. Results across all model pairs are reported for Caltech-101 (Figure 48), CIFAR-100 (Figure 49), and Oxford Pets (Figure 50). We find that more expressive maps achieve better *pointwise* alignment: both the unconstrained linear map and the MLP yield higher text-text pointwise cosine similarity, reflecting their ability to approximate small distortions and scaling. However, this flexibility comes at the cost of geometric fidelity. The orthogonal map consistently outperforms the alternatives on geometry-sensitive metrics—most notably image to aligned text and aligned-image to aligned-text classification accuracy—which are the metrics that matter for downstream use. By contrast, text-text and image-image retrieval are comparable across all three approaches, with no clear winner. These results indicate that while flexible maps can improve local fit, preserving global geometry through orthogonality can be critical for reliable cross-modal transfer.

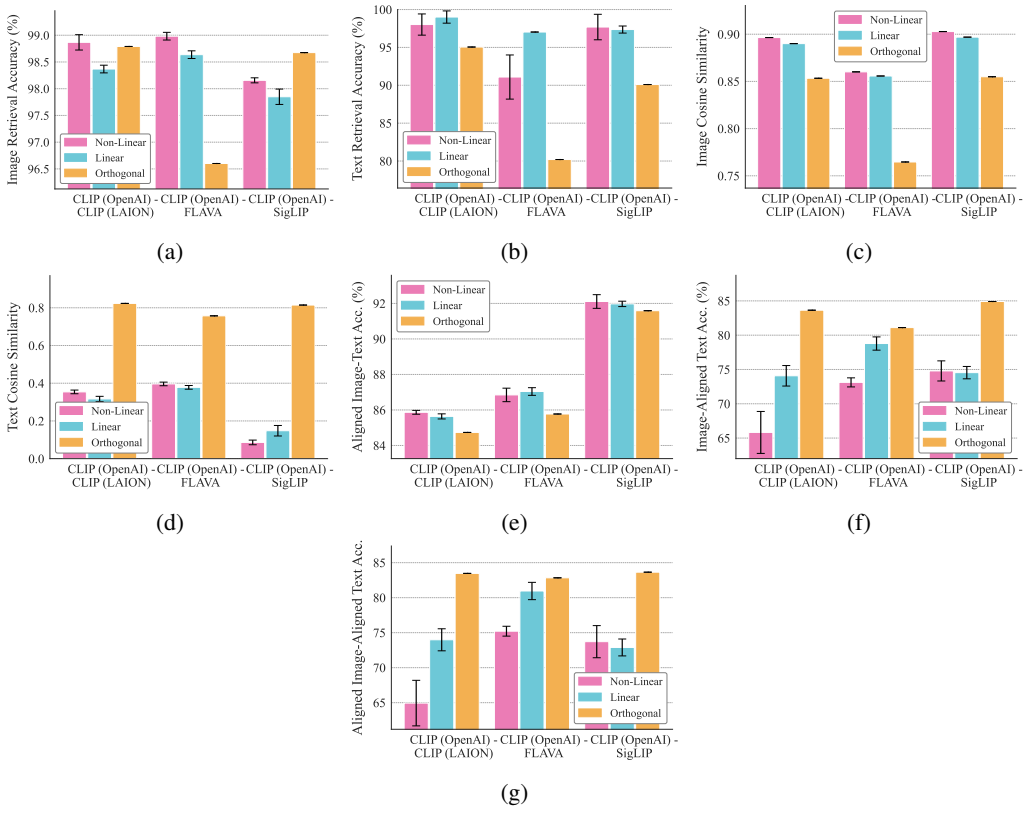

Figure 48: *Comparison of alignment strategies (Non-Linear, Linear, and Orthogonal) on Caltech-101 across model pairs.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A. Unlike linear and non-linear maps, orthogonal transformations preserve the underlying image-text geometry favorable for downstream tasks and transfers across modalities (text in this case).

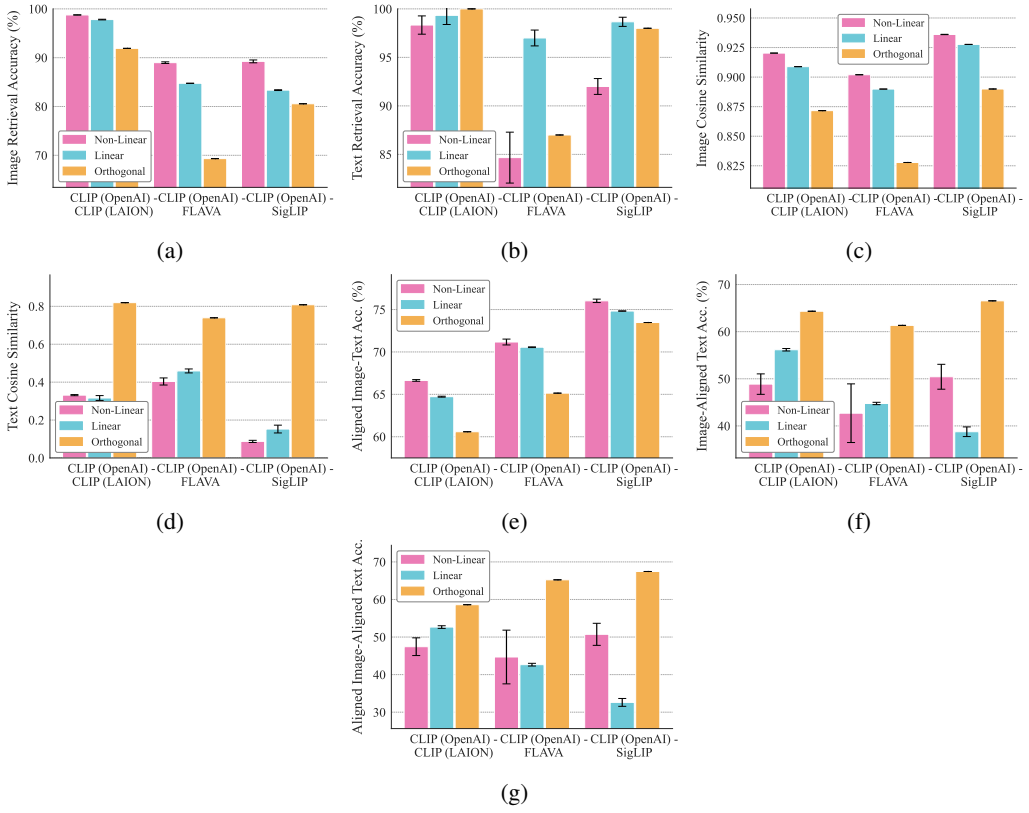

Figure 49: *Comparison of alignment strategies (Non-Linear, Linear, and Orthogonal) on CIFAR-100 across model pairs.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A. Unlike linear and non-linear maps, orthogonal transformations preserve the underlying image-text geometry favorable for downstream tasks and transfers across modalities (text in this case).

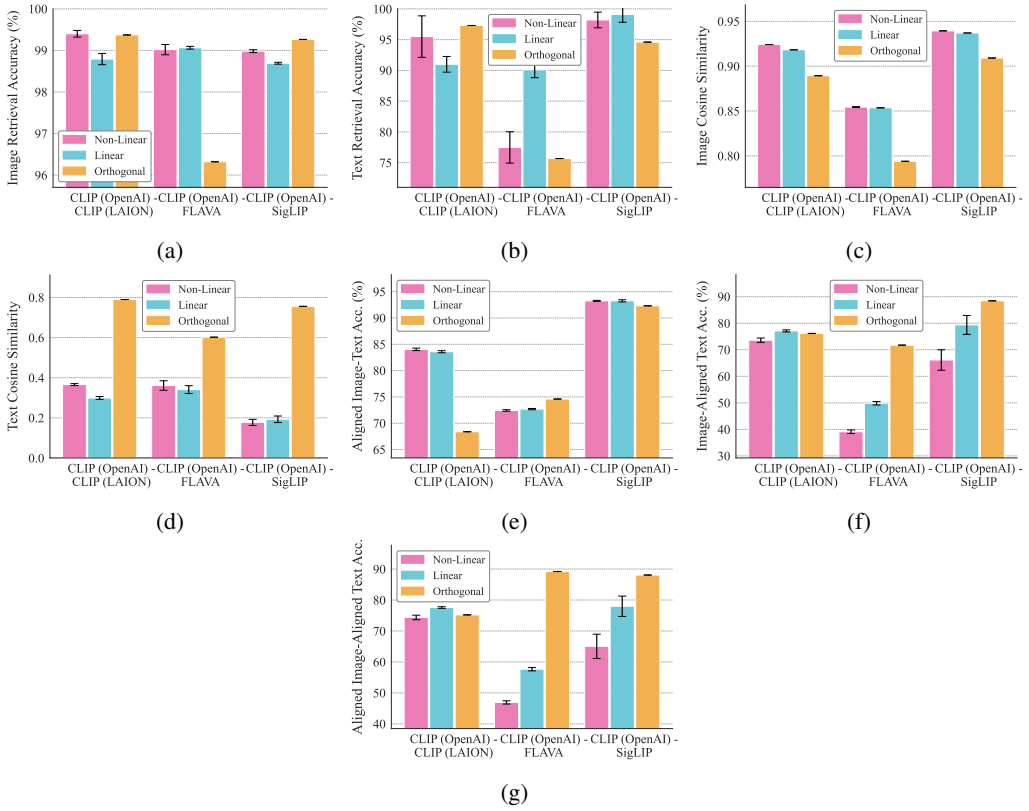

Figure 50: *Comparison of alignment strategies (Non-Linear, Linear, and Orthogonal) on Oxford Pets across model pairs (CLIP OpenAI vs. LAION, FLAVA, and SigLIP).* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean image–image cosine similarity. (d) Mean text–text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A. Unlike linear and non-linear maps, orthogonal transformations preserve the underlying image-text geometry favorable for downstream tasks and transfers across modalities (text in this case).

E.14 Orthogonal Alignment With and Without Centering

All results thus far use our *centered* orthogonal Procrustes map (i.e., we fit Q on mean-centered embeddings and re-center at deployment). Here, we ablate this centering step by comparing it to a pure orthogonal map applied directly to raw embeddings. Specifically, we evaluate: (i) *Orthogonal (no centering)*: $z \mapsto Qz$, a pure change of coordinates that preserves inner products exactly; and (ii) *Orthogonal (with centering)*: $z \mapsto Q(z - \mu_A^{(\cdot)}) + \mu_B^{(\cdot)}$, where $\mu_A^{(\cdot)}$ and $\mu_B^{(\cdot)}$ are modality-specific means of the source and target models (image or text, as appropriate). Throughout, we use "Q" to denote the centered variant, consistent with the rest of the paper.

Results on Caltech-101, CIFAR-100 and Oxford Pets are presented in Figures 51 to 53 respectively. Across all figures, the orthogonal map with mean centering (denoted by Q) improves downstream metrics such as aligned-image-to-text and image-to-aligned-text retrieval, suggesting that centering helps correct first-order inter-model offsets in these comparisons. In contrast, aligned-image-to-aligned-text accuracy is typically higher without centering, since orthogonal w/o centering exactly preserves model A's geometry. That said, these gains are modest, and the overall retrieval performance remains largely similar with or without centering. Finally, pointwise image-image and text-text cosine similarities increase significantly under centering, as subtracting and re-adding the global centroid shift removes a global centroid shift that otherwise depresses cosine agreement.

To separate the effect of mean correction from the orthogonal map, we additionally compare against a mean-shift-only (μ-only) baseline. We also include an oracle map $Q_{\text{img+txt}}$, fitted using both image and text anchors, to test whether supervising both modalities improves over fitting Q on images alone. Here, f, g are the image and text encoders of the source model, while $\tilde{f}, \tilde{g}$ are those of the target model; $\langle f, g \rangle$ denotes the cross-modal accuracy obtained by pairing the two encoders. The columns c_I^+, c_I^- and c_T^+, c_T^- report matched and mismatched image and text similarities, while z_I, z_T measure their normalized separation, as defined in Section E.5. As shown in Table 6, mean shift raises matched and mismatched similarities by nearly the same amount ($z_I = 0.02$, $z_T = -0.02$) and fails to recover cross-modal performance. In contrast, Q yields strong separation ($z_I = 4.52$, $z_T = 4.13$) and near-native performance. The joint oracle provides no consistent gain despite using additional text supervision, supporting our claim that anchors from one modality are sufficient to recover the shared cross-modal map.

Table 6: Comparison of orthogonal alignment with and without mean correction, a mean-shift-only baseline, and a joint image-text oracle on Caltech-101 for CLIP ViT-B/32 (OpenAI) and CLIP ViT-B/32 (LAION-400M).

Method	$\langle f, g \rangle$	$\langle \tilde{f}, \tilde{g} \rangle$	$\langle Qf, \tilde{g} \rangle$	$\langle \tilde{f}, Qg \rangle$	$\langle Qf, Qg \rangle$	c_I^+	c_I^-	c_T^+	c_T^-	z_I	z_T
Raw	82.4	87.0	–	–	–	0.015	0.015	0.269	0.270	0.00	-0.04
Q (w/o μ)	82.4	87.0	80.2	77.6	82.4	0.620	0.228	0.287	0.119	4.50	5.71
μ-only	82.4	87.0	0.5	1.2	77.2	0.423	0.422	0.590	0.591	0.02	-0.02
Q (w/ μ)	82.4	87.0	84.9	82.6	82.2	0.853	0.421	0.822	0.640	4.52	4.13
$Q_{\text{img+txt}}$ (w/ μ)	82.4	87.0	85.5	78.8	77.2	0.813	0.421	0.834	0.643	4.40	3.82

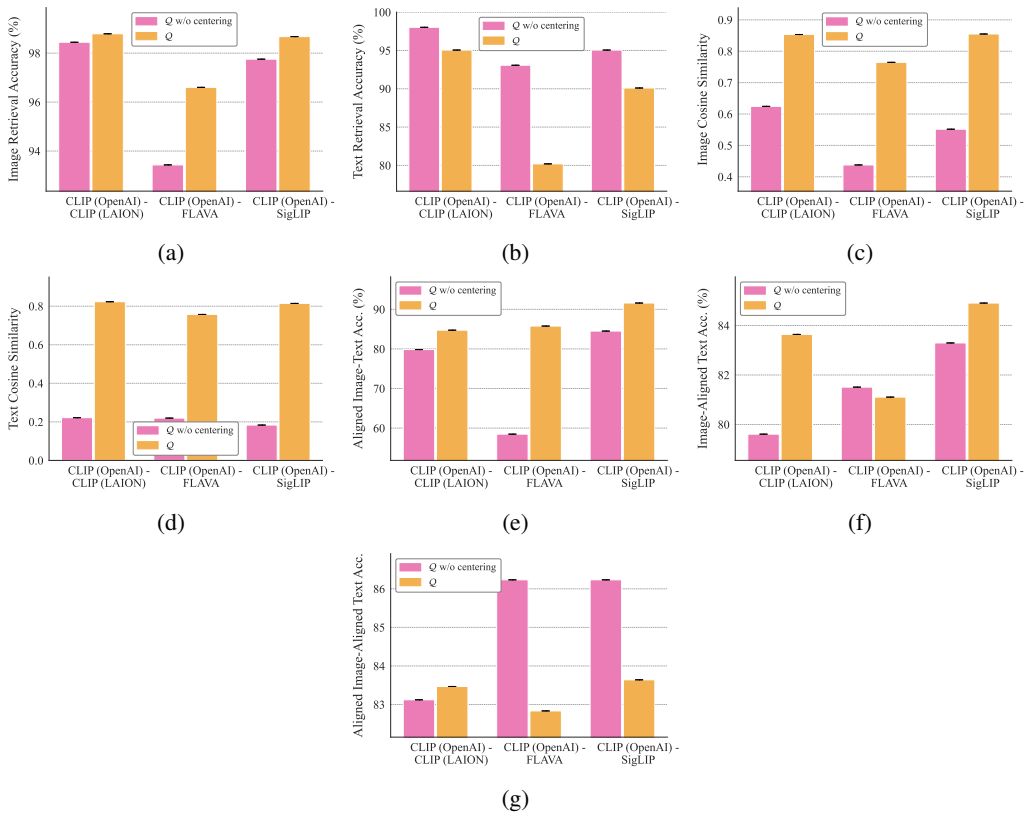

Figure 51: *Comparison between an orthogonal map with and without mean adjustment across model pairs on Caltech-101.* (a) Image–image class retrieval (b) text–text class retrieval (c) Mean image–image cosine similarity (d) Mean text–text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A. Mean adjustment boosts pointwise image-image (c) and text-text (d) cosine and downstream performance (e,f), but can slightly underperform on aligned-image-to-aligned-text and text-text retrieval (b,g).

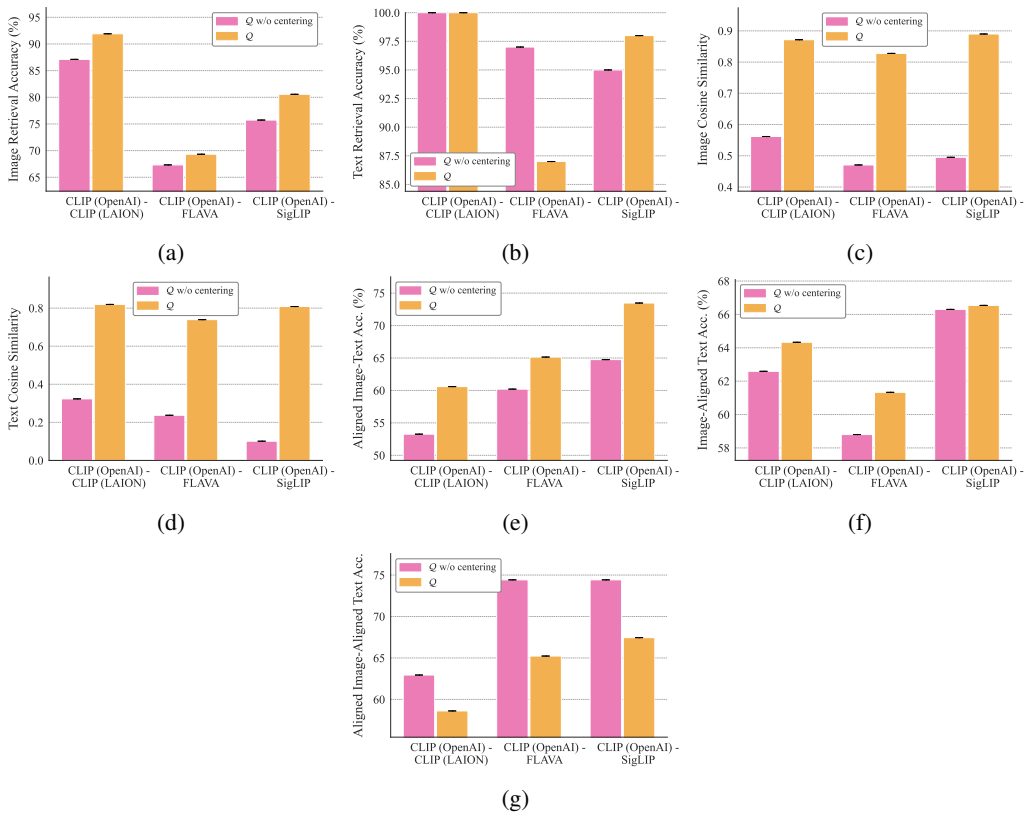

Figure 52: *Comparison between an orthogonal map with and without mean adjustment across model pairs on CIFAR-100.* (a) Image–image class retrieval (b) text–text class retrieval (c) Mean image–image cosine similarity (d) Mean text–text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A. Mean adjustment boosts pointwise image-image (c) and text-text (d) cosine and downstream performance (e,f), but can slightly underperform on aligned-image-to-aligned-text and text-text retrieval (b,g).

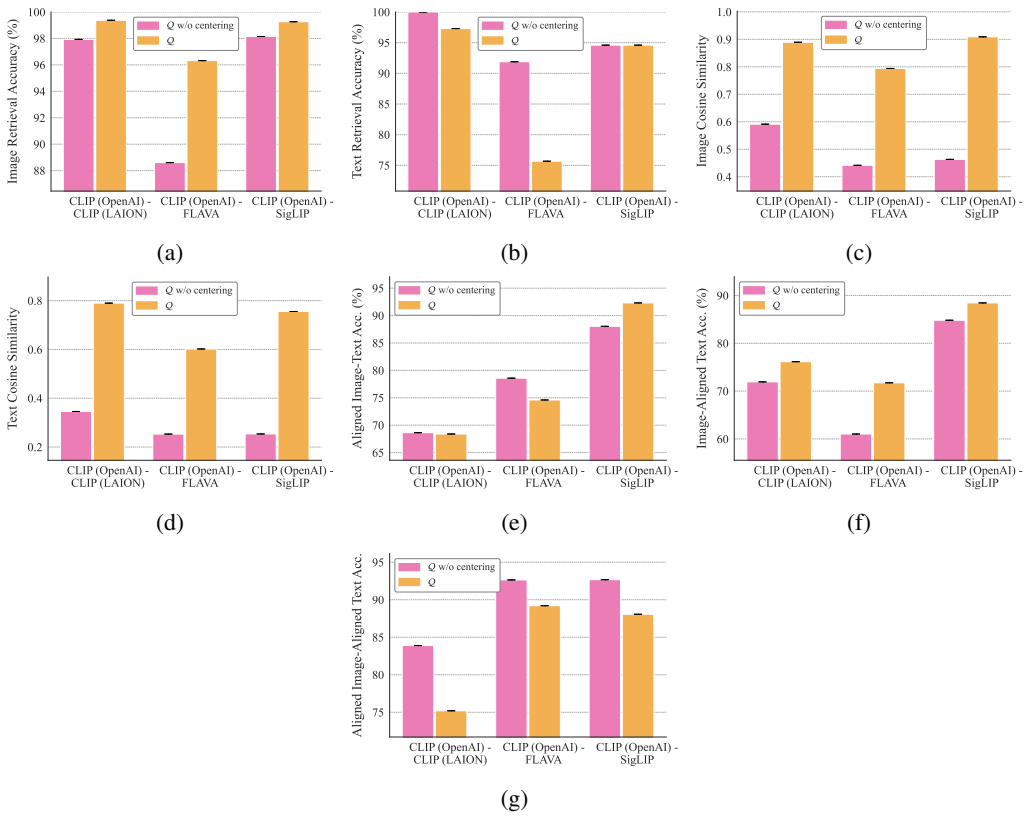

Figure 53: *Comparison between an orthogonal map with and without mean adjustment across model pairs on Oxford Pets.* (a) Image–image class retrieval (b) text–text class retrieval (c) Mean image–image cosine similarity (d) Mean text–text cosine similarity. (e) Image–text retrieval using aligned images from model A and text from model B. (f) Image–text retrieval using images from model B and aligned text from model A. (g) Image–text retrieval using aligned images and aligned text from model A. Mean adjustment boosts pointwise image-image (c) and text-text (d) cosine and downstream performance (e,f), but can slightly underperform on aligned-image-to-aligned-text and text-text retrieval (b,g).

E.15 Isolating Architecture Effects Under Fixed Training Data

All previous experiments compare model pairs that differ in both training distribution and design choices. Here we isolate the role of the data by aligning two CLIP models trained on (probably) the same distribution (dataset used by OpenAI (Radford et al., 2021)) but with different architectures (ViT-B/32 vs. ViT-B/16). Results on Caltech-101, CIFAR-100 and Oxford Pets are reported in Figures 54 to 56. As observed across these figures, in this controlled setting, transfer is consistently stronger, compared to when using models trained on different distributions. This aligns with our theoretical results showing that when two contrastive models optimize the same underlying P_{XY}, their induced multimodal kernels agree up to an additive constant at the population optimum Corollary C.6, tightening the conditions under which a single orthogonal map aligns both modalities.

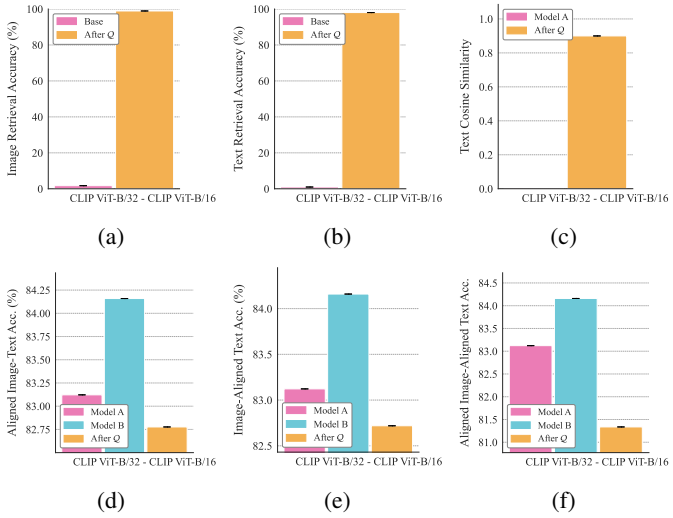

Figure 54: *Cross-model alignment between identical training datasets with differing architectures (Caltech-101; CLIP ViT-B/32 to ViT-B/16, OpenAI) before and after fitting a single orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A. When both models are trained on the same data distribution and differ only in architecture, applying $\mathcal{Q}$ yields substantially stronger modality transfer than in settings where the training distributions differ.

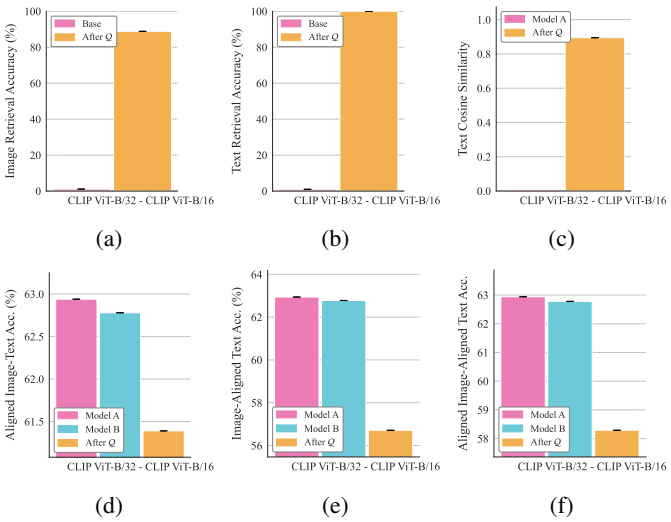

Figure 55: *Cross-model alignment between identical training datasets with differing architectures (CIFAR-100; CLIP ViT-B/32 to ViT-B/16, OpenAI) before and after fitting a single orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A. When both models are trained on the same data distribution and differ only in architecture, applying $\mathcal{Q}$ yields substantially stronger modality transfer than in settings where the training distributions differ.

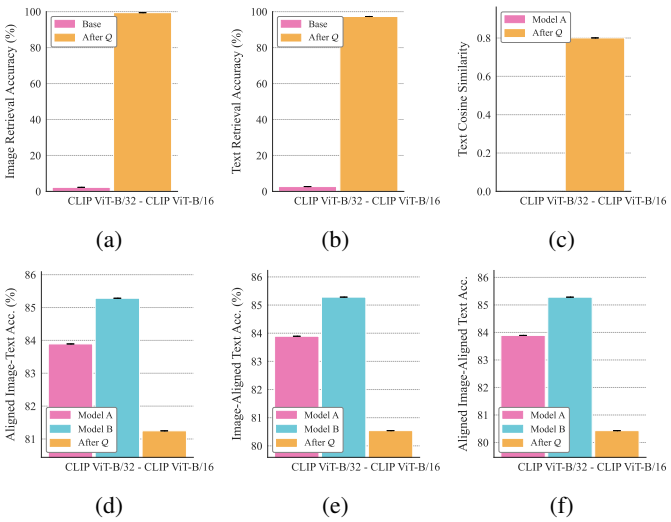

Figure 56: *Cross-model alignment between identical training datasets with differing architectures (Oxford Pets; CLIP ViT-B/32 to ViT-B/16, OpenAI) before and after fitting a single orthogonal map $\mathcal{Q}$.* (a) Image–image class retrieval and (b) text–text class retrieval. (c) Mean text–text cosine similarity. (d) Image–text retrieval using aligned images from model A and text from model B. (e) Image–text retrieval using images from model B and aligned text from model A. (f) Image–text retrieval using aligned images and aligned text from model A. When both models are trained on the same data distribution and differ only in architecture, applying $\mathcal{Q}$ yields substantially stronger modality transfer than in settings where the training distributions differ.

E.16 Qualitative Evidence of Commutativity of Image-Text Alignment Paths

In this section, we qualitatively test whether the learned map $\mathcal{Q}$ induces a consistent, modality-invariant geometry by comparing two alignment paths from a source image x. In the *direct* path, we map the image embedding via $\mathcal{Q}$ and retrieve its top-5 nearest images in the target model's space. In the *text-mediated* path, we first retrieve the nearest text to x in the source model, map this text embedding via $\mathcal{Q}$, retrieve its top-1 nearest text in the target model, and then retrieve the top-5 images associated with that text.

As shown in Figure 57 and Figure 58, both paths yield highly consistent semantic neighborhoods: images retrieved via the text-mediated route closely match those obtained by direct image alignment. Equivalently, in terms of the commuting diagram induced by cross-modal nearest-neighbor operators, the two routes from x—direct image transport and transport via a retrieved caption—lead to nearly the same target-space neighborhood, suggesting that $\mathcal{Q}$ approximately commutes with these retrieval operators on this domain.



Figure 57: *Qualitative k-NN retrieval under two alignment paths for Oxford Pets.* We compare two routes from a source image x into the target model's image space after alignment by $\mathcal{Q}$: a direct image-alignment path (top row) and a text-mediated path (bottom row). Both routes yield highly consistent target-space neighborhoods, showing that the two alignment paths approximately close as a commuting diagram.



Figure 58: *Qualitative k-NN retrieval under two alignment paths for Caltech-101.* We compare two routes from a source image x into the target model's image space after alignment by $\mathcal{Q}$: a direct image-alignment path (top row) and a text-mediated path (bottom row). Both routes yield highly consistent target-space neighborhoods, showing that the two alignment paths approximately close as a commuting diagram.

E.17 Fine-grained Semantic Preservation Under Orthogonal Maps.

We further probe whether the learned map $\mathcal{Q}$ preserves semantic structure beyond coarse class labels. Starting from a source image x, we retrieve its nearest caption in Model A's text space and then construct a minimally perturbed variant by changing only the class token (e.g., *Birman* to *Russian Blue*) while keeping the remaining description fixed. We map this modified text via $\mathcal{Q}$, retrieve its nearest text in Model B's space, and finally retrieve the nearest image in Model B's image space.

As shown in Figure 59, the retrieved images not only reflect the edited class (Russian Blue) but also somewhat preserve finer visual attributes implied by the original description. In particular, different captions lead to distinct Russian Blue images that emphasize different features (e.g., eye prominence

vs. body pose), despite sharing the same class label. Similarly, in Figure 60, the retrieved Basset Hound images emphasize body mass and posture consistent with the original Bulldog description, rather than collapsing to a generic class prototype. Together with the previous example, this indicates that alignment by $\mathcal{Q}$ preserves fine-grained semantic structure encoded in text and image embeddings, enabling controlled semantic edits to transfer predictably across models.



Figure 59: *Fine-grained semantic preservation under controlled text edits on Oxford Pets.* After changing only the class token in the caption (Birman to Russian Blue) while keeping descriptive attributes fixed, alignment via $\mathcal{Q}$ retrieves target-model images that match the edited class while preserving caption-specific visual cues (e.g., eye prominence and facial structure), indicating preservation of fine-grained semantics beyond class identity.

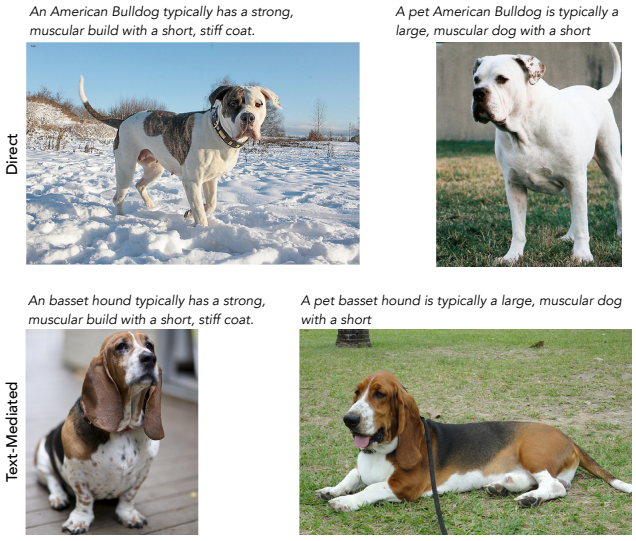

Figure 60: *Fine-grained semantic preservation under controlled text edits on Oxford Pets.* After changing only the class token in the caption (American Bulldog to Basset Hound) while keeping descriptive attributes fixed, alignment via $\mathcal{Q}$ retrieves target-model images that match both the edited class and the original attribute cues (e.g., muscular build and posture), indicating preservation of fine-grained semantics beyond class identity.