MULTI-WAY REPRESENTATION ALIGNMENT

Akshit Achara
King's College London
akshit.achara@kcl.ac.uk

Tatiana Gaintseva
Queen Mary University of London
t.gaintseva@qmul.ac.uk

Mateo Mahaut
Universitat Pompeu Fabra
mateo.mahaut@gmail.com

Pritish Chakraborty
Indian Institute of Technology, Bombay
pritish@cse.iitb.ac.in

Viktor Stenby Johansson
Technical University of Denmark
vstenby@dtu.dk

Melih Barsbey
Imperial College London
m.barsbey@imperial.ac.uk

Emanuele Rodolà
Sapienza University of Rome
Paradigma
rodola@di.uniroma1.it

Donato Crisostomi
Sapienza University of Rome
crisostomi@di.uniroma1.it

ABSTRACT

The Platonic Representation Hypothesis suggests that independently trained neural networks converge to increasingly similar latent spaces. However, current strategies for mapping these representations are inherently pairwise, scaling quadratically with the number of models and failing to yield a consistent global reference. In this paper, we study the alignment of $M \geq 3$ models. We first adapt Generalized Procrustes Analysis (GPA) to construct a shared orthogonal universe that preserves the internal geometry essential for tasks like model stitching. We then show that strict isometric alignment is suboptimal for retrieval, where agreement-maximizing methods like Canonical Correlation Analysis (CCA) typically prevail. To bridge this gap, we finally propose Geometry-Corrected Procrustes Alignment (GCPA), which establishes a robust GPA-based universe followed by a post-hoc correction for directional mismatch. Extensive experiments demonstrate that GCPA consistently improves any-to-any retrieval while retaining a practical shared reference space.

1 INTRODUCTION

Deep networks succeed largely because of the representation spaces they build. Strikingly, these spaces often resemble each other even when the models differ in architecture, data, optimization, or random seed (Kornblith et al., 2019; Moschella et al., 2022; Maiorca et al., 2023; Huh et al., 2024). This observation is captured by the Platonic Representation Hypothesis: independently trained models tend to recover a shared statistical structure of the world in their latent representations (Huh et al., 2024).

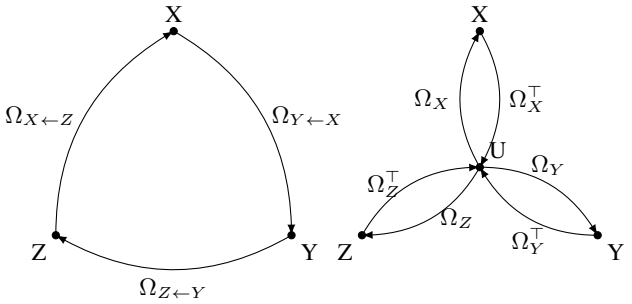

Figure 1: Pairwise alignment (left) learns a separate map for each ordered pair, which does not enforce consistency when maps are composed. Universe alignment (right) learns one map per model into a shared reference U, enabling translation between models by composition.

Motivated by these findings, a growing body of work studies representa-

tion alignment as a way to make independently trained models interoperable. Methods range from simple linear or orthogonal mappings (Maiorca et al., 2023) to richer functional correspondences (Fumero et al., 2024). When alignment succeeds, it unlocks practical use cases including model stitching, cross-modal transfer, and zero-shot composition across models (Maiorca et al., 2023; Moschella et al., 2022; Norelli et al., 2023).

Yet most alignment pipelines are still pairwise (Figure 1, left): they fit one map for each ordered pair of models and train each map independently. This is poorly matched to settings where many models are used together and translations are obtained by composing maps. In the M-model regime, pairwise alignment scales quadratically in the number of maps, requires fitting $(M-1)$ new maps to incorporate a new model, and provides no mechanism to ensure that compositions are consistent. As a result, the translation from X to Z can differ depending on whether it is learned directly or obtained by composing through intermediate models, even when each pairwise map performs well under pairwise evaluation. This can make pairwise alignment unreliable as a shared reference for larger collections of models.

A natural alternative is to factor translation through a shared universe space (Figure 1), learning one map per model to and from a common reference. In the orthogonal case this yields a generalized Procrustes formulation. Translation between arbitrary model pairs is obtained via the universe, which reduces the number of learned maps to $O(M)$ and provides a shared coordinate system that can be reused across tasks. Importantly, this construction encourages consistency under composition: the translation between any two models is unique and independent of the path taken through intermediate models. This supports multi-model use, simplifies model addition, and enables shared-coordinate analyses such as probing and aggregation.

In this work, we study the alignment of three or more models through a shared universe reference. We first adapt Generalized Procrustes Analysis (GPA) (Schönemann, 1966; Gower, 1975) to the representation alignment setting, demonstrating that its orthogonal nature yields a universe that preserves each model's internal structure, a property essential for tasks requiring geometric fidelity such as model stitching. We then empirically show that geometry preservation alone is insufficient for tasks like zero-shot retrieval, as methods maximizing cross-model agreement such as Generalized Canonical Correlation Analysis (GCCA) (Horst, 1961; Kettenring, 1971) outperform strict isometries. To bridge this gap, we propose Geometry-Corrected Procrustes Alignment (GCPA), which combines the best of both worlds. GCPA first establishes a robust orthogonal universe via GPA, then applies a post-hoc shared correction to minimize residual directional mismatch.

We evaluate these approaches on benchmarks spanning inter-architecture probing, multilingual translation, and cross-camera retrieval. Our results show that while naturally multi-way techniques are necessary for scaling, the specific choice of objective matters: while GPA provides a consistent geometric reference, the proposed geometry-corrected universe (GCPA) is required to achieve state-of-the-art performance in any-to-any retrieval while retaining the practical benefits of the shared Procrustes frame.

Contributions.

1. We extend orthogonal representation alignment to three or more models by adapting Generalized Procrustes Analysis (GPA) to the neural representation setting.
2. We demonstrate that while GPA provides a consistent global reference superior to pairwise maps, it lags behind agreement-maximizing methods (like GCCA) on retrieval tasks.
3. We introduce Geometry-Corrected Procrustes Alignment (GCPA) to bridge this gap, establishing a robust universe via GPA and applying a post-hoc correction for directional mismatch to achieve superior performance across diverse datasets and settings.

2 Related Work

Representational similarity Work on representational similarity typically (i) measures how closely different learned spaces correspond (Kornblith et al., 2019; Klabunde et al., 2025; Glielmo et al., 2022; Huh et al., 2024; Acevedo et al., 2025), (ii) leverages such similarity to align spaces via explicit maps (Moschella et al., 2022; Maiorca et al., 2023; Fumero et al., 2024; Cannistraci et al.),

or (iii) studies similarity and alignment jointly (Fumero et al., 2024). The applications span multi-modal alignment (Norelli et al., 2023; Cicchetti et al., 2025; Yue et al., 2025; Gröger et al., 2025), cross-view alignment (Huang et al.), as well as cross-lingual alignment (Jawanpuria et al., 2019). While there are methods not requiring paired data (Grave et al., 2019; Jha et al., 2025; Schnaus et al., 2025), we study in this paper the common case where the correspondence is available. A common and effective alignment tool in this line is Procrustes analysis, which fits an orthogonal map between two spaces and preserves their internal distances and angles (Hurley & Cattell, 1962; Grave et al., 2019; Maiorca et al., 2023). More broadly, our setting can be viewed as a special case of multi-space representation learning focused on alignment (Li et al., 2018), and is connected to backward-compatible representation learning, which aims to keep updated embeddings comparable to earlier ones (Shen et al., 2020; Zhang et al., 2022), as well as to communication between pretrained networks (Mahaut et al., 2025; Ramesh & Li).

Multiset alignment An alternative to pairwise mapping is to learn a shared latent space from multiple views via multiset canonical correlation analysis (GCCA) and related variants (Horst, 1961; Kettenring, 1971; Fu et al., 2017; Sørensen et al., 2021). These methods provide a principled M-way objective and are natural baselines for multi-view alignment, especially in retrieval-style settings where matched items should be directly comparable in a common set of coordinates. In contrast, our focus is on a single shared reference learned with one map per model, so that translations between models are obtained by composing through the reference and new models can be incorporated by fitting only one additional map.

3 Methodology

We consider the problem of aligning $M \geq 3$ distinct representation spaces $\{X_m\}_{m=1}^M$, where $X_m \in \mathbb{R}^{N \times d}$ contains N matched samples (or is padded to a common dimension). In our setup, all representations are standardized within each space. Our goal is to establish a shared coordinate system that enables any-to-any translation.

3.1 The Universe Factorization

Standard alignment approaches are inherently pairwise. To map space n to space m, one typically minimizes the discrepancy between matched pairs:

$$\mathcal{L}_{\text{pair}}(\Omega) = \sum_{m<n} \|X_m - X_n \Omega_{m \leftarrow n}\|_F^2. \tag{1}$$

While effective for two models, this approach scales quadratically, requiring $O(M^2)$ maps for M models. Furthermore, it lacks a global reference; the composition of learned maps often violates cycle consistency (i.e., $\Omega_{m \leftarrow k}\Omega_{k \leftarrow n} \neq \Omega_{m \leftarrow n}$), making multi-hop translation unreliable.

To address this, we adopt an **object-to-universe factorization**. Rather than learning pairwise maps directly, we assume the existence of a shared statistical "universe" U and learn a single map Ω_m from each model into this reference:

$$\Omega_{m \leftarrow n} := \Omega_m \Omega_n^\top. \tag{2}$$

This reduces the complexity from $O(M^2)$ to $O(M)$ and guarantees cycle consistency by design. The central question then becomes: *What are the geometric properties of this universe, and how should U be constructed?*

3.2 Constructing an Isometric Universe (GPA)

For tasks such as model stitching or latent probing, it is critical that alignment does not distort the internal topology of the individual models. We therefore first consider the construction of an orthogonal universe, where each map Ω_m is constrained to the orthogonal group $O(d)$.

This formulation aligns with **Generalized Procrustes Analysis (GPA)**. We seek a consensus centroid U and a set of rotations $\{\Omega_m\}$ that minimize the global dispersion:

$$\min_{\{\Omega_m \in O(d)\},\, U} \sum_{m=1}^{M} \|X_m \Omega_m - U\|_F^2. \tag{3}$$

Optimization proceeds via alternating minimization:

1. **Update Consensus:** $U \leftarrow \frac{1}{M}\sum_{m=1}^{M} X_m \Omega_m$.
2. **Update Maps:** Solve the orthogonal Procrustes problem for each Ω_m to align X_m to the current U.

Because Ω_m is an isometry, distances and angles within each model are preserved exactly in the universe ($\|\Omega_m x - \Omega_m y\| = \|x - y\|$). This makes GPA the natural choice for geometric interoperability.

3.3 The Limits of Isometry in Retrieval

While GPA provides a robust geometric reference, empirical evidence suggests that strict isometry is suboptimal for **retrieval tasks**. In cross-modal or zero-shot retrieval, maximizing the cosine agreement between matched samples often requires reshaping the feature spaces to suppress modality-specific noise, a degree of freedom that orthogonal maps lack.

To quantify this "retrieval gap," we contrast GPA with **Generalized Canonical Correlation Analysis (GCCA)**. GCCA relaxes the orthogonality constraint, allowing linear projections Φ_m that deform the spaces to maximize shared correlation:

$$\min_{\{\Phi_m\}} \sum_{i<j} \|X_i\Phi_i - X_j\Phi_j\|_F^2 \quad \text{s.t.} \quad \sum_{m=1}^{M} \Phi_m^\top \Phi_m = I. \tag{4}$$

The following proposition highlights why GCCA often outperforms isometries in pure retrieval:

Proposition 3.1 (GCCA minimizes squared discrepancy)**.** *The global minimizer of Eq. 4 is obtained via a spectral decomposition of a matrix constructed from cross-view correlations (Theorem B.1). Among linear shared-basis embeddings satisfying the constraint, this solution globally minimizes the total pairwise mismatch energy.*

These considerations pinpoint a trade-off: GCCA achieves minimal mismatch (ideal for retrieval) but does so by distorting the internal geometry of the models (harmful for stitching/probing); GPA preserves geometry but suffers from residual directional mismatch.

3.4 Geometry-Corrected Procrustes Alignment (GCPA)

To bridge the gap between geometric fidelity and retrieval performance, we propose **Geometry-Corrected Procrustes Alignment (GCPA)**.

Our insight is that we can retain the robust orthogonal universe of GPA as a "scaffold," and then apply a shared, non-linear correction to "polish" the alignment. This correction minimizes the residual directional mismatch that GPA leaves behind, without discarding the universe coordinate system.

The Consensus Principle. We observe that in the GPA universe, the average direction of a matched sample across all M models serves as a robust target. Let $\hat{u}_{m,i}$ be the unit-normalized vector for sample i mapped into the universe by model m. We define the **consensus direction** c_i:

$$c_i = \text{norm}\left(\frac{1}{M}\sum_{m=1}^{M} \hat{u}_{m,i}\right). \tag{5}$$

Moving individual representations toward this consensus strictly improves global agreement. The following identity links consensus agreement to total pairwise agreement:

Proposition 3.2. *Fix a sample index i and let $\{\hat{u}_{m,i}\}_{m=1}^{M} \subset \mathbb{S}^{d-1}$ denote the unit-normalized universe directions for that sample across spaces. Define the consensus direction*

$$c_i = \text{norm}\left(\frac{1}{M}\sum_{m=1}^{M} \hat{u}_{m,i}\right).$$

Then

$$\frac{1}{M}\sum_{m=1}^{M}\langle \hat{u}_{m,i}, c_i\rangle = \frac{1}{M}\left\|\sum_{m=1}^{M}\hat{u}_{m,i}\right\|_2, \tag{6}$$

$$\sum_{m<n}\langle \hat{u}_{m,i}, \hat{u}_{n,i}\rangle = \frac{1}{2}\left(\left\|\sum_{m=1}^{M}\hat{u}_{m,i}\right\|_2^2 - M\right). \tag{7}$$

i.e. increasing the cosine similarity of each view to the consensus c_i monotonically increases the sum of pairwise cosine similarities between all views.

The GCPA Algorithm. GCPA operates in two stages: i) **GPA Initialization:** We first solve the standard GPA problem to obtain the orthogonal maps $\{\Omega_m\}$ and the base universe representations $X_m^\star = X_m\Omega_m$; ii) **Geometry Correction:** We train a lightweight, shared residual map T_θ (a small MLP) that acts on the universe coordinates. The objective minimizes the distance to the consensus while strictly penalizing deviation from the trusted GPA geometry:

$$\mathcal{L}_{\text{correct}} = -\sum_{m,i}\langle T_\theta(\hat{u}_{m,i}), c_i\rangle + \lambda\mathcal{L}_{\text{trust}}, \tag{8}$$

where $\mathcal{L}_{\text{trust}}$ penalizes T_θ if the angular deviation from the original GPA vector exceeds a tight threshold. By construction, T_θ returns a row-normalized direction and c_i is the per-sample consensus direction.

By sharing T_θ across all models, GCPA refines the universe itself rather than individual models. This allows us to achieve the high agreement typically associated with CCA-style methods while remaining anchored to the geometrically consistent Procrustes frame.

Additional theoretical details (including related theorems and corollaries) are deferred to Appendix B. Algorithmic summaries for GCCA and GCPA are provided in Appendix D and Appendix E.

4 Experiments

We study multi-model alignment through the lens of a shared universe reference. We begin with probing and stitching tasks, where preserving each model's internal structure is important. We then consider aggregation in a common coordinate system through clustering. Finally, we study the task of retrieval, first in settings with strong shared structure where shared-basis methods are effective, and then in more heterogeneous settings where an orthogonal universe remains valuable but residual directional mismatch benefits from a universe-level correction.

4.1 Baselines and evaluation

We compare **No Alignment (NA)**, **Pairwise (PW)** orthogonal Procrustes, M**-way GPA**, **GCCA** (retrieval settings), and **GCPA**; see Section 3 for definitions.

All alignment parameters are fit on the training split using sample identity to define correspondences across spaces, and results are reported on held-out test sets. For probing, we report stitching accuracy by training a linear probe on one space and evaluating it on another after mapping. For retrieval, we average performance over all ordered pairs $(i \to j)$ and include pairwise breakdowns where relevant. For retrieval experiments, we apply PCA on features to a common dimension d prior to alignment (only when model features have different dimensions), fitting PCA on the training split and applying it to validation/test. For probing and clustering, we use the original feature spaces without PCA.

We use a fixed, dataset-agnostic GCPA configuration for all main results. Appendix C.1 shows that additional gains are possible with an alternative set of parameters.

4.2 Geometric interoperability

4.2.1 Multi-way Alignment Strengthens Weak Pairwise Links

We investigate the ability of a shared universe to stabilize the alignment between two models that share little surface-level similarity. To test this, we isolate a fragile pair of models with poor direct correspondence induced by training them on disparate views of the data. We then attempt to align them via a universe of healthy anchor models. As a dataset, we utilize a version of CIFAR-100 (Krizhevsky et al., 2009) where we introduce a distribution shift by replacing 85% of the images with binary edge maps, as described in further detail in Appendix A.1.

We first establish a baseline by fitting a direct pairwise orthogonal map between the fragile pair. Due to the distribution shift between their training representations, this direct link performs poorly. We then construct an M-way GPA universe containing the fragile pair alongside a set of standard models. We progressively expand this set by adding more anchor models to the universe and re-evaluate the translation quality between the original fragile pair when routed through the shared reference.

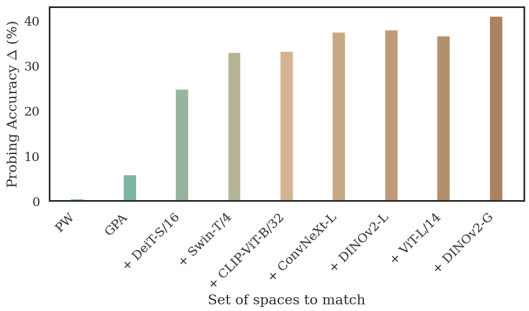

Figure 2: **Multi-way alignment stabilizes fragile connections.** On edge-heavy CIFAR-100, we isolate a "weak" model pair with poor alignment. By progressively expanding the universe with robust models and refitting the universe, we observe a monotonic increase in stitching accuracy between the original fragile pair.

As shown in Figure 2, the universe acts as a geometric hub. While the direct pairwise map fails to capture the correspondence, increasing the number of diverse models in the universe stabilizes the alignment. The shared reference leverages the structural consensus of the anchor models to triangulate the relationship between the fragile pair, effectively healing the weak link. (See Section A.1 for experimental details on the expanding set). We also replicate this experiment for GCPA in Section A.1, showing a similar trend.

4.2.2 Efficiently Extending the Universe

Once a universe is learned, it can be reused rather than rebuilt. Pairwise alignment requires learning $M-1$ new maps when extending an aligned set of M models, whereas the shared-universe formulation integrates a new model X_{M+1} by fitting a single orthogonal map Ω_{M+1} into the existing universe. Translation to and from any existing model then follows by composition:

$$\begin{aligned} \Omega_{m\leftarrow(M+1)} &= \Omega_m \Omega_{M+1}^\top, \\ \Omega_{(M+1)\leftarrow m} &= \Omega_{M+1} \Omega_m^\top. \end{aligned} \tag{9}$$

Figure 3: **Cross-model probing on CIFAR-100.** Adding a new model by fitting only Ω_{M+1} into a fixed universe (GPA-ADD) approaches refitting the universe (GPA-REFIT) and outperforms PW alignment. To cover diverse scenarios, we use four different base model sets where the first two (from the left) sets consist of three models and the next two consist of five models each.

On CIFAR-100 probing, we learn an M-way universe on a base set of models and then introduce a new model X_{M+1}. We report the average cross-model test accuracy over all directed transfers between the new model and the base models. We compare **PW**, which fits the incident pairwise maps; **GPA-REFIT**, which refits the universe on $M+1$ models; and **GPA-ADD**, which fits only Ω_{M+1} while keeping the base universe fixed (Appendix A.2).

The results in Figure 3 demonstrate the efficacy of this approach. Across diverse architectures, including ViT-B/16, Swin-T/4, ResNet18, and MobileNetv2, simply fitting the new model into the

fixed universe (**GPA-ADD**) consistently matches or outperforms the computationally expensive baselines. Specifically, fully refitting the universe (**GPA-REFIT**) achieves higher average stitching accuracy than independent pairwise alignment (**PW**) and **GPA-ADD** as expected and GPA-ADD slightly improves over pairwise alignment. This validates that the shared coordinate system remains robust and can be extended efficiently without the need for global retraining.

4.3 Aggregation and Clustering

Clustering tests a different property than probing or retrieval. Rather than asking whether two models are interchangeable after alignment, it asks whether mapping multiple models into a shared coordinate system yields a space where semantic class structure is easy to recover without supervision. This directly evaluates if multiple pretrained spaces can be aligned into a shared coordinate system consistently across encoders, which can be verified if examples with the same label form tighter groups after alignment.

We evaluate intent clustering on MASSIVE (FitzGerald et al., 2022), which consists of short user queries annotated with one of 60 intent classes. We embed the same evaluation split with $M=10$ pretrained multilingual models (one space per model), so each example provides a matched row across all spaces. We fit alignment on the training split using example identity as correspondence, then map each model's test split embeddings into a shared space and run k-means clustering with k set as the number of intents in the test split (for 5 different seeds).

We compare clustering in the original encoder spaces (**NA**) against clustering after mapping with **GPA**, **GCCA**, and our geometry-corrected universe (**GCPA**). Clustering quality is measured by Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). We report mean$\pm$std across the 10 models to reflect how consistently each alignment method supports aggregation across different pretrained spaces.

Table 1 presents the quantitative results. We observe that standard alignment methods like GPA and GCCA do not improve clustering quality over the original unaligned spaces (NA), suggesting that global rotation or linear projection alone does not enhance the separability of intent classes. In contrast, GCPA achieves a significant improvement in both metrics, boosting the Adjusted Rand Index by over 50% relative to the baselines. This indicates that the post-hoc geometric correction successfully pulls semantically similar items together, creating a shared representation where class structure is more distinct and easier to recover.

Method	ARI	NMI
NA	0.198$\pm$0.035	0.509$\pm$0.041
GPA	0.198$\pm$0.036	0.509$\pm$0.041
GCCA	0.194$\pm$0.032	0.508$\pm$0.038
GCPA (ours)	**0.300$\pm$0.024**	**0.617$\pm$0.022**

Table 1: **Clustering on MASSIVE.** We run k-means separately for each mapped model and report mean$\pm$std across models. GCPA outperforms the other approaches on all 10 models.

4.4 Retrieval

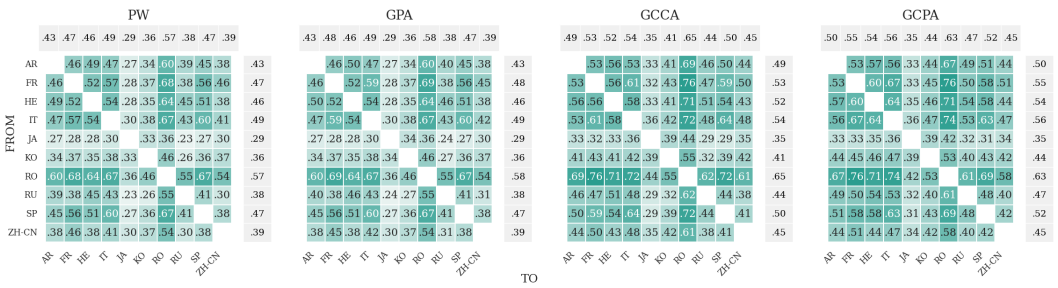

Figure 4: **Cross-lingual retrieval on TED-Multi (rank-1).** GCPA outperforms GCCA, GPA, and pairwise orthogonal alignment.

	$M = 3$		$M = 5$		$M = 10$	
Method	Avg ↑	Worst ↑	Avg ↑	Worst ↑	Avg ↑	Worst ↑
PW	0.571	0.468	0.471	0.303	0.430	0.230
GPA	0.572	0.469	0.474	0.308	0.433	0.236
GCCA	0.628	0.509	0.531	0.366	0.487	0.288
GCPA (ours)	**0.637**	**0.532**	**0.553**	**0.419**	**0.503**	**0.314**

Table 2: **Scaling with the number of languages on TED-MULTI.** Cross-lingual rank-1 retrieval for fixed language subsets at M=3, M=5, and M=10. We report the average over the evaluated A→B language pairs and the worst-case (minimum) pair performance.

4.4.1 MULTILINGUAL TRANSLATION

We evaluate cross-lingual sentence retrieval on TED-MULTI[1], a multilingual corpus of TED talk transcripts with aligned sentence translations across ten languages. Each language is embedded with a dedicated pretrained model (Appendix A.4, Table 5), producing matched-dimensional representations. Given a query sentence in language A, the task is to retrieve its paired translation in language B using cosine similarity; we measure performance as rank-1 retrieval accuracy over the evaluated A→B language pairs.

Direct comparison in the original representation spaces yields near-zero accuracy, indicating that independently trained multilingual models are not natively comparable. Pairwise orthogonal alignment (PW) substantially improves retrieval, and M-way GPA improves further by aligning all languages into a shared universe reference. Figure 4 shows that Geometry-Corrected Procrustes Alignment (GCPA) achieves the best retrieval performance, outperforming both GCCA and purely orthogonal universes (GPA).

Table 2 examines how alignment quality scales with the number of languages ($M \in (3, 5, 10)$). For M=3, we align (EN, FR, ES) and (EN, AR, ZH); for M=5 we align (EN, FR, ES, IT, RU) and (EN, AR, RU, JA, ZH). While retrieval difficulty naturally increases as more spaces are added, GCPA consistently yields the highest performance across all settings. Crucially, it improves both the average accuracy and the worst-case pair performance relative to GCCA and GPA, demonstrating that the corrected universe remains robust even as the collection of models grows.

4.4.2 ROBUSTNESS TO CORRESPONDENCE NOISE

We stress-test alignment under imperfect training correspondences by corrupting the pairing structure used to fit the alignment. On TED-MULTI, we select a language triad and, for each language independently, randomly permute 75% of the training sentences before fitting alignment; evaluation is performed on the unmodified test split. We consider three disjoint triads (English/French/Spanish; Arabic/Hebrew/Russian; Korean/Japanese/Chinese) and report results averaged over the three trials.

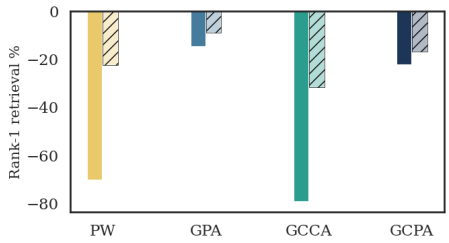

Figure 5: **Robustness to correspondence noise on TED-MULTI.** Rank-1 retrieval accuracy (%) on the clean test split relative to the unshuffled baseline. Solid bars average over the evaluated pairs within the triad; hatched bars average over all evaluated pairs that involve at least one shuffled language. Results are averaged over three disjoint triads.

Figure 5 summarizes the resulting change in cross-lingual rank-1 retrieval accuracy (percentage points) relative to the unshuffled baseline. The solid bar reports the mean change over evaluated retrieval directions restricted to the triad. The hatched bar reports a broader aggregate, namely the mean change over all evaluated retrieval directions where at least one side is one of the shuffled languages. Across both aggregates, GCCA is more sensitive to correspondence noise, while GCPA degrades more gracefully and retains higher retrieval performance under the same corruption level.

Although GPA can show a slightly smaller average drop under this corruption, GCPA attains the highest absolute retrieval accuracy, indicating that the correction remains beneficial even when correspondences are imperfect.

[1] https://huggingface.co/datasets/neulab/ted_multi

4.4.3 PERSON RE-IDENTIFICATION

We evaluate cross-camera person re-identification on MARKET-1501 (Zheng et al., 2015) in a zero-shot setting with identity level correspondences. We treat each camera as a separate space and learn an M-way alignment across camera-specific representations using training identities only and use the query and gallery sets for retrieval (Appendix A.3). At test time, we perform cross-camera retrieval where queries from camera C_i are matched against galleries from camera C_j ($j \neq i$) using cosine similarity in the aligned coordinates. To reduce camera imbalance, we subsample each camera to match the minimum number of images per identity.

Figure 6 presents the retrieval results in terms of Mean Average Precision (mAP). We observe that without alignment (NA), independently trained encoders fail to generalize across views, yielding negligible performance. While standard orthogonal alignments (PW and GPA) successfully restore interoperability, they are consistently outperformed by the agreement-maximizing baseline (GCCA), confirming that strict isometry limits retrieval flexibility. However, our proposed GCPA yields the highest performance across all camera pairs. By correcting the residual directional mismatch within the shared universe, GCPA achieves a distinct margin over GCCA, particularly in transitions such as $6 \to 3$, $3 \to 5$ and $1 \to 3$, demonstrating that combining geometric stability with a consensus-driven correction is crucial for robust cross-view retrieval.

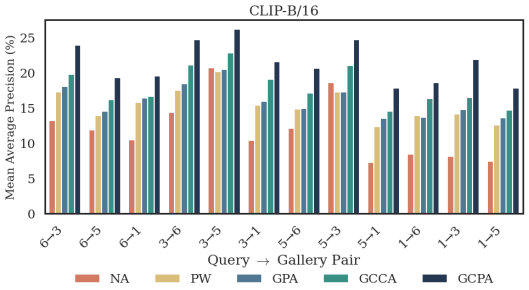

Figure 6: **Cross-camera retrieval on MARKET-1501 (mAP, %)**. Geometry-Corrected Procrustes Alignment (GCPA) improves over GCCA, GPA, and PW.

We also conduct an analysis of hyperparameters for cross-camera retrieval and highlight their mild impact on the performance. While we use a fixed GCPA setting across experiments, further performance gains may be obtained by dataset-specific tuning. See Appendix C for a sensitivity sweep over (τ, λ) and the induced geometry drift and retrieval performance.

4.4.4 MULTIMODAL ALIGNMENT

We study zero-shot cross-modal retrieval on FLICKR8K (Harwath & Glass, 2015), which pairs images with text and spoken captions. We construct modality-specific spaces using BERT (text) (Devlin et al., 2019), DINOv2 (image) (Caron et al., 2021), and HuBERT (audio) (Hsu et al., 2021), alongside CLIP (image) (Radford et al., 2021) to serve as a strong pivotal anchor. We learn an M-way alignment across these modalities, comparing NA, PW, GPA, GCCA, and GCPA. To enforce one-to-one correspondence, we use per-image mean features for text and audio.

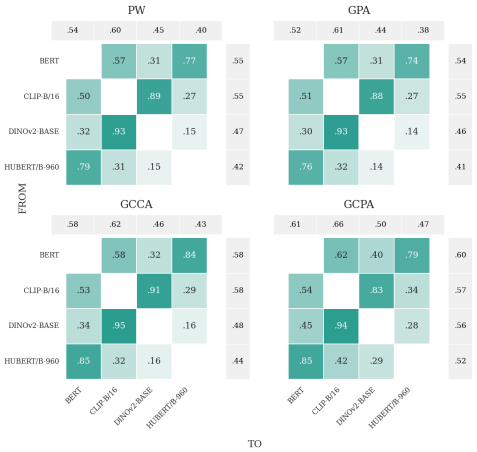

Figure 7: **Rank-1 cross-modal retrieval (%) on FLICKR8K**. GCPA improves audio↔image ↔text retrieval while retaining the GPA universe.

Figure 7 visualizes the pairwise retrieval performance (Rank-1 accuracy) as heatmaps. We observe that the audio modality (HuBERT) presents the most significant alignment challenge, showing weak connectivity to visual models in the standard pairwise (PW) and GPA baselines. For instance, HuBERT retrieval of DINOv2 targets is only $14 - 15\%$. While GCCA improves some connections, it struggles to bridge the gap between audio and pure vision encoders. In contrast, GCPA yields substantial improvements in these hard settings: HuBERT→DINOv2 retrieval nearly doubles to 29%, and HuBERT→CLIP improves from 32% (GPA) to 42% (GCPA). Overall, GCPA achieves the highest column averages across all modalities, demonstrating that the geometric correction effectively pulls audio representations into alignment with the vision-text backbone.

Method	1	2	3	Mean
	$\Delta^+ \downarrow$ ($\Gamma_{90} \uparrow$)	$\Delta^+ \downarrow$ ($\Gamma_{90} \uparrow$)	$\Delta^+ \downarrow$ ($\Gamma_{90} \uparrow$)	$\Delta^+ \downarrow$ ($\Gamma_{90} \uparrow$)
GPA	0.776 (0.732)	0.689 (0.787)	0.721 (0.761)	0.729 (0.760)
GCCA	0.779 (0.728)	0.730 (0.758)	0.724 (0.758)	0.744 (0.748)
GCPA (ours)	**0.623 (0.824)**	**0.503 (0.880)**	**0.554 (0.854)**	**0.560 (0.853)**

Table 3: **Agreement analysis (Universe) for multimodal retrieval.** We report the mean matched cross-modal distance Δ^+ (lower is better) together with the upper-tail matched triplet agreement Γ_{90} (higher is better), across three encoder triplets.

To understand the geometric mechanism behind these gains, we analyze how alignment changes (i) the typical distance between matched cross-modal pairs and (ii) the strength of high-agreement matched triplets. We consider images (I), text (T), and audio (A) modalities, and evaluate all three cross-modal pairs ($I \leftrightarrow T, I \leftrightarrow A, T \leftrightarrow A$).

We use cosine distance on row-wise normalized vectors. Let (z_i^I, z_i^T, z_i^A) denote the aligned embeddings for sample i in the three modalities after normalization, and define $d(u, v) = 1 - \langle u, v \rangle$. For each sample i, we define a matched cross-modal distance by averaging the three within-triplet distances:

$$d_i^+ = \tfrac{1}{3}\Big(d(z_i^I, z_i^T) + d(z_i^I, z_i^A) + d(z_i^T, z_i^A)\Big),$$

$$\Delta^+ = \tfrac{1}{N}\sum_{i=1}^{N} d_i^+$$

where lower Δ^+ indicates that matched items are closer on average across modalities.

To summarize three-way agreement, we measure the magnitude of the mean direction of each matched triplet:

$$\gamma_i \;=\; \left\|\tfrac{1}{3}\big(z_i^I + z_i^T + z_i^A\big)\right\|_2 .$$

We then report a high-agreement summary Γ_{90} defined as the value of γ_i at the 90th percentile over the dataset, i.e., the agreement level attained by the 10% most coherent matched triplets. Larger Γ_{90} indicates that a substantial fraction of matched triplets achieve strong cross-modal coherence.

Table 3 reports Δ^+ together with Γ_{90} across three encoder triplets. The results confirm that GCPA consistently reduces the mean cross-modal distance Δ^+ compared to GPA ($0.729 \rightarrow 0.560$) and GCCA ($0.744 \rightarrow 0.560$), indicating a much tighter clustering of matched modalities. Furthermore, GCPA significantly increases Γ_{90} ($0.760 \rightarrow 0.853$), showing that the correction successfully concentrates the multimodal universe, creating a high-coherence core that supports improved retrieval performance.

5 Conclusions

In this work, we addressed the challenge of aligning multiple representation spaces, demonstrating that standard pairwise strategies fail to scale and lack the consistency required for the M-model regime. To overcome these limitations, we proposed a framework that aligns all models into a single shared coordinate system. This formulation reduces alignment complexity from quadratic to linear, facilitates the efficient addition of new models, and enforces cycle consistency by design. Our analysis revealed a critical trade-off in constructing this universe: while Generalized Procrustes Analysis (GPA) yields an orthogonal reference that preserves the internal geometry necessary for model stitching, it is outperformed in zero-shot retrieval by agreement-maximizing methods like GCCA. We resolved this tension with **Geometry-Corrected Procrustes Alignment (GCPA)**. By using the GPA universe as a robust geometric scaffold and applying a consensus-driven correction, GCPA bridges the gap between geometric fidelity and semantic agreement. Our experiments across multilingual, cross-camera, and multimodal benchmarks confirm that GCPA effectively synthesizes the best of both worlds, achieving state-of-the-art retrieval performance while retaining a practical, composable reference space.

Acknowledgments

This project was ideated as a part of the London Geometry and Machine Learning Summer School (LOGML) 2025. We are grateful to the organizers for hosting the summer school. Akshit Achara was

supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant reference number EP/Y035216/1] Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) at King's College London. Matéo Mahaut received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101019291), and from the Catalan government (AGAUR grant SGR 2021 00470). This paper reflects the authors' view only, and the ERC is not responsible for any use that may be made of the information it contains. This work was also supported by MUR FIS2 grant n. FIS-2023-00942 "NEXUS" (cup B53C25001030001) and by Sapienza University of Rome via the Seed of ERC grant "MINT.AI" (cup B83C25001040001).

REFERENCES

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. An approach to identify the most semantically informative deep representations of text and images, 2025. URL https://arxiv.org/abs/2505.17101. 2

Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà. From bricks to bridges: Product of invariances to enhance latent space communication. In *The Twelfth International Conference on Learning Representations*. 2

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009. 14

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. 9

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A TRIANGLE enables multimodal alignment beyond cosine similarity. *arXiv [cs.LG]*, 2025. URL http://arxiv.org/abs/2509.24734. 3

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019. 9

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 14

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022. 7

Xiao Fu, Kejun Huang, Mingyi Hong, Nicholas D Sidiropoulos, and Anthony Man-Cho So. Scalable and flexible multiview max-var canonical correlation analysis. *IEEE Transactions on Signal Processing*, 65(16):4150–4165, 2017. 3

Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. *Advances in Neural Information Processing Systems*, 37:66178–66203, 2024. 2, 3

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gábor Csányi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS Nexus*, 1(2):pgac039, 04 2022. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgac039. URL https://doi.org/10.1093/pnasnexus/pgac039. 2

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 2

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1880–1890. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/grave19a.html. 3

Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With limited data for multimodal alignment, let the structure guide you. *arXiv preprint arXiv:2506.16895*, 2025. 3

David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, pp. 237–244. IEEE, 2015. 9

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 14

Paul Horst. *Generalized canonical correlations and their application to experimental data*. Number 14. Journal of clinical psychology, 1961. 2, 3

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021. 9

Kuan Wei Huang, Brandon Li, Bharath Hariharan, and Noah Snavely. C3po: Cross-view cross-modality correspondence by pointmap prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 3

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *CoRR*, 2024. 1, 2

John R Hurley and Raymond B Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral science*, 7(2):258, 1962. 3

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. Learning multilingual word embeddings in latent metric space: a geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120, 2019. 3

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings, 2025. URL https://arxiv.org/abs/2505.12540. 3

Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. 2, 3

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. 2

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019. 1, 2

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 14

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018. 3

Matéo Mahaut, Roberto Dessi, Francesca Franzon, and Marco Baroni. Referential communication in heterogeneous communities of pre-trained visual deep networks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=8L3khbpUJL. 3

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023. 1, 2, 3

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *Advances in Neural Information Processing Systems*, 36:15303–15319, 2023. 2, 3

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 9

Vignav Ramesh and Kenneth Li. Communicating activations between language model agents. In *Forty-second International Conference on Machine Learning*. 3

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24983–24992, 2025. 3

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 2

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. 3

Mikael Sørensen, Charilaos I Kanatsoulis, and Nicholas D Sidiropoulos. Generalized canonical correlation analysis: A subspace intersection approach. *IEEE Transactions on Signal Processing*, 69:2452–2467, 2021. 3

Yisong Yue, Been Kim, et al. Escaping platos cave: Jam for aligning independently trained vision and language models. *arXiv preprint arXiv:2507.01201*, 2025. 3

Binjie Zhang, Yixiao Ge, Yantao Shen, Shupeng Su, Fanzi Wu, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. Towards universal backward-compatible representation learning. *arXiv preprint arXiv:2203.01583*, 2022. 3

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015. 9

A Additional details

A.1 Multi-Way Alignment Strengthens Weak Pairwise Links

To simulate a fragile pair (Section 4.2.1), we utilize a ResNet-18 (He et al., 2016) and a ViT-T/16 (Dosovitskiy, 2020) trained on a corrupted version of CIFAR-100 (Krizhevsky et al., 2009). We introduce a distribution shift by replacing 85% of the training images for these two models with binary Canny edge maps (Canny, 2009), retaining only 15% as original RGB. This degrades the correlation between their learned features compared to models trained on standard data. The anchor models added to the universe are trained on the standard uncorrupted RGB training set. We use a 45,000/5,000 stratified train/validation split for alignment fitting. Crucially, the corruption is applied only to the training samples used to learn the alignment. The stitching accuracy is reported on the standard, uncorrupted CIFAR-100 test set to evaluate generalization. Now, we replicate this experiment for GCPA.

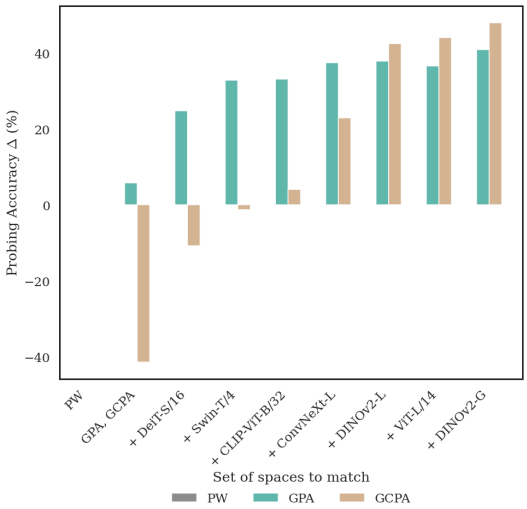

Figure 8: **Weak-link probing on edge-heavy CIFAR-100 under an expanding alignment set.** We plot the change in probing accuracy (in %) for the same fragile pair as additional models are added. GPA improves steadily with more anchors, while GCPA can be unstable for very small sets but surpasses GPA once the universe is supported by sufficiently strong models.

Figure 8 shows that GCPA can underperform when correspondences are weak. GCPA nudges universe directions toward a per-sample multi-model consensus; if this consensus is distorted by imperfect correspondences (e.g., natural image versus edge-map), the correction can reinforce the mismatch. As additional models are added, the consensus is better supported and GCPA recovers, surpasses GPA, and continues improving as more anchors are introduced. This behavior is consistent with the intended role of the correction, which is most beneficial when multi-way agreement reflects shared semantics rather than corruption-specific variation.

Since the GCPA geometry correction is directional, Fig. 8 reports the unit-direction form to isolate the effect of directional correction in the shared universe coordinates. For scale-sensitive linear probing, we additionally find that rescaling corrected universe coordinates to match the pre-correction GPA norm yields further consistent improvements. Even in this unit-direction form, GCPA can match or exceed the pairwise probing accuracy of the same architecture pair trained on uncorrupted RGB data, and rescaling improves further.

A.2 Incremental addition

For a base set of M models and an added model X_{M+1}, we report the mean cross-model test accuracy over directed transfers involving the added model:

$$\text{AvgNew} = \frac{1}{2M} \sum_{m=1}^{M} \Big(\text{Acc}(X_{M+1} \to X_m) + \text{Acc}(X_m \to X_{M+1}) \Big).$$

PW fits $\Omega_{m \leftarrow (M+1)}$ and $\Omega_{(M+1) \leftarrow m}$ independently for each m. **GPA-REFIT** fits $\{\Omega_m\}_{m=1}^{M+1}$ jointly on $\{X_m\}_{m=1}^{M+1}$. **GPA-ADD** keeps $\{\Omega_m\}_{m=1}^{M}$ fixed and learns only Ω_{M+1}, inducing pairwise maps by $\Omega_{m \leftarrow (M+1)} = \Omega_m \Omega_{M+1}^\top$ and $\Omega_{(M+1) \leftarrow m} = \Omega_{M+1} \Omega_m^\top$. We list all model sets used in Table 4.

Added model	Base model set
ViT-B/16	DINOv3 ViT-L/16, ConvNeXt-L, ResNet-50
Swin-T/4	DINOv2-B, CLIP ViT-B/32, ViT-B/16
ResNet-18	ConvNeXt-L, DINOv2-L, CLIP ViT-B/32, ViT-B/16, Swin-L/12
MobileNetV2	ConvNeXt-B, DINOv3 ViT-L/16, ViT-B/16, CLIP ViT-B/16, ResNet-34

Table 4: **Incremental addition configurations.** Each row defines a base model set used to learn a universe, and an added model integrated into that universe.

Language	Encoder
English (EN)	RoBERTa-base
French (FR)	CamemBERT-base
Spanish (ES)	Spanish BERT-base
Italian (IT)	Italian BERT-base
Arabic (AR)	AraBERTv2 (base)
Hebrew (HE)	HeBERT (base)
Russian (RU)	RuBERT (base)
Japanese (JA)	Japanese BERT-base
Korean (KO)	Korean BERT-base
Chinese (ZH)	Simplified Chinese BERT-base

Table 5: **Language-specific encoder set used in our multilingual experiments.** Each row contains the language and the corresponding encoder used in our multilingual experiments.

A.3 Person Re-Identification

Market-1501 provides images from six cameras. We select four cameras with the largest identity overlap, yielding 437 training identities and 465 identities for both query and gallery. Query and gallery identities coincide and are disjoint from train. Each selected camera defines one space (C_1, C_2, C_3, C_4). We learn the multi-way alignment using per-image embeddings computed on the training split. For evaluation, we subsample each camera to match the per-identity minimum number of images in query and gallery, and assess all cross-camera pairs $C_i \mapsto C_j$ with $i \neq j$. Note that this dataset does not provide cross-camera image-level correspondences and therefore, our evaluation aligns person identities with no assumption of per-image pairing across cameras.

A.4 Multilingual encoder set

We use one language-specific Transformer encoder per language. This encoder set is used across our multilingual experiments, including the multilingual retrieval(Section 4.4.1), correspondence-noise stress test(Section 4.4.2) and the massive multilingual clustering experiment(Section 4.3).

A.5 Cross-modal protocol details.

Flickr8k contains 8,000 images, each paired with five text and five spoken captions. We construct modality-specific spaces (image, text, audio) and form per-image mean representations for text and audio to obtain one-to-one correspondence with image features. We learn an M-way alignment with one space per modality and evaluate zero-shot cross-modal retrieval by querying from one modality and retrieving in another (e.g., image→audio, audio→image).

B Theorem proofs

Here, we present some theoretical properties of GCCA which are useful for retrieval and why consensus based correction is an effective way to correct directional mismatch.

Theorem B.1 (Optimal multi-space alignment under squared cross-space discrepancy)**.** *Let $S \in \mathbb{R}^{(\sum_i r_i)\times(\sum_i r_i)}$ be the block matrix with diagonal blocks $S_{ii} = (M-1)I_{r_i}$ and off-diagonal blocks $S_{ij} = -U_i^\top U_j$. Then the global minimizer of Eq. 4 is given by the eigenvectors corresponding to*

the R smallest eigenvalues of S, stacked into $\Phi^\star = [\Phi_1^\star; \ldots; \Phi_M^\star]$. Moreover, the objective in Eq. 4 admits the equivalent trace form:

$$\sum_{i<j} \|U_i\Phi_i - U_j\Phi_j\|_F^2 = \mathrm{Tr}(\Phi^\top S \Phi),$$

where $\Phi = [\Phi_1; \ldots; \Phi_M]$.

Proof. Recall the objective in Eq. 4:

$$\min_{\{\Phi_i\}} \sum_{1\le i<j\le M} \|U_i\Phi_i - U_j\Phi_j\|_F^2 \quad \text{s.t.} \quad \sum_{i=1}^{M} \Phi_i^\top \Phi_i = I_R,$$

where each $U_i \in \mathbb{R}^{N\times r_i}$ has orthonormal columns ($U_i^\top U_i = I_{r_i}$). Define the stacked variable

$$\Phi := \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_M \end{bmatrix} \in \mathbb{R}^{(\sum_i r_i)\times R}.$$

First, we prove that the objective can be rewritten as a trace quadratic form. For any $i < j$,

$$\begin{aligned} \|U_i\Phi_i - U_j\Phi_j\|_F^2 &= \mathrm{Tr}\big((U_i\Phi_i - U_j\Phi_j)^\top (U_i\Phi_i - U_j\Phi_j)\big) \\ &= \mathrm{Tr}(\Phi_i^\top U_i^\top U_i \Phi_i) + \mathrm{Tr}(\Phi_j^\top U_j^\top U_j \Phi_j) \\ &\quad - 2\,\mathrm{Tr}(\Phi_i^\top U_i^\top U_j \Phi_j) \\ &= \|\Phi_i\|_F^2 + \|\Phi_j\|_F^2 - 2\,\mathrm{Tr}(\Phi_i^\top U_i^\top U_j \Phi_j), \end{aligned}$$

using $U_i^\top U_i = I_{r_i}$ and $U_j^\top U_j = I_{r_j}$.

Summing over all pairs $1 \le i < j \le M$, each term $\|\Phi_i\|_F^2$ appears exactly $(M-1)$ times, hence

$$\begin{aligned} \sum_{i<j} \|U_i\Phi_i - U_j\Phi_j\|_F^2 &= \sum_{i=1}^{M} (M-1)\|\Phi_i\|_F^2 \\ &\quad - 2\sum_{i<j} \mathrm{Tr}(\Phi_i^\top U_i^\top U_j \Phi_j). \end{aligned}$$

Now define the symmetric block matrix $S \in \mathbb{R}^{(\sum_i r_i)\times(\sum_i r_i)}$ with blocks

$$S_{ii} = (M-1)I_{r_i}, \qquad S_{ij} = -U_i^\top U_j \;\; (i \ne j).$$

A direct block expansion yields the identity

$$\mathrm{Tr}(\Phi^\top S \Phi) = \sum_{i=1}^{M} (M-1)\|\Phi_i\|_F^2 - 2\sum_{i<j} \mathrm{Tr}(\Phi_i^\top U_i^\top U_j \Phi_j),$$

which matches the expression above. Therefore,

$$\sum_{i<j} \|U_i\Phi_i - U_j\Phi_j\|_F^2 = \mathrm{Tr}(\Phi^\top S \Phi).$$

The constraint $\sum_i \Phi_i^\top \Phi_i = I_R$ is exactly $\Phi^\top \Phi = I_R$.

Now let us solve the constrained trace minimization. We have reduced Eq. 4 to

$$\min_{\Phi^\top\Phi = I_R} \mathrm{Tr}(\Phi^\top S \Phi).$$

Since S is symmetric, by the Rayleigh–Ritz/Ky Fan variational characterization, the minimum is attained when the columns of Φ span the eigenspace associated with the R smallest eigenvalues of S. Equivalently, an optimal solution $\Phi^\star$ is obtained by taking as columns the eigenvectors corresponding to the R smallest eigenvalues of S (orthonormalized), and partitioning $\Phi^\star$ into blocks $\{\Phi_i^\star\}_{i=1}^M$.

This proves that the global minimizer of Eq. 4 is given by the bottom-R eigenvectors of S, and the objective admits the trace form claimed in the theorem. □

Corollary B.2. *Let $Y_i^\star = U_i\Phi_i^\star$ denote the aligned embeddings obtained from Theorem B.1. Among all embeddings representable within the retained subspaces $\{U_i\}$ and satisfying the global orthonormality constraint, $\{Y_i^\star\}$ minimize the total cross-space mismatch energy.*

Proof. By definition, the total cross-space mismatch energy is exactly the objective in Eq. 4,

$$\sum_{1\le i<j\le M} \|Y_i - Y_j\|_F^2 = \sum_{1\le i<j\le M} \|U_i\Phi_i - U_j\Phi_j\|_F^2, \qquad Y_i = U_i\Phi_i.$$

The feasible set in Corollary B.2 coincides with the feasible set of Eq. 4, namely all collections $\{\Phi_i\}$ satisfying the global orthonormality constraint $\sum_{i=1}^M \Phi_i^\top\Phi_i = I_R$ (equivalently $\Phi^\top\Phi = I_R$ for $\Phi = [\Phi_1;\dots;\Phi_M]$). Theorem B.1 states that $\{\Phi_i^\star\}$ is a *global minimizer* of Eq. 4. Therefore, the corresponding embeddings $Y_i^\star = U_i\Phi_i^\star$ globally minimize the mismatch energy over the stated admissible class. □

Corollary B.3. *Let $G := [Y_1^\star|\dots|Y_M^\star] \in \mathbb{R}^{N\times MR}$. The matrix $B \in \mathbb{R}^{N\times R}$ formed by the top R left singular vectors of G provides an orthonormal basis for the dominant shared latent subspace induced by the aligned spaces.*

Proof. Let $G := [Y_1^\star \mid \cdots \mid Y_M^\star] \in \mathbb{R}^{N\times MR}$. We consider the optimization problem

$$\max_{B^\top B = I_R} \|B^\top G\|_F^2.$$

Observe that

$$\|B^\top G\|_F^2 = \mathrm{Tr}\big((B^\top G)(B^\top G)^\top\big) = \mathrm{Tr}\big(B^\top GG^\top B\big).$$

Define the symmetric positive semidefinite matrix $A := GG^\top \succeq 0$. Then the problem becomes

$$\max_{B^\top B = I_R} \mathrm{Tr}(B^\top AB).$$

By the Ky Fan maximum principle (equivalently the variational characterization of the sum of the top eigenvalues), the maximizer B is given by the eigenvectors of A associated with its largest R eigenvalues. Since the eigenvectors of $GG^\top$ are precisely the left singular vectors of G, the solution is the matrix formed by the top R left singular vectors of G. □

Corollary B.4. *The corresponding linear maps from the original feature spaces into the shared latent space are given by $Q_i := V_i\Phi_i^\star$, so that for any feature vector $z \in \mathbb{R}^{d_i}$, the embedding zQ_i realizes the same aligned coordinates as the sample-side embedding $U_i\Phi_i^\star$, up to the scaling induced by the SVD convention. Equivalently, defining $Q_i := V_i\Sigma_i^{-1}\Phi_i^\star$ yields $X_iQ_i = U_i\Phi_i^\star$ exactly.*

Proof. Let $X_i = U_i\Sigma_iV_i^\top$ be a (possibly truncated) SVD with $U_i^\top U_i = I$ and $V_i^\top V_i = I$. First consider the map $\widetilde{Q}_i := V_i\Sigma_i^{-1}\Phi_i^\star$ (assuming Σ_i is invertible on the retained rank). Then

$$X_i\widetilde{Q}_i = (U_i\Sigma_iV_i^\top)(V_i\Sigma_i^{-1}\Phi_i^\star) = U_i\Phi_i^\star = Y_i^\star,$$

which shows that $\widetilde{Q}_i$ yields an *exact* feature-space realization of the aligned sample embedding.

For the map $Q_i := V_i\Phi_i^\star$, we similarly obtain

$$X_iQ_i = (U_i\Sigma_iV_i^\top)(V_i\Phi_i^\star) = U_i\Sigma_i\Phi_i^\star.$$

Thus, Q_i realizes the same aligned coordinates up to the singular-value scaling induced by the SVD convention (i.e., absorbing Σ_i into the coordinates). This is the stated "up to scaling" relationship. □

Proof of Proposition 3.2. For unit vectors $\{\widehat{u}_{m,i}\}_{m=1}^{M}$, let $s := \sum_{m=1}^{M} \widehat{u}_{m,i}$ and $c_i = s/\|s\|_2$. Then

$$\frac{1}{M}\sum_{m=1}^{M}\langle \widehat{u}_{m,i}, c_i\rangle = \frac{1}{M}\left\langle \sum_{m=1}^{M}\widehat{u}_{m,i}, \frac{s}{\|s\|_2}\right\rangle = \frac{1}{M}\|s\|_2,$$

which gives Eq. equation 6. Moreover,

$$\|s\|_2^2 = \sum_{m=1}^{M}\|\widehat{u}_{m,i}\|_2^2 + 2\sum_{m<n}\langle \widehat{u}_{m,i}, \widehat{u}_{n,i}\rangle = M + 2\sum_{m<n}\langle \widehat{u}_{m,i}, \widehat{u}_{n,i}\rangle,$$

which rearranges to Eq. equation 7. □

C Ablations

C.1 Hyperparameters

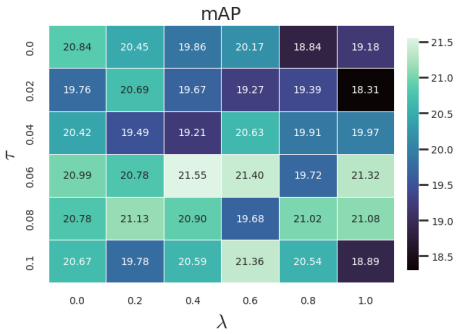

Figure 9: **Sensitivity of GCPA to the trust penalty on cross-camera retrieval.** We sweep the trust-region parameters (τ, λ) and report cross-camera mAP (%, higher is better).

The trust penalty in GCPA acts as a soft trust-region as it allows the corrector to nudge a point toward the per-sample consensus direction, while discouraging excessive deviation from the original GPA universe direction. In our implementation, τ specifies a tolerance on drift (no penalty is incurred below this threshold), and λ controls the strength of the penalty once the drift exceeds τ.

Rather than tuning these parameters per benchmark, we fix a single conservative (τ, λ) for all experiments. The intent is to allow mild corrections (so retrieval can benefit from consensus nudging) while maintaining a clear notion of geometric trust in the underlying GPA universe. To make the effect of these parameters transparent, we report a sensitivity sweep on Market-1501 in Figure 9. It can be seen that there are no abrupt performance differences and mild settings can help in achieving a reasonable retrieval performance.

C.2 Geometry changes on nudging

Figure 10: **Sensitivity of GCPA to the trust penalty on drift.** We sweep the trust-region parameters (τ, λ) and report median drift (%, lower is better). The top-right corner consists of higher λ and lower τ and therefore has lower drift values.

GCPA improves retrieval by applying a small correction in the GPA universe, which intentionally relaxes strict isometry. To make this trade-off explicit, we quantify how much the correction changes the geometry of the universe embeddings.

Let $\hat{u}$ denote the row-wise ℓ_2-normalized universe embedding produced by GPA, and let $\tilde{u}$ be the corresponding corrected embedding after applying the GCPA update (and renormalization).

Specifically, we measure drift as the angular deviation between $\tilde{u}$ and $\hat{u}$, using $\mathbb{E}[1 - \langle \tilde{u}, \hat{u}\rangle]$.

Figure 10 summarizes the drift percentage for different values of λ and τ providing insights on how increasing λ results in a decrease in drift whereas increase τ results in an increase.

Algorithm 1 GCCA for M-way shared-basis alignment

Require: Matched representations $\{X_m \in \mathbb{R}^{N \times d_m}\}_{m=1}^{M}$, target rank R
Ensure: Space-to-shared-basis maps $\{Q_m \in \mathbb{R}^{d_m \times R}\}_{m=1}^{M}$

1: For each space m, compute a (possibly truncated) thin SVD:

$$X_m = U_m \Sigma_m V_m^\top.$$

2: Let r_m denote the retained rank of the thin SVD for space m (so $\Sigma_m \in \mathbb{R}^{r_m \times r_m}$ is invertible).

3: Form the block matrix $S \in \mathbb{R}^{(\sum_m r_m) \times (\sum_m r_m)}$ with blocks

$$S_{mm} = (M-1) I_{r_m}, \qquad S_{mn} = -U_m^\top U_n \;\; (m \neq n).$$

4: Compute the R eigenvectors of S with smallest eigenvalues and stack them as

$$\Phi^\star = \begin{bmatrix} \Phi_1^\star \\ \vdots \\ \Phi_M^\star \end{bmatrix}, \qquad \Phi_m^\star \in \mathbb{R}^{r_m \times R}.$$

5: Define the shared-basis embeddings for samples in each space:

$$Y_m := U_m \Phi_m^\star \in \mathbb{R}^{N \times R}.$$

6: Define the feature-side maps into the shared basis:

$$Q_m := V_m \Sigma_m^{-1} \Phi_m^\star, \quad \text{so that} \quad X_m Q_m = Y_m.$$

7: **return** $\{Q_m\}_{m=1}^{M}$

Overall, GCPA introduces a controlled amount of drift and this is expected as unlike GPA, which is exactly isometric, GCPA is designed to allow limited geometric deformation when it increases cross-view agreement and improves retrieval. We fit the alignment using three different text encoders for this computation.

D Shared-basis alignment module (GCCA) algorithm

We provide a compact algorithmic summary of the shared-basis alignment used in our experiments.

E Geometry correction (GCPA) algorithm

This section specifies the universe-level geometry correction used by **GCPA**. We first fit an M-way orthogonal universe using GPA, obtaining per-space rotations $\{\Omega_m\}_{m=1}^{M}$ and universe embeddings $X_m^\star = X_m \Omega_m \in \mathbb{R}^{N \times d}$. GCPA then applies a shared correction in universe coordinates that encourages agreement across spaces while keeping the corrected directions close to the GPA universe.

Consensus target. Let $u_{m,i}$ denote the ith row of $X_m^\star$. We form per-space unit directions $\hat{u}_{m,i} = u_{m,i}/\|u_{m,i}\|_2$ and define a per-sample consensus direction

$$c_i = \mathrm{norm}\left(\frac{1}{M}\sum_{m=1}^{M} \hat{u}_{m,i}\right), \qquad C = [c_1; \ldots; c_N] \in \mathbb{R}^{N \times d}, \tag{10}$$

where $\mathrm{norm}(\cdot)$ denotes row-wise ℓ_2 normalization.

Correction module. We learn a shared residual map $T_\theta : \mathbb{R}^d \to \mathbb{R}^d$ applied to row-normalized universe embeddings. It is initialized near the identity and outputs a row-normalized corrected direction,

$$T_\theta(u) = \mathrm{norm}\big(\hat{u} + \Delta_\theta(\hat{u})\big), \qquad \hat{u} = \mathrm{norm}(u), \tag{11}$$

where Δ_θ is a small MLP shared across all spaces.

Training objective with trust penalty. We train T_θ to match the consensus direction c_i from universe vectors $u_{m,i}$ sampled across spaces and indices. For a sampled pair $(u_{m,i}, c_i)$, let $y = T_\theta(u_{m,i})$ and $\hat{u} = \mathrm{norm}(u_{m,i})$. We use the cosine loss

$$\ell_{\mathrm{align}}(y, c_i) = 1 - \langle y, c_i \rangle,$$

and measure deviation from the original GPA direction by

$$\ell_{\mathrm{drift}}(y, \hat{u}) = 1 - \langle y, \hat{u} \rangle.$$

The trust penalty activates only when drift exceeds a tolerance τ:

$$\ell_{\mathrm{trust}}(y, \hat{u}) = \max\{0, \ell_{\mathrm{drift}}(y, \hat{u}) - \tau\}. \tag{12}$$

The final objective for a minibatch is

$$\mathcal{L}_{\mathrm{GCPA}} = \mathbb{E}\big[\ell_{\mathrm{align}}\big] \; + \; \lambda\, \mathbb{E}\big[\ell_{\mathrm{trust}}\big], \tag{13}$$

with fixed (τ, λ) across all benchmarks.

Optimization. We optimize Eq. equation 13 on the training split using a standard minibatch procedure.

Using the corrector at inference. For a source space m, the universe map becomes

$$\mathrm{toU}(x; m) = \begin{cases} x\Omega_m, & \text{GPA} \\ T_\theta(x\Omega_m), & \text{GCPA}. \end{cases} \tag{14}$$

Mapping back into a target space n is $\mathrm{fromU}(u; n) = u\Omega_n^\top$. Because T_θ is not restricted to be orthogonal, strict orthogonal cycle consistency need not hold after correction; we therefore analyze the correction separately (Appendix C).

F Visualisation of Structure Differences

Figure 11 provides a qualitative view of how alignment changes class structure. We apply UMAP to embeddings in each aligned space and visualize the resulting class clusters. Consistent with the quantitative results in Table 1, GCPA produces more clearly separated class-specific clusters, reflecting a controlled deformation of the universe that improves cross-space comparability. In contrast, GPA preserves within-space geometry by construction, while GCCA emphasizes shared directions through a learned projection; both often yield cluster layouts that remain closer to the original structure and can exhibit more overlap between classes.

G LLM Usage

LLMs were used to help with writing-related tasks (e.g., improving clarity, grammar, and phrasing) and to assist in designing parts of our visualizations (e.g., suggesting layouts and rephrasing captions). All technical content, experimental design, results, and conclusions are the authors' own. The authors reviewed and verified the final manuscript and figures, including any LLM-assisted text, and ensured that all claims and references are accurate.

H Future Directions

A natural extension of this work is to explore more expressive alignment regimes to construct the shared universe. While this study focused on linear and near-linear alignment, future work could investigate M-way functional alignment to capture non-isometric correspondences inherent in highly heterogeneous model collections. Furthermore, the ability of the universe to "heal" weak links between disparate models may indicate applications in decentralized learning and model merging, where maintaining a coherent global state from fragmented views is critical.

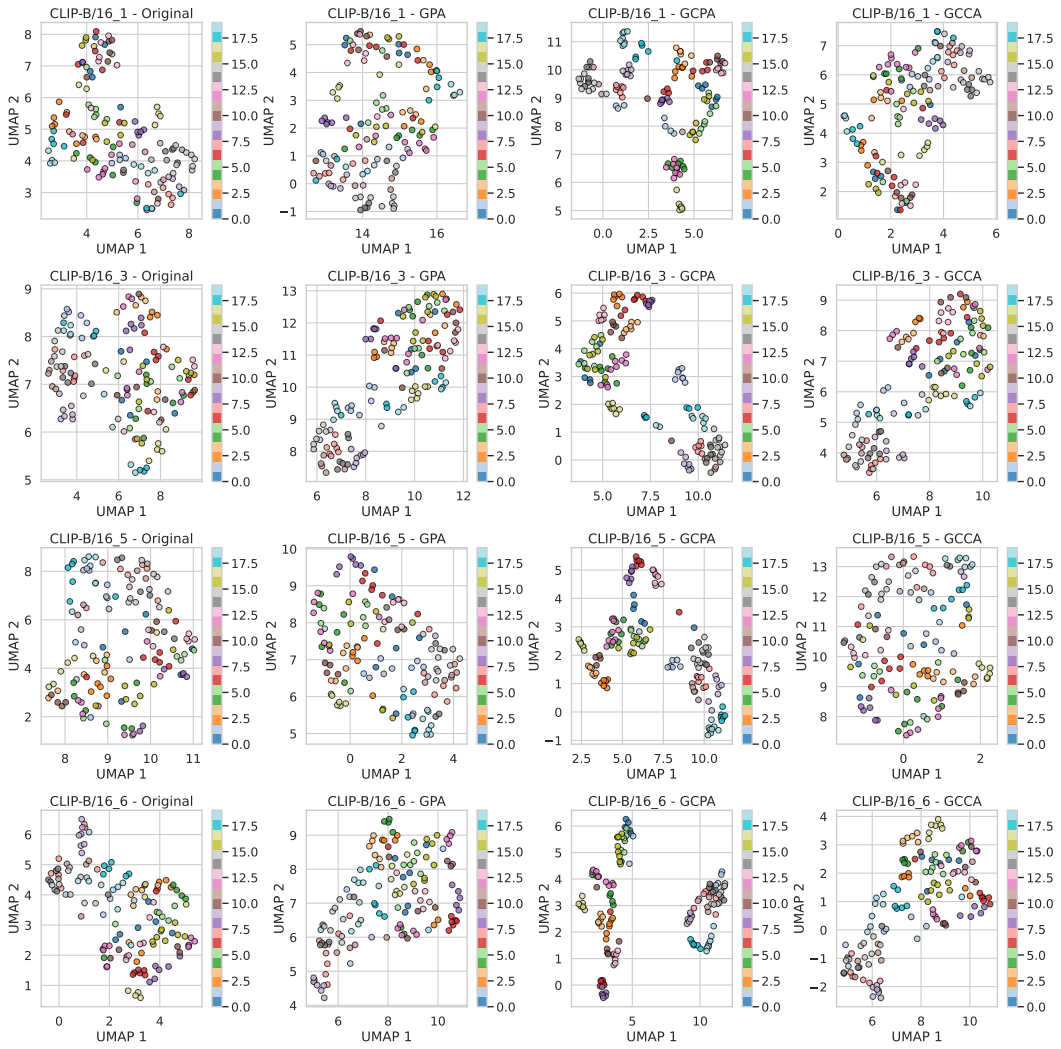

Figure 11: **UMAP representation of images from 20 representative classes from the Market-1501 dataset.** The first column computes distances between images directly in the original model space, while the others are the universal spaces made using GPA, GCPA, and GCCA. Each line of subplots is made using input from a different camera. Dots with the same colors represent images from the same classes.