BARYCENTRIC ALIGNMENT FOR STUDYING VARIATION IN REPRESENTATIONAL CONVERGENCE ACROSS INPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparing representations across neural networks is challenging because representations admit symmetries, such as arbitrary reordering of units or rotations of activation space, that obscure underlying equivalence between models. We introduce a barycentric alignment framework based on the Generalized Procrustes Algorithm that quotients out these nuisance symmetries to construct a universal embedding space across many models. Unlike existing similarity measures, which summarize relationships over entire stimulus sets, this framework enables similarity to be defined at the level of individual stimuli, revealing inputs that elicit convergent versus divergent representations across models. Using this instance-level notion of similarity, we identify input properties that predict representational convergence versus divergence across vision and language model families. We also construct universal embedding spaces for brain representations across individuals and cortical regions, enabling instance-level comparison of representational agreement across stages of the human visual hierarchy. Finally, we apply the barycentric alignment framework to purely unimodal vision and language models and find that post-hoc alignment into a shared space yields image–text similarity scores that closely track human cross-modal judgments and approach the performance of contrastively trained vision-language models. This strikingly suggests that independently learned representations already share sufficient geometric structure for human-aligned cross-modal comparison. Together, these results show that resolving representational similarity at the level of individual stimuli reveals phenomena that cannot be detected by set-level comparison metrics.

1 INTRODUCTION

Across domains, deep neural networks map stimuli into high-dimensional internal representations that support perception, reasoning, and action. Despite their central role, our understanding of these representations remains limited. Most analyses probe individual models or narrow architectural families, identifying patterns or mechanisms that explain behavior in a specific system. This leaves open a fundamental scientific question: when a representational phenomenon is observed in a single network, which aspects reflect idiosyncratic implementation choices, and which reflect shared representational structure that persists across architectures, training regimes, scales, or even modalities? As models continue to diversify, the field lacks principled tools for determining where universality ends and model-specificity begins in neural representations.

Addressing this question requires tools for comparing representations across models, rather than explaining them in isolation. A substantial body of prior work has therefore focused on representational similarity analysis, introducing measures such as RSA (Kriegeskorte et al.), CKA (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), and related metrics to quantify alignment between neural networks (Yamins et al., 2014; Laakso & Cottrell, 2022; Li et al., 2015; Morcos et al., 2018; Wang et al., 2018). These approaches have yielded important insights into how models relate with one another, revealing broad trends across architectures, training regimes, and layers. However, they fundamentally characterize representational similarity as a set-level property: a single score that summarizes how an entire collection of stimuli is arranged in two representation spaces. In doing so, they obscure how agreement and disagreement are distributed across the input space. This

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

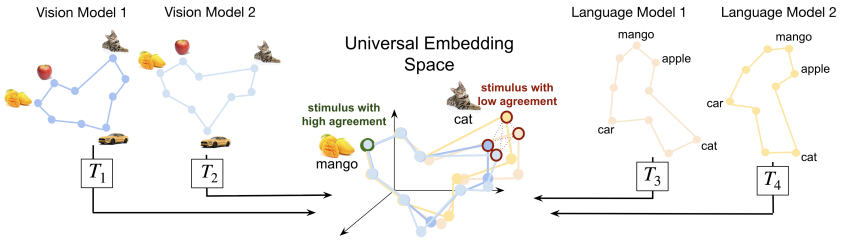


Figure 1: Barycentric alignment into a universal embedding space. Representations from a diverse set of models, spanning modalities, architectures, training objectives, and scales, are aligned via a barycentric procedure into a shared embedding space by factoring out nuisance symmetries.

limitation precludes a finer-grained view of representational structure - one that treats alignment as an instance-level phenomenon and examines its variation across stimuli and model families. Such a view is essential for characterizing the *ecology of representational alignment*: which inputs induce consensus and which expose divergence. Moreover, because most existing approaches operate pairwise, they do not provide a principled way to align or analyze representations from an entire population of models simultaneously. As a result, we currently lack tools for defining stimulus-level notions of agreement that are stable, interpretable, and comparable across large model ensembles.

A small number of recent studies have begun to move beyond purely aggregate comparison, but important limitations remain (e.g., (Hosseini et al., 2024; Kolling et al.)). Hosseini et al. partition a fixed stimulus set into "high-agreement" and "low-agreement" subsets and show that stimuli on which artificial models disagree most strongly are also those on which models and human brain representations diverge. However, whether a particular stimulus is labeled "agreeable" depends on the composition of the evaluation set: the same item can appear highly consistent or highly divergent depending on which other stimuli are present. Agreement therefore reflects not only intrinsic properties of a stimulus, but also contextual contrasts induced by the stimulus pool, undermining a stimulus-specific interpretation. Pointwise Normalised Kernel Alignment (PNKA; (Kolling et al.)) takes an important step toward finer granularity by assigning scores to individual stimuli. Yet PNKA still defines similarity relative to the global neighborhood structure of the evaluation set and remains inherently pairwise, with scores tied to a specific pair of models. As a result, extending the analysis to larger model populations or to new stimuli requires recomputing kernels, limiting its utility for studying representational agreement at the level of model ensembles.

More broadly, these limitations mean that the field lacks a systematic account of how representational agreement is distributed across inputs. As a result, the ecology of representational similarity, particularly how agreement and disagreement are structured across stimuli, how they depend on input properties, and how they vary across model populations, remains surprisingly underexplored.

Here we introduce a Generalized Procrustes Analysis (GPA)-style barycentric framework for representational alignment that treats symmetry-quotienting as a first-class design principle, with Procrustes barycenters providing a natural solution for rotational invariances (Gower, 1975; Ten Berge, 1977). Our framework is related to recent latent-space alignment approaches (Moschella et al., 2023; Maiorca et al., 2023; Norelli et al., 2023), but differs in constructing a single explicit shared space across a population of models rather than relying on anchors or pairwise translations. This barycentric alignment enables a stimulus-specific measure of representational agreement, defined by how tightly a stimulus's representations cluster across models in the shared space. Unlike prior approaches, this instance-wise measure is independent of the surrounding stimulus set at inference time and well-defined for populations of models rather than model pairs. Using this framework, we characterize which inputs elicit convergent versus divergent representations across vision and language models, construct universal embedding spaces for brain representations across individuals and regions, and show that post-hoc alignment of unimodal vision and language models recovers meaningful image–text similarity. Together, these results demonstrate how instance-level analysis reveals structure in representational similarity that is inaccessible to set-level comparison metrics.

2 METHOD

We propose a *barycentric alignment* framework for representational analysis (Figure 1). We treat each model's representation of the same stimulus set as a point cloud whose coordinates are only identifiable up to transformations induced by architectural and training choices (e.g., rotations). Using a *GPA-style* barycentric alignment procedure, we align all models to a common *barycenter* under an appropriate invariance class, obtaining a universal space in which representations become directly comparable across models. Building on this shared space, we define an *instance-level representational consistency* measure that quantifies how consistently a stimulus is represented across models.

2.1 SYMMETRIES AND THE CHOICE OF ALIGNMENT

We treat two representations as equivalent if they are related by a transformation from a specified group $\mathcal{G}$. Formally, let $X_i \in \mathbb{R}^{n \times d}$ denote the representation matrix from model i, where rows correspond to n stimuli and columns to d features. We consider $X_i \sim X_j$ if and only if there exists a transformation $T \in \mathcal{G}$ such that X_i can be approximated well with $X_j T$.

The choice of $\mathcal{G}$ reflects assumptions about which transformations preserve meaningful representational content. Permutation invariance ($\mathcal{G} = \mathcal{P}(d)$, the group of permutation matrices) captures the intuition that neuron ordering is arbitrary. Orthogonal invariance ($\mathcal{G} = O(d)$, the orthogonal group) additionally accounts for rotations and reflections of the activation space. Throughout this work, we adopt orthogonal invariance as our default equivalence relation, as it provides a principled middle ground: it quotients out rotational degrees of freedom while preserving the metric structure of the representation space. Critically, the orthogonal group consists of linear isometries, i.e., transformations that preserve Euclidean distances. This ensures that aligning representations via orthogonal transformations does not distort the geometric relationships among stimuli within each model's space.

2.2 BARYCENTRIC ALIGNMENT VIA GPA-STYLE PROCRUSTES ITERATION

Generalized Procrustes Analysis (GPA) (Gower, 1975; Ten Berge, 1977) is a classical method for aligning multiple matched configurations into a common reference frame under a specified transformation group. In our setting, we use a GPA-style orthogonal alignment procedure to estimate a shared barycenter of neural representations while factoring out rotational symmetry. Given N models, each producing representations for the same set of n training stimuli, we seek a common embedding space and a set of alignment transformations $\{T_i\}_{i=1}^N$ such that the aligned representations $\{X_i T_i\}$ are mutually consistent. We formulate this as the problem of computing the *Procrustes barycenter* of the N representation matrices, i.e., the mean shape in the quotient space induced by orthogonal equivalence. To ensure consistent dimensionality across models, we set $d = \max(d_1, \ldots, d_N)$, where d_i denotes the original dimensionality of model i. Representations from models with $d_i < d$ are zero-padded to match the maximum dimensionality. We initialize the template $M^{(0)}$ as the arithmetic mean of the representations and then iterate the following procedure:

Alignment step: For each model i, compute the orthogonal Procrustes transformation $T_i = \arg\min_{R \in O(d)} \|X_i R - M^{(t)}\|_F$, which admits a closed-form solution via the singular value decomposition.

Update step: Set $X_i^{(t)} \leftarrow X_i T_i$ and recompute the template as $M^{(t+1)} = \frac{1}{N} \sum_{i=1}^N X_i^{(t)}$.

This alternating procedure converges to a local minimum of the sum of squared Procrustes distances, yielding aligned representations that reside in a shared coordinate system. We stop early when $\|M^{(t+1)} - M^{(t)}\|_F / \|M^{(t)}\|_F < \varepsilon$. The resulting template M can be interpreted as the Fréchet mean of the N representations under the Procrustes metric. Its barycenter coordinates constitute what we refer to as the *universal embedding space*, and $\{T_i\}$ provide model-specific maps into that space. Algorithm A summarizes the training phase. The n training stimuli used to learn the alignment transformations are held separate from the test stimuli used for inference.

Table 1: Averaged Cross-model correlation, Procrustes Residual, Heldout Residual, RMS score, and stimulus retrieval accuracy for the different vision model pools and language model pools.

Model Pool	Corr. ↑	Procrustes Res. ↓	Heldout Res. ↓	RMS ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑
Self Supervised	0.5970	0.2763	0.2840	0.0178	0.9740	0.9957	0.9986
ResNet	0.7099	0.2158	0.2188	0.0159	0.9929	0.9999	1.0000
ConvNext	0.6824	0.2738	0.2779	0.0180	0.9558	0.9976	0.9994
SWIN Transformers	0.4903	0.1564	0.1587	0.0213	0.9785	0.9987	0.9999
VIT	0.4068	0.3817	0.3868	0.0304	0.8206	0.9414	0.9634
VGG	0.6795	0.2117	0.2375	0.0124	0.9899	0.9999	1.0000
Supervised CNNs	0.5897	0.3299	0.3445	0.0133	0.9555	0.9964	0.9989
Supervised Transformers	0.3678	0.4088	0.4148	0.0298	0.8141	0.9478	0.9716
Big Model	0.5017	0.4280	0.4369	0.0136	0.7734	0.9443	0.9718
Small Model	0.4690	0.4489	0.4570	0.0132	0.8221	0.9584	0.9785
Misc1	0.4901	0.3419	0.3495	0.0200	0.8828	0.9851	0.9959
Misc2	0.4441	0.3733	0.3868	0.0144	0.8559	0.9706	0.9886
Misc3	0.3782	0.3739	0.3757	0.0146	0.7677	0.9416	0.9756
Supervised	0.4864	0.4595	0.4670	0.0138	0.8094	0.9602	0.9819
OpenLLaMA	0.5699	0.0489	0.3157	0.0113	0.9594	0.9750	0.9772
GPT-2	0.4778	0.1302	0.2482	0.0174	0.6753	0.8372	0.8889
Qwen	0.4358	0.3151	0.5124	0.0092	0.7358	0.8372	0.8730
ALBERT	0.3289	0.3592	0.7745	0.0154	0.6450	0.6997	0.7231
Autoregressive	0.5148	0.2250	0.5738	0.0107	0.8497	0.9018	0.9224

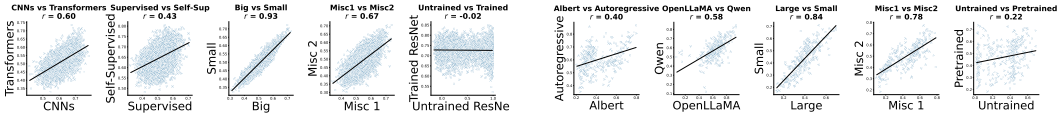
2.3 Inference and stimulus-wise agreement

The shared universal space enables comparison of representations at the level of individual stimuli across an entire population of models. Given a new set of m test stimuli, each model produces representations $Y_i \in \mathbb{R}^{m \times d}, \quad i = 1, \ldots, N$, which we project into the universal space via the learned transforms: $Y_i' = Y_i T_i$. For each test stimulus $j \in \{1, \ldots, m\}$, we then quantify representational agreement by averaging pairwise similarity across the aligned representations: $S_j = \frac{1}{N(N-1)} \sum_{p \neq q} \mathrm{SIM}\big(Y'_{pj}, Y'_{qj}\big)$, where Y'_{pj} denotes the row of Y'_p corresponding to stimulus j. The similarity function is left generic; unless otherwise stated, we use cosine similarity throughout. Throughout, we refer to this stimulus-wise representational agreement as the *instance-level representational consistency score*, which quantifies the extent to which a stimulus admits a consistent representation across models after quotienting out nuisance symmetries.

We evaluate the quality of the learned universal embedding by measuring cross-model representational alignment after projection into the shared space. Alignment is quantified using four complementary metrics: cross-model correlation, root-mean-square (RMS) discrepancy and Procrustes residuals, all computed per embedding dimension across stimuli and then averaged, and cross-model stimulus retrieval accuracy, which measures whether a stimulus in one model maps nearest to the same stimulus in another. See Appendix Section A.1 for formal definitions and implementation details.

3 Results

3.1 Quantitative evaluation of barycentric alignment



(a) (b)

Figure 2: **Instance-level representational consistency is conserved across model pool constructions.** Pairwise correlations of instance-level representational consistency scores computed from different model pools: (a) vision models and (b) language models, defined by different criteria (architectural family, scale, random mixtures).

Alignment quality across model pools. We first assess the learned universal embedding space as a diagnostic of how much representational variation across models can be reduced to nuisance symmetries. If differences between models are largely attributable to symmetry transformations (e.g., rotations of feature space), barycentric alignment should bring corresponding representations into

close correspondence in the shared space. Table 1 reports pool-level averaged evaluation metrics for multiple vision and language model pools. Retrieval provides the clearest evidence of successful alignment: top-k stimulus retrieval is consistently high, especially in homogeneous vision pools, and remains far above chance in heterogeneous settings. Correlation and RMS show the same overall pattern. Procrustes residuals provide a complementary fit diagnostic: they are smallest for homogeneous vision pools and increase with heterogeneity, whereas language pools show a more variable pattern, especially in held-out residuals. Corresponding per-model results for representative pools are provided in Appendix Table 15 and show the same qualitative trends. Overall, these results show that, after quotienting out orthogonal symmetries, representations from diverse architectures can be brought into meaningful correspondence at the level of individual stimuli. Alignment is strongest for small or homogeneous pools, as expected, but remains substantial even for heterogeneous pools spanning architectures, training objectives, and model scales. This suggests that much of the inter-model variability is symmetry-reducible.

Robustness to baseline choice. We additionally compare barycentric alignment against several baselines (Appendix Section A.5.1, Figure 6–7). Barycentric alignment substantially outperforms anchor-based approaches (Moschella et al., 2023), while approaching the performance of independently optimized pairwise Procrustes alignment (Maiorca et al., 2023). Thus, a single shared-space construction recovers much of the alignment quality achievable by pairwise fits while additionally supporting population-level instance-wise analysis in one explicit common coordinate system.

Stability of instance-level scores across pool construction. Consistency scores are highly correlated across vision model pools varying in architecture, training objective, model size, and mixed pool conposition (Figure 2a; see Appendix Figure 10 for all pairwise comparisons). Thus, the same stimuli tend to drive disagreement across very different model classes, suggesting that instance-level disagreement reflects shared axes of representational variation rather than idiosyncrasies of a particular architecture or training regime. Consistency also depends on learned model structure: when the pool is composed of randomly initialized networks, the resulting scores are largely uncorrelated with those from trained models (Figure 2a; Appendix Figure 11). For language model pools (Figure 2b), instance-level consistency is also broadly shared across pool constructions, but correlations vary more than in vision: they are moderate across substantially different architectures and model families, and higher across pools with more similar composition or size. Untrained versus trained comparisons are weakly positive, suggesting that a small scaffold of instance-level structure may already be present from initialization or tokenization statistics, although training substantially reshapes it.

From a geometric perspective, these results indicate that the *alignability* of representations (how well different systems can be brought into correspondence after quotienting by symmetries) is structured over the input space. Some stimuli occupy stable equivalence classes across models, while others expose irreducible differences in representations.

3.2 Instance-level representational consistency tracks stimulus difficulty

We next use the instance-level representational consistency score as a stimulus-level diagnostic, asking how it relates to interpretable properties of individual inputs.

3.2.1 Vision

We first examine how instance-level representational consistency varies across natural images. Ranking ImageNet validation images by their consistency scores reveals a striking qualitative structure (Figure 3A, Appendix Figure 16): images with high consistency are visually unambiguous and strongly aligned with their nominal categories, whereas low-consistency images tend to be perceptually ambiguous, cluttered, or atypical, even for human observers. This qualitative pattern suggests that instance-level representational consistency may reflect stimulus difficulty.

To formalize this relationship, we focus on consistency scores computed from the "Supervised" vision model pool and relate them to established computational measures of stimulus difficulty. We first consider *prototypicality* which quantifies how representative an image is of its semantic category, with more prototypical images lying closer to the category centroid in feature space. Instance-level consistency is strongly correlated with prototypicality (Figure 3C, left), indicating that images that are more visually canonical also elicit more convergent representations across models. This

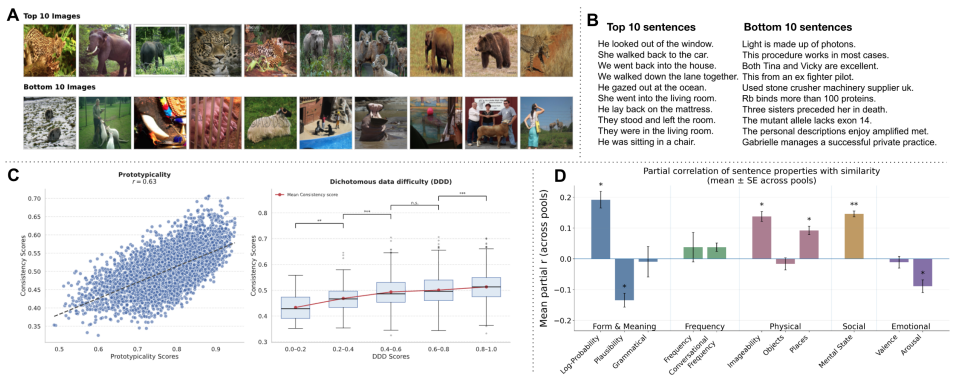


Figure 3: **(A)-(B)** Representative images from the Animals parent category with the highest and lowest instance-level representational consistency scores, derived from a pooled ensemble of supervised models with varied architectures and scales. **(C)** Instance-level representational consistency as a function of computational difficulty metrics (Prototypicality and DDD). Paired t-tests across DDD bins show significant increases in consistency scores. **(D)** Partial correlations between sentence properties and instance-level representational consistency. Bars show the mean partial correlation r across k=4 language-model pools; error bars indicate ±1 SE across pools.

correspondence suggests that representational consensus concentrates on stimuli that occupy stable, central regions of category structure.

We next relate instance-level representational consistency to *dichotomous data difficulty (DDD)*, an agreement-based measure that quantifies how many models assign the same label to a stimulus, irrespective of correctness. Consistency scores increases systematically with DDD (Figure 3C, right), demonstrating that stimuli on which models agree at the level of decision outputs are also those on which their internal representations align most closely. This correspondence indicates that instance-level representational consistency captures a shared notion of stimulus difficulty that is reflected both in models' internal representations and in their downstream decisions, clarifying how population-level behavioral agreement emerges from shared structure in representation space.

Finally, we investigated the relationship between instance-level representational consistency and human perceptual judgments from the THINGS dataset (Hebart et al., 2023), including memorability, nameability, and recognizability. Memorability reflects the likelihood that an image is remembered after brief exposure, nameability quantifies the consistency of object labels assigned by human observers, and recognizability measures how easily an object can be correctly identified. We found that instance-level representational consistency shows no significant relationship with any of these measures (Appendix Figure 17). Overall, these results suggest that instance-level representational consistency is more closely tied to difficulty as it arises for a population of models than to human perceptual experience.

3.2.2 Language

We next investigate diverse language model pools using the curated six-word stimulus set described in Section A.2.2. We compute a consistency score for each sentence and analyze both qualitative exemplars at the extremes and quantitative relationships with annotated linguistic and semantic properties. Sentences with the highest consistency scores (Figure 3B, left) are conventional and scene-based, describing everyday physical actions and locations using simple syntax and high-frequency lexical items (e.g., walked back, went into the living room, sitting in a chair). In contrast, sentences with the lowest consistency scores (Figure 3B, right) tend to be stylistically and semantically less homogeneous, including technical/scientific statements, fragment-like constructions, and web/SEO-like artifacts, which introduce rarer tokens, abbreviations, and domain-specific terminology.

To characterize this structure more formally, we computed both marginal and partial correlations between sentence-level consistency and a suite of lexical and human-rated norms. Because all measured sentence properties show positive marginal associations with consistency (Appendix Figure 18), we use the partial analysis to identify which properties explain unique variance after ac-

counting for the others (Figure 3D). The strongest independent predictor is distributional predictability: consistency increases with log-probability (mean partial $r = 0.19$), showing that sentences more expected under language statistics elicit more convergent representations across model pools even after accounting for frequency and semantic factors. By contrast, frequency shows only weak partial associations (general and conversational frequency: mean partial $r \approx 0.04$), suggesting that its marginal relationship with consistency is largely explained by shared variance with predictability.

Beyond predictability, consistency is also positively related to grounded semantic content, with positive partial correlations for imageability (mean partial $r = 0.14$) and place-related content (mean partial $r = 0.09$). Social-cognitive content shows a similar pattern (mental state: mean partial $r = 0.15$), whereas emotional dimensions are weaker (valence: mean partial $r \approx 0$; arousal: mean partial $r = -0.09$). Plausibility shows a negative partial association (mean partial $r = -0.13$), indicating a residual effect after controlling for predictability and the other correlated norms.

To complement the sentence-level analysis, we also performed a word-level evaluation using Lancaster sensorimotor norms (Lynott et al., 2020) (Appendix Figure 19). This yields a clearer characterization pattern: instance-level consistency is systematically related to grounded semantic content, with especially strong effects for visual and action-related dimensions. Together, these results suggest that cross-model representational agreement in language is strongest for inputs that are distributionally predictable and rich in physical, visual, action-related, and social-cognitive content.

3.3 Cross-modal barycentric alignment

Recent work on the Platonic Representation Hypothesis (Huh et al., 2024) has proposed that sufficiently large models trained on different modalities may converge on a shared representational structure, as evidenced by increasing overlap in local neighborhoods across representation spaces as model scale grows. This result suggests a form of cross-modal convergence at scale, but leaves open a critical question: to what extent are differences between vision and language models reducible to symmetry transformations, such that their representations admit a joint alignment to a common reference space?

To address this question, we extend our barycentric alignment framework to a cross-modal setting by jointly aligning a pool of vision-only and language-only models into a shared quotient space defined by a common barycenter. This approach can be contrasted with learned cross-modal models such as CLIP (Radford et al., 2021), which acquire a shared image–text space through large-scale paired supervision during representation learning. In contrast, our approach does not learn a new cross-modal encoder; instead, it makes explicit how much cross-modal structure is already present in pretrained unimodal representations and can be revealed through symmetry-aware alignment.

We constructed a joint universal embedding space using a pool of vision transformers (ViTs) and language models (LLMs). For each MS-COCO image–caption pair, we extracted (i) image embeddings from each vision model and (ii) caption embeddings from each language model. We then learned a set of model-specific alignment transforms that map all representations from both modalities into a common universal space (Figure 4a).

3.3.1 Cross-modal Barycentric Alignment Quality and Retrieval

Overall, the universal space yields retrieval performance comparable to CLIP across retrieval depths (Figure 4b). In the text→image direction, our aligned universal space slightly outperforms CLIP at top-1 retrieval (54.0% vs. 52.5%) while remaining close at top-5 (87.5% vs. 82.5%) and top-10 (97.0% vs. 94.5%). In the image→text direction, CLIP remains higher at each k (top-1: 58.0% vs. 41.5%; top-5: 89.5% vs. 77.0%; top-10: 99.0% vs. 93.5%), but the universal space still achieves strong retrieval accuracy despite relying on pretrained unimodal models and using paired data only to estimate alignment transforms rather than to learn representations. These findings reveal that unimodal vision and language models already encode, implicitly and independently, much of the structure needed for cross-modal matching. Barycentric alignment exposes this latent compatibility—achieving near-CLIP performance without contrastive pretraining.

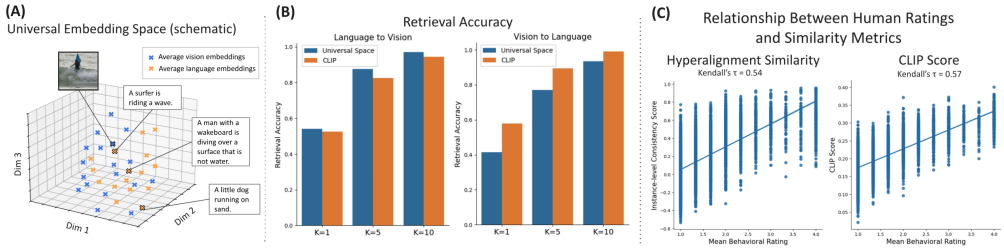


Figure 4: (A) Universal embedding space aligning vision-only image embeddings (blue) and language-only caption embeddings (orange), bringing matched image–caption pairs close together. (B) Bidirectional MS-COCO retrieval accuracy for the universal space and CLIP (text→image and image→text; top-k, $k \in 1, 5, 10$). (C) Flickr8k-Expert comparison: unimodal-pool sentence-level consistency (left) and CLIP image–text similarity (right) versus mean human ratings. Models in the unimodal pool: OpenLLaMA-7b, OpenLLaMA-13b, ViT-L/14 DINOv2, and ViT-g/14 DINOv2.

3.3.2 Cross-Modal Consistency Scores Correlate with Human Judgements

To assess whether instance-level representational consistency scores capture human-perceived image–caption matching beyond retrieval, we evaluated the resulting compatibility scores on Flickr8k-Expert, which provides human ratings for image–caption pairs. We computed the similarity between an image and caption as the cosine similarity between their projections in the joint universal space and compared these scores to human judgments. In Figure 4c, similarities in the universal space were significantly correlated with human ratings (Kendall's tau = 0.54), closely approaching CLIP's correspondence with human judgments on the same benchmark (Kendall's tau = 0.57). While CLIP remains modestly stronger in absolute terms, the gap is small relative to the difference in training regime: CLIP learns a joint space through large-scale paired supervision, whereas the universal embedding is obtained by post hoc alignment of independently trained unimodal models.

Importantly, this result shows that barycentric alignment yields a cross-modal similarity measure that tracks how humans relate meaning across modalities, rather than merely reflecting low-level statistical overlap. That such correspondence emerges without training a dedicated multimodal encoder suggests that substantial alignment-relevant structure is already implicit in unimodal representations and can be made explicit through symmetry-aware alignment.

Notably, these cross-modal findings are robust to several design choices. Additional pools yield qualitatively similar results (Appendix Figure 34). Performance also improves systematically with representation depth (Appendix Figure 25) and model capability (Appendix Figure 35– 36), and remains stable across a broad range of alignment-set training sizes (Appendix Figure 28–29), suggesting that the shared cross-modal structure is both general and data-efficient.

Table 2: Average NSD performance by region pool. Chance top-1/5/10 accuracies are 0.01/0.05/0.10.

Region	Corr.	RMS	Top-1	Top-5	Top-10
V1v	0.277	0.048	0.243	0.458	0.579
V1d	0.259	0.043	0.255	0.448	0.559
V2v	0.233	0.041	0.216	0.467	0.590
V2d	0.209	0.050	0.179	0.374	0.505
V3v	0.213	0.049	0.173	0.412	0.540
V3d	0.210	0.053	0.161	0.369	0.498
V4	0.195	0.048	0.198	0.417	0.544
Ventral	0.159	0.016	0.203	0.503	0.668
Lateral	0.150	0.023	0.117	0.318	0.446
Dorsal	0.170	0.015	0.139	0.389	0.575

Taken together, these findings demonstrate that the barycentric alignment framework extends naturally to the cross-modal setting, supporting both paired retrieval and human-aligned compatibility scoring. More broadly, it provides a principled lens for decomposing cross-modal alignment into components that are recoverable via geometry-preserving, post hoc transformations of pretrained representations (a form of shallow alignment) versus components that require contrastive multimodal training and a deeper reshaping of representational geometry.

3.4 Barycentric Alignment of Neural Representations

Next, we extend the universal embedding framework to biological data by constructing pools of neural representations derived from human visual cortex responses from the Natural Scenes Dataset Allen et al. (2022) (Appendix Table 6). Each cortical region from each subject is treated as a dis-

tinct model. These models are grouped by functional stream (e.g., pooling all ventral visual regions across subjects). These biological pools span distinct functional roles: early visual cortex primarily encodes low-level features such as edges and orientations, whereas higher-order visual regions support more abstract, image-centric representations that more closely align with the objectives typically optimized in artificial vision models. We compute instance-level representational consistency scores using MSCOCO images included in the Natural Scenes Dataset (Allen et al., 2022). Of these, 1,000 images were viewed by four subjects, and we construct a train–test split over this shared stimulus set, evaluating all metrics on held-out images.

We first assess the quality of the resulting universal embedding using the same metrics applied to artificial systems. Table 2 reports these metrics for pools drawn from each individual visual cortical region. For each region, scores are averaged across the four subjects (Appendix Table 13 reports individual-subject results). Despite the substantial noise inherent in fMRI-derived representations, correlation scores remain relatively high, RMS errors are low, and retrieval accuracies are consistently above chance. This shows that barycentric alignment yields a stable and meaningful shared space even for noisy biological population codes.

We next examine whether instance-level representational consistency scores are preserved across different biological pools. Strikingly, scores computed within pools drawn from comparable cortical regions are highly correlated: early visual region pools show strong agreement with one another, as do pools drawn from higher-order visual areas. In contrast, correlations between early and higher visual pools are substantially weaker (Figure 5, Appendix Figure 13). This pattern mirrors the known stratification of the visual hierarchy and indicates that instance-level representational agreement is structured by cortical stage rather than reflecting a generic property of the stimulus set.

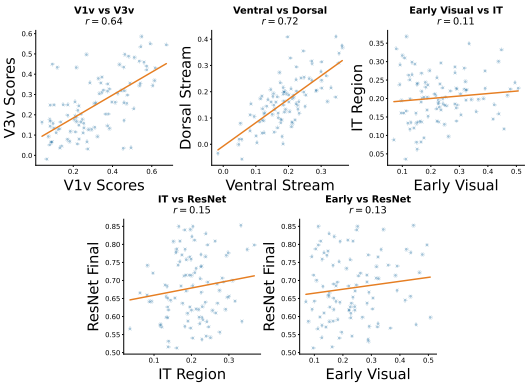


Figure 5: Instance-level representational consistency scores computed from different cortical region pools.

Notably, consistency scores computed from biological model pools show little correlation with those obtained from artificial model pools, such as the ResNet family (Figure 5, Appendix Figure 14). This dissociation indicates that the stimuli on which biological models exhibit strong consensus differ markedly from those that elicit agreement among artificial vision models. Strikingly, this divergence persists even for the IT (Later) model pool, which comprises high-level visual cortical regions whose functional role most closely parallels the objectives optimized in artificial networks, including object classification. Together, these results suggest that the dominant sources of variability captured by contemporary artificial model pools (e.g. architecture) do not yet provide a sufficient basis for explaining the structure of representational variability observed across biological individuals.

4 Discussion

We introduced barycentric alignment as a framework for co-registering populations of representation spaces into a shared quotient space while factoring out nuisance symmetries. This enables representational agreement to be measured at the instance level rather than only as a set-level summary. Across vision and language model pools, instance-level consistency is stable across pool constructions, varies with interpretable stimulus properties, and separates trained from untrained networks. In the cross-modal setting, post-hoc alignment of unimodal vision and language models produces image-text similarity scores that track human judgments, suggesting substantial recoverable shared structure in pretrained unimodal representations. In biological data, the framework yields stable within-region alignment but weak correspondence between cortical and artificial-model pools, suggesting that current model ensembles do not yet capture key sources of inter-individual biological variation.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118.

John C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/BF02291478.

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.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers*, pp. 199–209, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1019/.

Camila Kolling, Till Speicher, Vedant Nanda, Mariya Toneva, and Krishna P Gummadi. Investigating the effects of fairness interventions using pointwise representational similarity. *Transactions on Machine Learning Research*.

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.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. 2. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full.

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. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Aarre Laakso and Garrison W Cottrell. How can i know what you think?: Assessing representational similarity in neural systems. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 591–596. Routledge, 2022.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *NIPS 2015 Workshop on Feature Extraction: Modern Questions and Challenges*, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291, 2020. doi: 10.3758/s13428-019-01316-z.

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.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

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 (ICLR)*, 2023.

Lothar D. Narins, Andrew T. Scott, Aakash Gautam, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, James M. Mason, Alexander Blum, and Ilmi Yoon. Validated image caption rating dataset. In *Advances in Neural Information Processing Systems (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=xKYtTmtyI2.

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.

OpenLM Research. Openllama: An open reproduction of llama. https://github.com/openlm-research/open_llama, 2023. Accessed 2026-01-29.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Andrew Taylor Scott, Lothar D. Narins, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, and Ilmi Yoon. Improved image caption rating – datasets, game, and model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*, 2023. doi: 10.1145/3544549.3585632. Article 172.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jos M. F. Ten Berge. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42 (2):267–276, 1977. doi: 10.1007/BF02294053.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. doi: 10.48550/arXiv.2302.13971.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. doi: 10.1038/s41562-023-01783-7. Epub 2024-01-03.

Liwei Wang, Lei Hu, Jiayuan Gu, Yue Wu, Zhihua Hu, Kun He, and John E. Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. In *Advances in Neural Information Processing Systems*, 2018.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

A Appendix

Phase I: Training (Barycenter)

Input : $X_i \in \mathbb{R}^{n \times d}, T, \varepsilon$
Output: $M^{(t+1)}, \{T_i\}$

Initialize $M^{(0)} \leftarrow \frac{1}{N} \sum_i X_i$;
for $t = 0$ **to** $T - 1$ **do**
  for $i = 1$ **to** N **do**
    $T_i \leftarrow \arg\min_{R \in O(d)} \|X_i R - M^{(t)}\|_F$;
    $X_i^{(t)} \leftarrow X_i T_i$;
  end
  $M^{(t+1)} \leftarrow \frac{1}{N} \sum_{i=1}^{N} X_i^{(t)}$;
  if $\frac{\|M^{(t+1)} - M^{(t)}\|_F}{\|M^{(t)}\|_F} < \varepsilon$ **then**
    break;
  end
end

Phase II: Inference (Scoring)

Input : $Y_i \in \mathbb{R}^{m \times d}, \{T_i\}$
Output: $\{S_1, \ldots, S_m\}$

for $i = 1$ **to** N **do**
  $Y_i' \leftarrow Y_i T_i$;
end
for $j = 1$ **to** m **do**
  $S_j \leftarrow \frac{1}{N(N-1)} \sum_{p \neq q}^{N} \mathrm{SIM}(\mathbf{Y}'_{pj}, \mathbf{Y}'_{qj})$;
end

A.1 Metrics for evaluating Barycentric Alignment

We quantify the *quality of the learned universal embedding itself* by measuring how well representations from different models align once mapped into the shared space. Let $\{M_i\}_{i=1}^{X}$ denote a pool of X models evaluated on the same set of N stimuli. Model i produces a representation matrix $M_i \in \mathbb{R}^{N \times D_i}$. After projection into the universal space, we obtain $M_i' \in \mathbb{R}^{N \times D}$, where D is the shared dimensionality.

Correlation Score. For each model pair (i, j) with $i \neq j$ and each universal dimension $d \in \{1, \ldots, D\}$, we compute the Pearson correlation across stimuli: $\rho_d^{(i,j)} = \mathrm{corr}\big(M_i'[:, d], M_j'[:, d]\big)$. We report a per-model summary by averaging across dimensions and across all other models: $\mathrm{Corr}(i) = \frac{1}{(X-1)D} \sum_{\substack{j=1 \\ j \neq i}}^{X} \sum_{d=1}^{D} \rho_d^{(i,j)}$. High values indicate that corresponding universal axes capture similar stimulus-wise variation across models.

Root Mean Square (RMS) Score. For each pair (i, j) and dimension d, we compute the RMS discrepancy across stimuli: $\mathrm{RMS}_d^{(i,j)} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \big(M_i'[n, d] - M_j'[n, d]\big)^2}$. The per-model summary averages across dimensions and partners: $\mathrm{RMS}(i) = \frac{1}{(X-1)D} \sum_{\substack{j=1 \\ j \neq i}}^{X} \sum_{d=1}^{D} \mathrm{RMS}_d^{(i,j)}$. Low values indicate tight co-registration in the shared coordinates.

Procrustes Residuals. While RMS measures absolute discrepancy in the shared coordinates, *Procrustes residuals* measure how much of the universal template remains unexplained by the orthogonal alignment itself. Let $U \in \mathbb{R}^{N \times D}$ denote the learned universal template (barycenter). For model

i, we report the normalized Procrustes residual

$$\mathrm{Res}(i) \;=\; \frac{\|M_i' - U\|_F^2}{\|U\|_F^2} \;=\; \frac{\sum_{n=1}^{N}\sum_{d=1}^{D}\left(M_i'[n,d] - U[n,d]\right)^2}{\sum_{n=1}^{N}\sum_{d=1}^{D} U[n,d]^2},$$

where $M_i' \in \mathbb{R}^{N\times D}$ is the representation of model i after projection into the universal space. Equivalently, if T_i denotes the learned orthogonal transform for model i, then

$$\mathrm{Res}(i) \;=\; \frac{\|M_i T_i - U\|_F^2}{\|U\|_F^2}.$$

This quantity can be interpreted as the fraction of template variance left unexplained by the orthogonal fit of model i to the shared barycenter. Smaller values indicate better fit under the orthogonal alignment assumption.

Cross-model Stimulus Retrieval Accuracy. Cross-model stimulus retrieval operationalizes a strong notion of representational agreement: whether different models not only encode stimuli similarly in aggregate, but identify the same stimulus as corresponding to itself across representational spaces. For a stimulus $x \in \{1,\dots,N\}$ and a model pair (M_i, M_j) with $i \neq j$, we compute the Euclidean distances $d_{x,y}^{(i,j)} = \left\|M_i'[x] - M_j'[y]\right\|_2, \quad \forall y \in \{1,\dots,N\}$. Let $\mathcal{N}_K^{(i,j)}(x)$ denote the indices of the K smallest distances. The retrieval accuracy for stimulus x is $\mathrm{Acc}_K^{(i,j)}(x) = \mathbb{I}\Big[x \in \mathcal{N}_K^{(i,j)}(x)\Big]$, where $\mathbb{I}(\cdot)$ is the indicator function. The final top-K retrieval accuracy for model M_i is then computed by averaging over all stimuli and all other models in the pool: $\mathrm{Acc}_K(i) = \frac{1}{(X-1)N}\sum_{\substack{j=1\\ j\neq i}}^{X}\sum_{x=1}^{N}\mathrm{Acc}_K^{(i,j)}(x)$ High Top-K accuracy indicates that aligned representations preserve stimulus identity across models. This quantifies the extent to which cross-model representational differences are eliminable once nuisance symmetries are removed.

A.2 Datasets

A.2.1 Vision Dataset

We utilize two primary vision datasets in our experiments:

ImageNet. ImageNet Deng et al. (2009) is a large-scale image classification dataset comprising over 1.2 million images across 1,000 object categories. It is commonly used to train and evaluate convolutional neural networks and other vision models. For our experiments, we utilized only the ImageNet validation set, which contains 50,000 images. Of these, 45,000 images were used to train the transformation matrices for the pure vision model pools, while the remaining 5,000 images were reserved for inference.

THINGS. The THINGS dataset Hebart et al. (2023) is a curated collection of over 26,000 high-quality, naturalistic object images spanning 1,854 object concepts. Unlike ImageNet, THINGS emphasizes detailed, instance-level representations of everyday objects. We use the THINGS dataset solely for validation. Specifically, we apply the transformation matrices learned from the ImageNet dataset to compute instance-level consistency scores for THINGS. These scores are then compared to human behavioral ratings to assess their correlation with human judgments.

A.2.2 Language Dataset

Tuckute et al. We adopted the English sentence stimulus set introduced by Tuckute et al. (2024). This dataset is well-suited for evaluating universality of language-model representations because it was explicitly curated to span a wide range of linguistic content and (in the original study) a wide range of language-network brain responses, rather than reflecting a single genre or task distribution. All stimuli are short, corpus-extracted six-word sentences. The dataset contains:

- **Baseline sentences** ($n = 1{,}000$): naturalistic six-word sentences curated to maximize linguistic and semantic diversity, combining a semantically diverse subset ($n = 534$) with genre-diverse sentences sampled from sources such as news, web media, and transcribed speech ($n = 466$).

- **Drive / suppress sentences** ($n = 250$ each): out-of-distribution sentences selected (via a GPT2-XL–based encoding model) to elicit maximal (drive) or minimal (suppress) responses in the human language network; identified by ranking candidates from a large-scale search over $\sim$1.8M sentences spanning nine corpora.

Sentence-level properties. To relate universality scores to interpretable linguistic dimensions, we used the sentence-level properties released alongside the stimulus set. These include a model-based predictability measure (surprisal/log probability; estimated with GPT2-XL) and human-rated norms. Specifically, we analyze the following properties:

- **Surprisal / log probability**: model-estimated predictability of the sentence (higher surprisal indicates lower predictability).
- **Grammaticality**: how well the sentence conforms to English grammar.
- **Plausibility**: how much the sentence "makes sense" semantically.
- **Mental-state content**: the extent to which the sentence evokes others' thoughts, beliefs, or intentions.
- **Physical-object content**: the extent to which the sentence evokes physical objects and their interactions.
- **Place content**: the extent to which the sentence evokes places and environments.
- **Valence**: how positive versus negative the sentence content is.
- **Arousal**: how exciting or emotionally activating the sentence content is.
- **Visual imagery**: how easy it is to form a visual mental image of the sentence content.
- **General frequency**: perceived commonness of the sentence in everyday language overall.
- **Conversational frequency**: perceived commonness of the sentence specifically in spoken conversation.

Word-level sensorimotor properties (Lancaster norms). To complement the sentence-level analyses, we also performed a word-level evaluation using the Lancaster Sensorimotor Norms (Lynott et al., 2020). This resource provides human ratings for a large vocabulary of English words along multiple perceptual and action-related dimensions, making it useful for testing whether representational consistency is associated with grounded semantic content. For each target word, we constructed a simple prompt of the form *"the meaning of word x"*, extracted the corresponding model representations, computed instance-level consistency scores, and related those scores to the Lancaster norm dimensions.

We focus on the following groups of sensorimotor properties:

- **Perceptual dimensions**: auditory, gustatory, haptic, interoceptive, olfactory, and visual strength.
- **Action dimensions**: foot/leg, hand/arm, head, mouth, and torso/action strength.

These norms provide a complementary characterization of consistency at the word level, allowing us to test whether words with stronger perceptual or action-grounded semantic content tend to elicit more convergent representations across model pools.

A.2.3 Multimodal Datasets

MS-COCO. MS-COCO is a large-scale image captioning dataset of natural images depicting complex everyday scenes, with each image annotated with five human-authored captions Lin et al. (2014). We use the official 2017 validation split (Val2017; 5,000 images) and randomly sample 1,000 images with their associated captions for constructing the cross-modal universal embedding space.

Flickr8k-Expert. Flickr8k-Expert is a human-judgment benchmark for image–caption compatibility, consisting of 1,000 images from the Flickr8k test set paired with 5,822 candidate captions. Each image–caption pair is rated by expert annotators on a 4-point scale (1 = unrelated; 4 = perfect match), with three ratings per pair Kilickaya et al. (2017); Scott et al. (2023); Narins et al. (2023). We use these ratings as ground-truth human judgments to evaluate whether similarity scores derived from the cross-modal universal space track human assessments of caption quality.

A.2.4 BRAIN DATASETS

A detailed description of the Natural Scenes Dataset (NSD; http://naturalscenesdataset.org) is provided elsewhere (Allen et al., 2022). The NSD dataset contains measurements of fMRI responses from 8 participants who each viewed 9,000–10,000 distinct color natural scenes (22,000–30,000 trials) over the course of 30–40 scan sessions. Scanning was conducted at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. Images were taken from the Microsoft Common Objects in Context (COCO) database (Lin et al., 2014), square cropped, and presented at a size of 8.4° x 8.4°. A special set of 1,000 images were shared across subjects; the remaining images were mutually exclusive across subjects. Images were presented for 3 s with 1-s gaps in between images. Subjects fixated centrally and performed a long-term continuous recognition task on the images. The fMRI data were pre-processed by performing one temporal interpolation (to correct for slice time differences) and one spatial interpolation (to correct for head motion). A general linear model was then used to estimate single-trial beta weights. Cortical surface reconstructions were generated using FreeSurfer, and both volume- and surface-based versions of the beta weights were created. In this study, we analyze manually defined regions of interest (ROIs) across both early and higher-level visual cortical areas. For early visual areas, we focus on ROIs delineated based on the results of the population receptive field (pRF) experiment - V1v, V1d, V2v, V2d, V3v, V3d, and hV4. For higher level visual cortex regions, we target the ventral, dorsal, and lateral streams, as defined by the streams atlas.

A.3 MODELS

A.3.1 VISION MODELS

1. **ResNet family** He et al. (2016) (ResNet-50, ResNet-34, ResNet-18, ResNet-101): Deep residual networks that use skip connections to mitigate vanishing gradients, enabling very deep convolutional architectures.
2. **ConvNeXt family** (ConvNeXt Small, Tiny, Base, Large) Liu et al. (2022): A modernized CNN design that incorporates architectural ideas from Transformers, such as large kernels and inverted bottlenecks.
3. **Swin Transformers** Liu et al. (2021) (Swin Small, Tiny, Base, Large): Hierarchical vision Transformers that process images using shifted windows, providing linear computational complexity with respect to image size.
4. **Vision Transformers (ViTs)** (ViT Small, Tiny, Base, Large) Dosovitskiy et al. (2020): Pure Transformer architectures that treat images as sequences of non-overlapping patches, capturing global context through self-attention.
5. **DINO models** (DINO Small, Base, ResNet-50) Caron et al. (2021): Self-supervised representations learned via knowledge distillation, producing strong, transferable visual features without labeled data and supporting both ViT and ResNet backbones.
6. **VGGs** (VGG 11,13,16,19) Simonyan & Zisserman (2014) - is characterized by its depth, consisting of 16 layers including 13 convolutional layers and 3 fully connected layers. VGG-16 is renowned for its simplicity and effectiveness as well as its ability to achieve strong performance on various computer vision tasks including image classification and object recognition.
7. **AlexNet** Krizhevsky et al. (2012) - A pioneering CNN that won the 2012 ImageNet competition, introducing ReLU activations, overlapping max-pooling, and dropout for regularization. AlexNet consists of five convolutional layers followed by three fully connected layers

A.3.2 LANGUAGE MODELS

1. **OpenLLaMA family** (OpenLLaMA-3B, 7B, 13B) OpenLM Research (2023); Touvron et al. (2023): Permissively licensed reproductions of the LLaMA-style decoder-only Transformer architecture, trained as causal language models. We use the 3B/7B/13B checkpoints as representative autoregressive LLMs with matched architecture and varying scale.
2. **Qwen2.5 family** (Qwen2.5-0.5B, 1.5B, 3B, 7B, 14B) Qwen Team et al. (2024): A modern family of decoder-only Transformer language models spanning a wide range of parameter scales, trained with large-scale pretraining and post-training to support strong general-purpose language understanding and generation.
3. **Gemma 2** (Gemma2-2B) Gemma Team (2024): A lightweight, state-of-the-art decoder-only Transformer model at a practical scale, trained with modern architectural and training refinements.
4. **ALBERT family** (ALBERT-base/large/xlarge/xxlarge-v2) Lan et al. (2020): Encoder-only Transformer models that introduce parameter-sharing and factorized embeddings to reduce memory while maintaining performance, trained with self-supervised objectives for general-purpose language representations.
5. **GPT-2 family** (GPT2-small/medium/large/xl) Radford et al. (2019): Decoder-only Transformer language models trained with next-token prediction, commonly used as baseline autoregressive LLMs across a range of model sizes.

A.3.3 MODEL POOLS

For the barycentric alignment experiments, we group individual models into several pools designed to isolate different sources of variation, including architecture, scale, training objective, and auxiliary mixed compositions. Appendix Tables 3, 4, and 5 summarize the vision, language, and cross-modal unimodal pools used throughout the paper.

For cross-modal experiments, each pool consists of pretrained vision-only and language-only models jointly aligned into a shared universal space. Unless otherwise noted, results in the main text are reported at the pool level, with per-model results provided separately where relevant.

Table 3: **Vision model pools.** Categorized by architecture, scale, and training objective for Procrustes alignment.

Group	Models
Architecture-based pools	
ResNets	resnet18, resnet34, resnet50, resnet101, resnet152
VGGs	vgg11, vgg13, vgg16, vgg19
ConvNeXts	convnext_tiny, convnext_small, convnext_base, convnext_large
ViTs	vit_t_16, vit_s_16, vit_b_16, vit_l_16
DINO ViTs	vit_s_16_dino, vit_b_16, vit_b_8
Swins	swin_t, swin_s, swin_b
Scale and objective-based pools	
Large	resnet101, resnet152, vgg16, vgg19, swin_b, vit_b_16, convnext_base, convnext_large
Small	resnet18, resnet34, resnet50, vgg11, vgg13, swin_t, swin_s, vit_s_16, vit_t_16
Supervised	Standard variants of ResNet, VGG, ViT, Swin, and ConvNeXt
Self-sup.	resnet50_dino, moco_v2, vit_s_16_dino, vit_b_16_dino
Auxiliary mixed pools	
Misc-1	resnet34, resnet50_dino, swin_b
Misc-2	vgg13, moco_v2, vit_s_16_dino
Misc-3	alexnet, vit_s_16_dino, swin_s

Table 4: **Language model pools.** Categorized by model family, scale, and auxiliary mixed constructions for barycentric alignment.

Group	Models
Family-based pools	
OpenLLaMA	OpenLLaMA-3B, OpenLLaMA-7B, OpenLLaMA-13B
Qwen2.5	Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B
GPT-2	GPT2-small, GPT2-medium, GPT2-large, GPT2-xlarge
ALBERT	ALBERT-base, ALBERT-xxlarge
Autoregressive	OpenLLaMA-3B, OpenLLaMA-7B, OpenLLaMA-13B, Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, Llama3-1B, Llama3-3B, Llama3-8B, Gemma2-2B
Scale-based pools	
Large	OpenLLaMA-13B, Qwen2.5-14B, Llama3-8B, ALBERT-xxlarge, GPT2-xlarge
Small	ALBERT-base, GPT2-small, Qwen2.5-0.5B, OpenLLaMA-3B, Llama3-1B
Auxiliary mixed pools	
Misc-1	GPT2-small, Llama3-3B, GPT2-xlarge, OpenLLaMA-13B, GPT2-large, Qwen2.5-1.5B, Qwen2.5-14B, Gemma2-2B
Misc-2	Qwen2.5-7B, GPT2-medium, Qwen2.5-3B, Qwen2.5-0.5B, Llama3-1B, Llama3-8B, OpenLLaMA-7B, OpenLLaMA-3B

Table 5: **Cross-modal unimodal pools.** Joint pools of vision-only and language-only models used for post-hoc cross-modal barycentric alignment.

Group	Models
OpenLLaMA-based cross-modal pool	
OpenLLaMA-large + ViT-large/giant	OpenLLaMA-7B, OpenLLaMA-13B + ViT-Large, ViT-Giant
Qwen-based cross-modal pools	
Qwen-small + ViT-small/base	Qwen2.5-0.5B, Qwen2.5-1.5B + ViT-Small, ViT-Base
Qwen-small + ViT-large/giant	Qwen2.5-0.5B, Qwen2.5-1.5B + ViT-Large, ViT-Giant
Qwen-medium + ViT-large/giant	Qwen2.5-3B, Qwen2.5-7B + ViT-Large, ViT-Giant
Qwen3-large + ViT-large/giant	Qwen3-8B, Qwen3-14B + ViT-Large, ViT-Giant
Reference multimodal baseline	
CLIP	CLIP-ViT-B/32

Table 6: **Biological model pools.** Categorized by visual region, architecture, and scale for Procrustes alignment.

Group	Models
Region-based pools	
V1v	V1v – Subjects 1,2,5,7
V1d	V1d – Subjects 1,2,5,7
V2v	V2v – Subjects 1,2,5,7
V2d	V2d – Subjects 1,2,5,7
V3v	V3v – Subjects 1,2,5,7
V3d	V3d – Subjects 1,2,5,7
V4	V4 – Subjects 1,2,5,7
Ventral Visual Stream	Ventral Visual Stream – Subjects 1,2,5,7
Dorsal Visual Stream	Dorsal Visual Stream – Subjects 1,2,5,7
Lateral Visual Stream	Lateral Visual Stream – Subjects 1,2,5,7
Functionality-based pools	
EARLY	V1v, V1d, V2v, V2d, V3v, V3d, V4 – Subjects 1,2,5,7
LATER	Ventral, dorsal, lateral streams – Subjects 1,2,5,7

A.4 EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

This subsection summarizes the default implementation choices used for the **main experiments**, including feature extraction, train/test splits, alignment hyperparameters, preprocessing, and randomness control. Unless otherwise noted, the appendix analyses use the same default setup and vary only the factor under study (e.g., alignment baseline, representational depth, or training-set size).

Feature extraction. For pure vision model pools, representations were taken from the model's penultimate pooled output, yielding a single vector per image. For **language models**, we extracted hidden states from all transformer layers using the HuggingFace `transformers` interface with `output_hidden_states=True`. We then applied attention-mask-weighted mean pooling over tokens to obtain a single vector per sentence.

To compare representations across depth, we mapped each model to a normalized depth grid $\{0.0, 0.1, \ldots, 1.0\}$, where depth δ corresponds to layer index $\mathrm{round}(\delta \cdot L)$ for a model with L layers. Unless otherwise noted, the main language-model results use normalized depth 0.9.

For the **main cross-modal experiments**, we used caption embeddings from `OpenLLaMA-7B` and `OpenLLaMA-13B` at normalized depth 0.9, and image embeddings from `ViT-L/14 DINOv2` and `ViT-g/14 DINOv2` at normalized depth 0.9. For the cross-modal vision models (ViTs from `timm`), we extracted features from the residual addition after each transformer block (`blocks.{i}.add_1`) using a `torchvision` feature extractor, and used the CLS token (index 0) as the representation at each depth. The CLIP baseline uses the standard `CLIP-ViT-B/32` `encode_image` / `encode_text` outputs (with 77-token chunked encoding and per-caption averaging for any caption longer than CLIP's context window).

Train/test splits.

- **Vision-model pools**: transformation matrices were learned using $45{,}000$ ImageNet validation images and evaluated on the disjoint $5{,}000$-image held-out split.
- **Language-model pools**: alignment was learned on the Tuckute et al. six-word stimulus set (`greta_baseline_search`, $1{,}500$ sentences = $1{,}000$ baseline + 250 drive + 250 suppress) using a stratified $80/20$ split over the three condition labels (`T`, `D_search`, `B_search`) with seed 42, yielding $1{,}200$ training and 300 held-out test sentences.
- **Cross-modal experiments**: the universal space was learned on $1{,}000$ MS-COCO val2017 image–caption pairs (using one of the five captions per image as the alignment caption) with a random $80/20$ split (seed 42), giving 800 training and 200 held-out test pairs; the main retrieval numbers in Figure 4b are reported on these 200 pairs.
- **Flickr8k-Expert**: Flickr8k-Expert was used for cross-modal evaluation only. The transforms learned on MS-COCO were applied to all $5{,}822$ Flickr8k-Expert image–caption pairs to compute cosine similarities in the shared space, which were then correlated with mean human ratings (Kendall's τ, Figure 4c).

Unless otherwise noted, appendix ablations use the same underlying data sources and preprocessing, while varying only the factor under study. In particular, the depth analyses vary representational depth, the training-size analyses vary the amount of alignment data, and the baseline analyses replace the alignment method while keeping evaluation datasets fixed.

Alignment optimization. Barycentric alignment was run for at most $T = 5$ iterations and stopped early when the relative change in the barycenter satisfied $\|M^{(t+1)} - M^{(t)}\|_F / \|M^{(t)}\|_F < \varepsilon = 10^{-6}$. The orthogonal Procrustes step uses `scipy.linalg.orthogonal_procrustes`. Unless otherwise noted, all reported results use orthogonal alignment with zero-padding to the maximum dimensionality within each pool, followed by per-model mean-centering and L2 row-normalization. Convergence in practice is reached within a few iterations for the pool sizes used in this paper.

Cross-modal implementation details. In the main cross-modal experiments, we used the unimodal pool {`OpenLLaMA-7B`, `OpenLLaMA-13B`, `ViT-L/14 DINOv2`, `ViT-g/14 DINOv2`} at $(\delta_{\text{lang}}, \delta_{\text{vis}}) = (0.9, 0.9)$ (see also Table 5). The CLIP baseline used `CLIP-ViT-B/32`.

Before barycentric alignment, we applied a per-model, per-depth ridge pre-mapping that projected caption embeddings and image embeddings into a shared item-level target space defined by the corresponding MS-COCO image–caption group. This pre-mapping was fit separately for each model and depth on the training split and then reused at hyperalignment time.

For unimodal vision models, image preprocessing used the default `timm` transform associated with each pretrained backbone. For CLIP, we used the preprocessing pipeline returned by `clip.load`. For language models, we used each model's HuggingFace tokenizer with dynamic padding to the longest sequence in the batch. Retrieval was evaluated in both text→image and image→text directions using cosine similarity in the aligned shared space.

Baseline preprocessing. For *pairwise Procrustes alignment* (PA), each model pair was aligned independently using orthogonal Procrustes after feature standardization. For *Relative Representations* (RelRep), anchors were the alignment-set stimuli themselves, and each embedding was replaced by its cosine-similarity vector to those anchors. For *ASIF*, we used the same anchor sets as RelRep and selected sparsification/sharpening hyperparameters by grid search on retrieval performance. All baselines used the same train/test splits, depth choices, and per-model embeddings as barycentric alignment; only the alignment method changed.

Randomness and seeds. We used a global random seed of 42 for all randomized train/test splits, anchor-set construction, and random-model initialization. Model-pool definitions were fixed in advance and not randomly sampled. Given the input representations, barycentric alignment itself is deterministic, since it is initialized with the arithmetic mean and introduces no additional stochasticity. Main reported results are from a single seed unless otherwise noted.

A.4.1 Compute resources

Experiments were run on a single workstation with $8\times$ NVIDIA A40 GPUs (46 GB VRAM each) and $2\times$ AMD EPYC 9224 24-core CPUs (755 GB system RAM). Individual extraction and alignment jobs used a single GPU each, while alignment itself was primarily CPU-bound. Reported single modality pool experiments required approximately 1.5 GPU-hours for activation extraction plus 1.5 CPU-hours for alignment/evaluation, and reported cross-modal experiments required approximately 2.5 GPU-hours plus 25 CPU-minutes. The full project also involved additional exploratory compute beyond the reported runs, which was not systematically logged.

A.4.2 Licenses and terms of use for existing assets

Our experiments use publicly available datasets and pretrained model checkpoints obtained from their official release sites or repositories. Table 7 summarizes the corresponding licenses or access terms. For assets governed by access agreements or dataset-specific terms rather than standard software licenses, we report the official terms of use instead.

A.5 Additional results

A.5.1 Baseline comparisons for shared-space alignment

We compare barycentric alignment against several related baselines that also aim to make representations from different models comparable. *Pairwise Procrustes alignment* (PA) (Maiorca et al., 2023) uses orthogonal Procrustes fits independently for each model pair and therefore provides a strong pairwise baseline, but does not produce a single shared space across the full model population. *Relative Representations* (RelRep) (Moschella et al., 2023) instead defines a shared comparison space through similarities to anchor stimuli. In the cross-modal setting, we additionally compare against *ASIF* (Norelli et al., 2023), which extends anchor-based relative representations with sparsification and sharpening. Figures 6–9 report these comparisons for vision, language, and cross-modal settings.

Table 7: Licenses and terms of use for existing assets used in this work.

Asset	Type	License / terms of use
ImageNet	Dataset	Official ImageNet access agreement / terms of use (not a standard open license).
THINGS	Dataset	Original THINGS images are for academic use only; the official site recommends using THINGSplus license-free images for publications.
MS-COCO	Dataset	Official COCO Terms of Use; images are subject to Flickr terms of use, and COCO annotations / website are distributed under CC BY 4.0.
Flickr8k-Expert	Dataset	Available for non-commercial academic research (images subject to original Flickr8k academic request terms).
NSD	Dataset	Access subject to the NSD Data Access Agreement.
Torchvision pretrained vision models	Model	BSD-3-Clause (official torchvision repository; includes standard pretrained models such as ResNet, VGG, AlexNet, and torchvision vision transformers where used).
timm pretrained vision models	Model	Official timm repository is Apache License 2.0; includes pretrained ConvNeXt / ViT / Swin models where used.
DINOv2 checkpoints	Model	Apache License 2.0.
OpenLLaMA	Model	Apache License 2.0.
Qwen2.5	Model	Checkpoint-specific official Hugging Face license files; verified checkpoints used here include Apache License 2.0.
Qwen3	Model	Apache License 2.0.
Gemma 2	Model	Access subject to Google's Gemma usage license / access agreement.
ALBERT	Model	Apache License 2.0.
GPT-2	Model	Hugging Face checkpoint license varies by released variant; verified pages show MIT for openai-community/gpt2 and Modified MIT License for gpt2-medium.
CLIP-ViT-B/32	Model	MIT license.

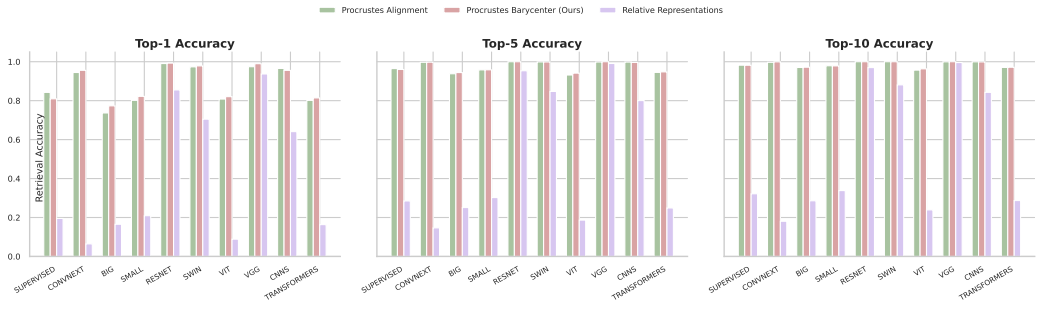


Figure 6: Comparison of retrieval accuracies across three alignment methods for vision model pools: pairwise Procrustes alignment, relative representations, and our approach based on Procrustes barycenters via Generalized Procrustes Analysis (GPA). Notably, while the barycenter and relative representation methods align multiple models into a shared universal space, pairwise Procrustes performs alignment independently for each model pair, without constructing a common embedding space.

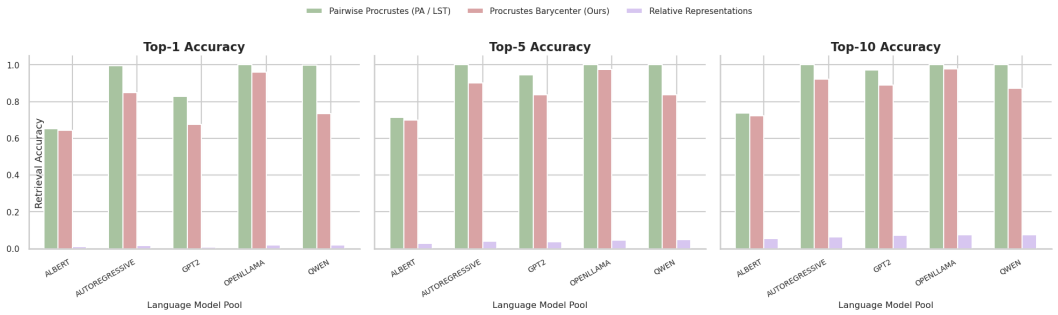


Figure 7: Comparison of retrieval accuracies across three alignment methods for language model pools: pairwise Procrustes alignment, relative representations, and our approach based on Procrustes barycenters via Generalized Procrustes Analysis (GPA). Notably, while the barycenter and relative representation methods align multiple models into a shared universal space, pairwise Procrustes performs alignment independently for each model pair, without constructing a common embedding space.

Table 8: Quantitative evaluation of alignment strategies across the Supervised Vision model pool. PA (Procrustes Alignment) aligns each pair of models in the pool using an optimal orthogonal Procrustes transformation, providing an upper bound on alignment performance under pairwise rotational transformations and serving as a reference for the best achievable performance. PF (Procrustes Fixed) fixes the barycenter to the representation space of a single model, using each model in the pool in turn and averaging performance across all choices, PB (Procrustes Barycenter Alignment), our proposed method via Pearson correlation, RMS error, and Top-k retrieval accuracy.

Method	**Corr.** ↑	**RMS** ↓	**Top-1** ↑	**Top-5** ↑	**Top-10** ↑
PB (Ours)	0.4864	0.0138	0.8094	0.9602	0.9818
PF	0.4262	0.0118	0.7911	0.9516	0.97664
PA	0.5128	0.2203	0.8425	0.9647	0.9822

Table 9: Quantitative evaluation of alignment strategies across the language model pool and cross-modal pool. The proposed method PB (Procrustes Barycenter Alignment) consistently out-perform the baseline PF (Procrustes Fixed), which fixes the barycenter to the representation space of a single model, on evaluation metrics, especially in Pearson correlation and Top-k retrieval accuracy.

Method	**Corr.** ↑	**RMS** ↓	**Top-1** ↑	**Top-5** ↑	**Top-10** ↑
Autoregressive Language Models					
PB (Ours)	0.3481	0.0135	0.6205	0.7278	0.7732
PF	0.3246	0.0133	0.6167	0.7214	0.7679
Vision (VIT) and Language (OpenLLaMA) Models					
PB (Ours)	0.7617	0.0081	0.4775	0.8225	0.9525
PF	0.6975	0.0065	0.4731	0.8181	0.9506

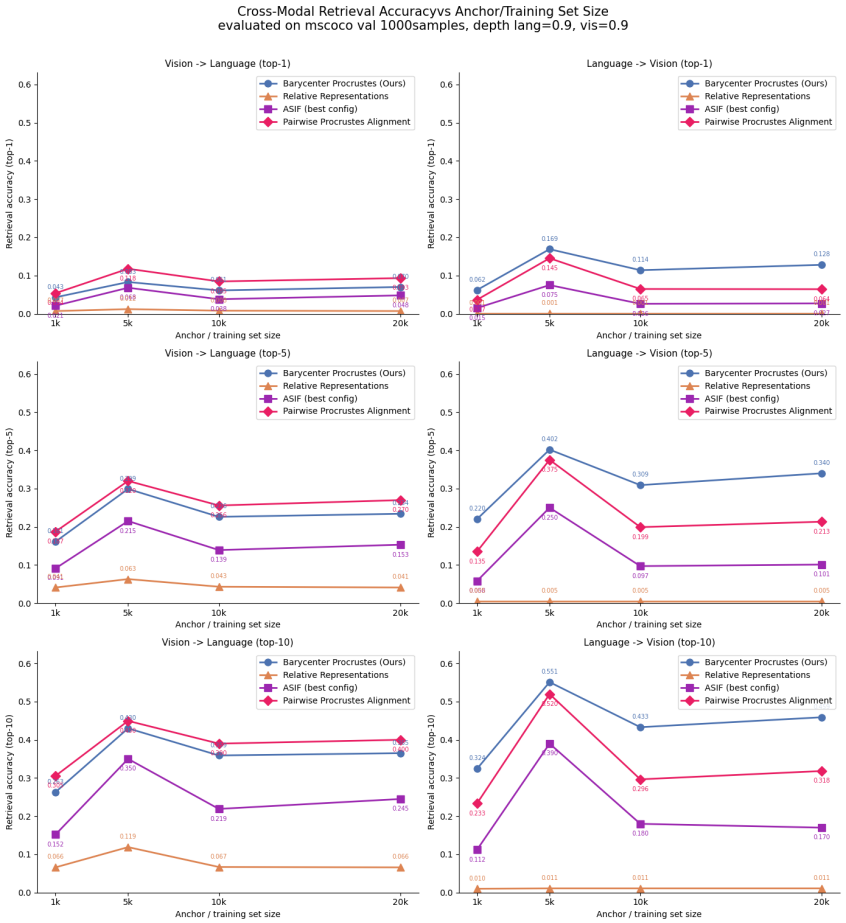


Figure 8: Cross-modal retrieval (text→image and image→text; top-1, top-5, top-10 accuracy) as a function of anchor/training set size for the OpenLLaMA-ViT(L) pool, evaluated on a fixed MS-COCO val2017 set (n=1000). **Barycenter Procrustes** (blue): Our iterative orthogonal alignment approach. All models are simultaneously aligned to a shared democratic barycenter space via iterative Procrustes rotations. **Relative Representations (RelRep)** (orange): Dense relative representations; each stimulus is encoded as a vector of cosine similarities to all anchor stimuli. No sparsification or sharpening. **ASIF** (purple): Sparse relative representations with top-k sparsification and power-law sharpening. Each model's test embedding is represented as a sparse vector of cosine similarities to anchor (training) stimuli; retaining only the k most similar anchors, raising similarities to the power p, and row-normalizing. At each anchor size, the best (k, p) configuration is selected by grid search over $k \in \{25, 50, 100, 200, 400, 800\}$ and $p \in \{1, 2, 4, 6, 8\}$, maximizing mean top retrieval accuracy across both directions. **Pairwise Procrustes Alignment (PA)** (pink): Pairwise orthogonal Procrustes alignment with StandardScaler preprocessing (close related to Latent Space Translation). Each (language, vision) model pair is aligned independently in both directions (lang→vis and vis→lang); retrieval accuracy is averaged across all cross-modal pairs. Unlike Barycenter Procrustes, pairwise Procrustes Alignment does not construct a shared multi-model space; it optimizes each pair separately, which is expected to yield higher per-pair retrieval since each transform is fit to just two models. Procrustes achieves retrieval accuracy comparable to pairwise PA despite constructing a multi-model shared space rather than optimizing pairwise transforms, especially in language→vision direction.

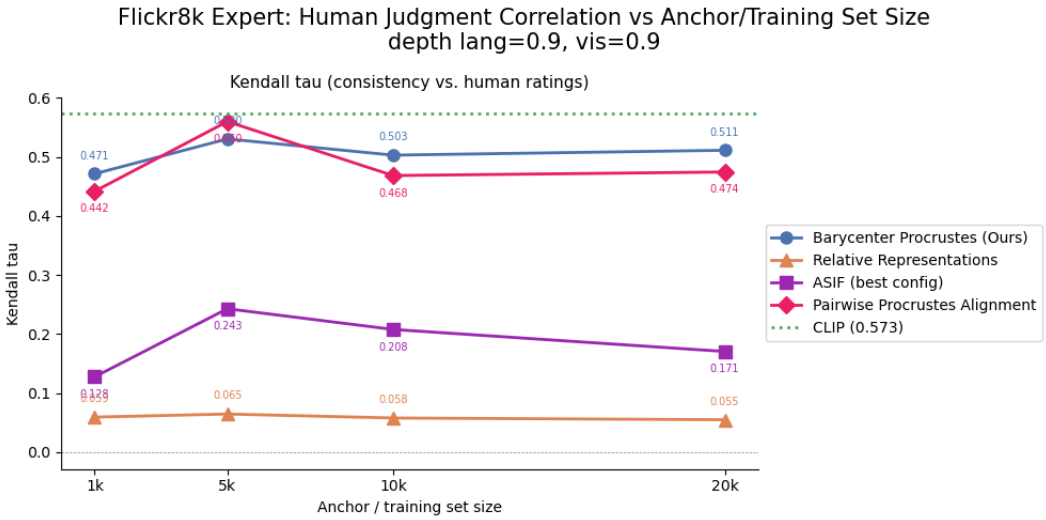


Figure 9: Agreement with Flickr8k-Expert human judgment ratings, measured by Kendall's tau, as a function of anchor/training set size for the OpenLLaMA-ViT(L) pool. Barycenter Procrustes (ours) achieves human judgment correlations comparable to pairwise PA across all training set sizes, despite pairwise PA's pairwise optimization advantage. Both methods show stable correlations even at small anchor sizes (5k), suggesting that orthogonal alignment generalizes well from limited training data. ASIF and Relative Representations correlations are lower, reflecting that similarity-based representations capture less of the instance-level structure that human raters attend to.

A.5.2 ADDITIONAL ANALYSES OF CONSISTENCY ACROSS MODEL POOLS

This section provides supplementary analyses extending the final part of Section 3.1, where we examine how instance-level representational consistency scores compare across different model-pool constructions. We report a more complete set of pairwise comparisons across vision model families and architectures (Figure 10), assess the dependence of these scores on learned model parameters by comparing trained and randomly initialized pools (Figure 11), and test whether the cross-pool consistency structure generalizes to out-of-distribution stimuli by comparing ImageNet- and THINGS-based scores (Figure 12). Finally, we provide the analogous across-pool comparison for biological model pools, showing how consistency scores relate across cortical regions (Figure 13).

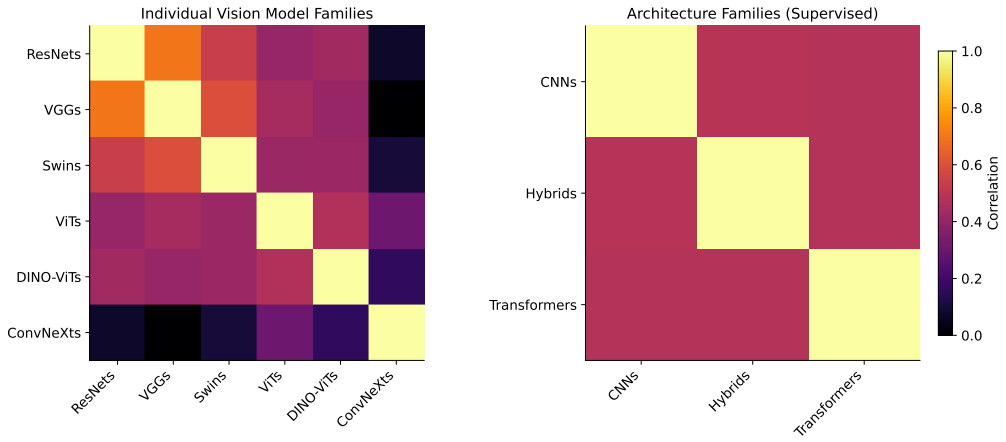


Figure 10: Heatmap showing consistency of universal similarity scores across different model families (left), and across different architectures (right)

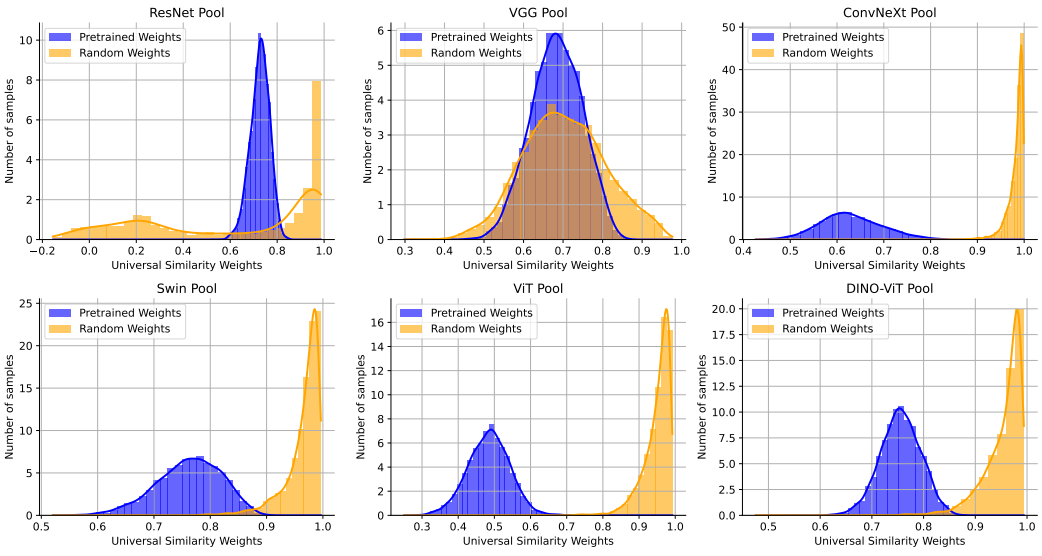


Figure 11: Instance-level consistency scores depend strongly on the model weights used in the model pool, with substantial differences observed between models with random weights and those with trained weights.

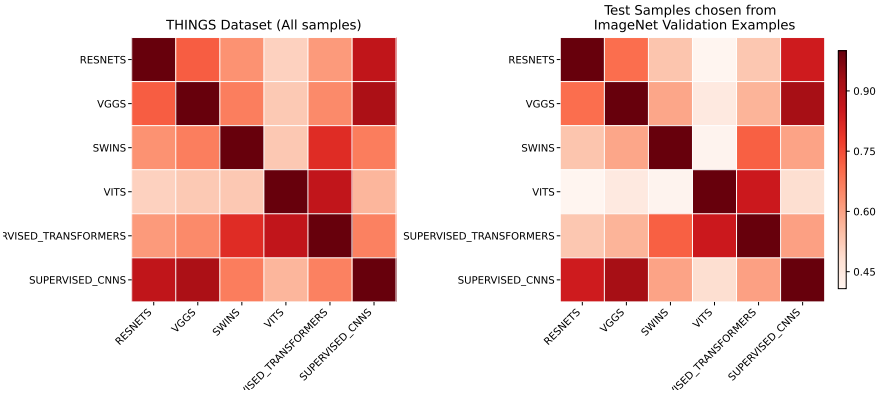


Figure 12: Transformation matrices were learned using training samples from the ImageNet validation set to map raw model representations into a universal embedding space. Universal scores were then computed on test samples from ImageNet and on samples from the THINGS dataset using the learned transformation matrices. Consistency across different model pools is shown for THINGS (left) and ImageNet (right). This consistency is highly similar across the two datasets, with a correlation of approximately 0.99 between the ImageNet- and THINGS-based score similarity matrices, indicating strong generalization to out-of-distribution data.

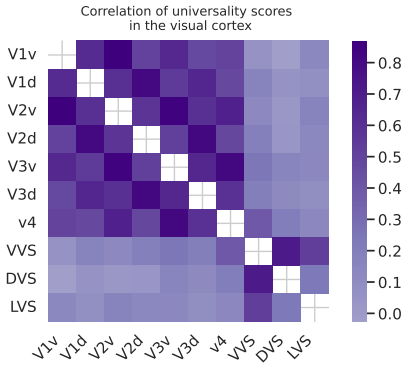


Figure 13: Universal similarity scores computed from different biological model pools. Scores within early visual region pools show strong consistency with one another, as do scores within higher-order visual region pools. In contrast, similarity scores between early and higher-order region pools are substantially weaker, reflecting functional differences across the visual hierarchy.

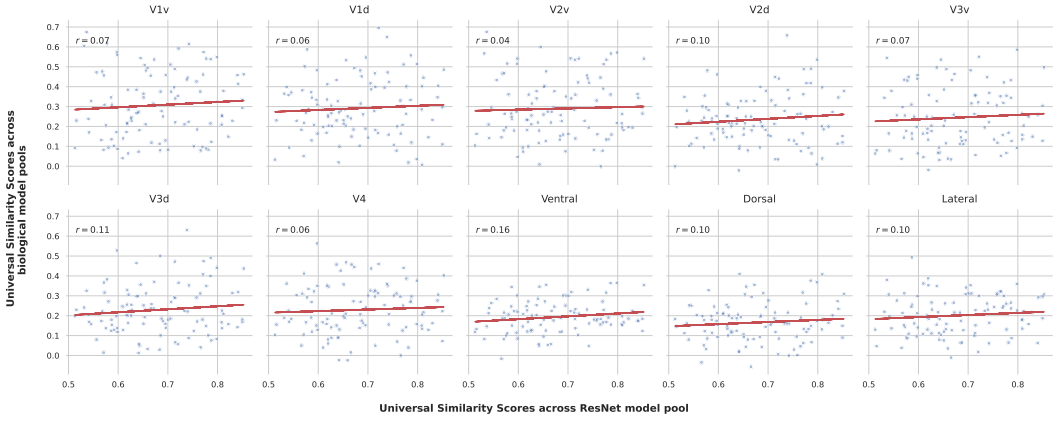


Figure 14: Comparing Universal Similarity scores obtained via biological model pools with those obtained via ResNet model pool

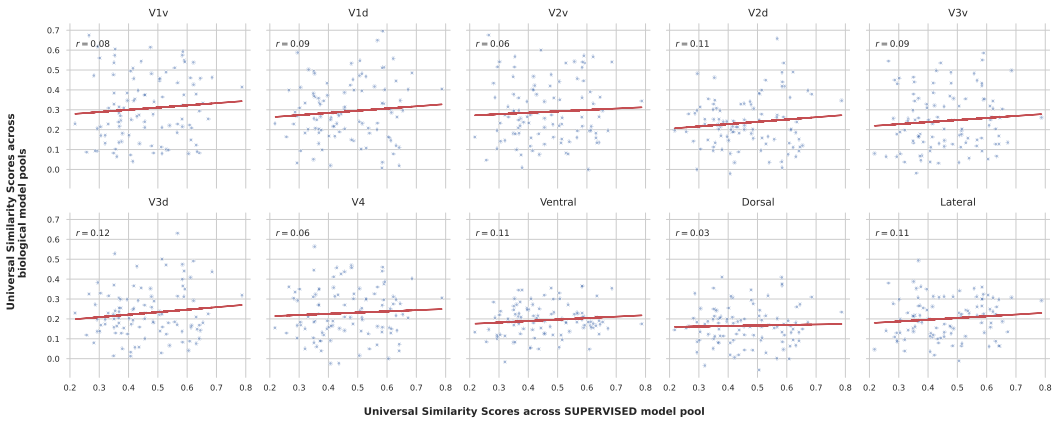


Figure 15: Comparing Universal Similarity scores obtained via biological model pools with those obtained via SUPERVISED model pool

A.5.3 Additional analyses of instance-level representational consistency

This section provides supplementary analyses extending the main results in Section 3.2 on instance-level representational consistency and its relationship to input properties. For vision, we include additional qualitative examples within a single semantic category and further compare consistency against human behavioral judgments. For language, we report complementary correlation analyses that help interpret the sentence-level results in the main text, including marginal correlations for the Greta/Tuckute sentence properties and a word-level analysis using Lancaster sensorimotor norms. Figures 16–**??** summarize these additional results.

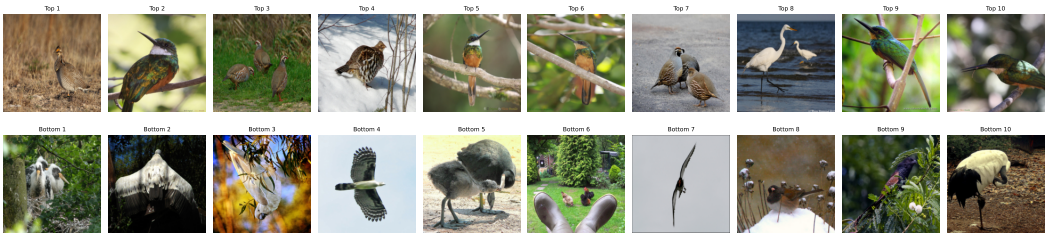


Figure 16: Representative images from the Birds parent category showing the highest and lowest universal similarity scores, derived from a pooled ensemble of supervised models with varied architectures and scales.

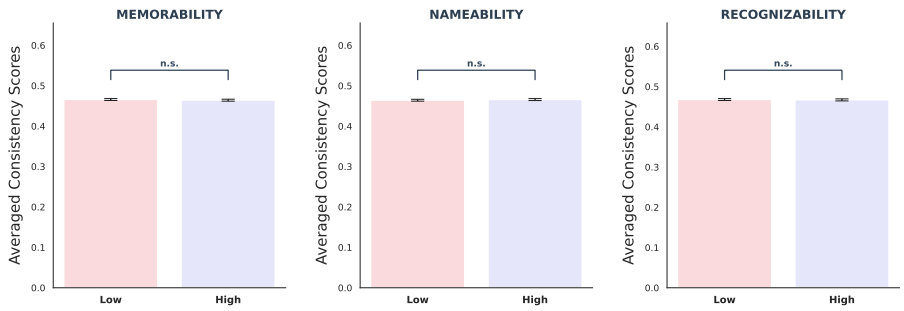


Figure 17: Statistical analysis of instance-level representational consistency as a function of human behavioral metrics (memorability, nameability, and recognizability). Paired t-tests indicate that the observed differences are not statistically significant.

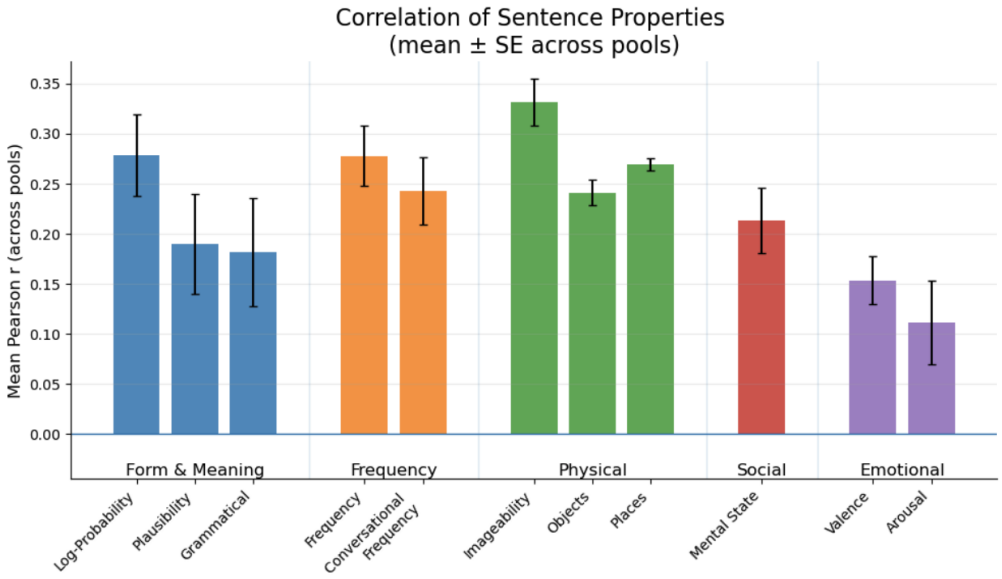


Figure 18: Marginal correlations between instance-level consistency and sentence-level properties from the Greta/Tuckute dataset, shown as mean Pearson r ± SE across model pools. This analysis is complementary to the partial-correlation analysis in the main paper: it captures the overall association of each property with consistency rather than its unique effect after controlling for the others. Notably, all properties are positively correlated with consistency.

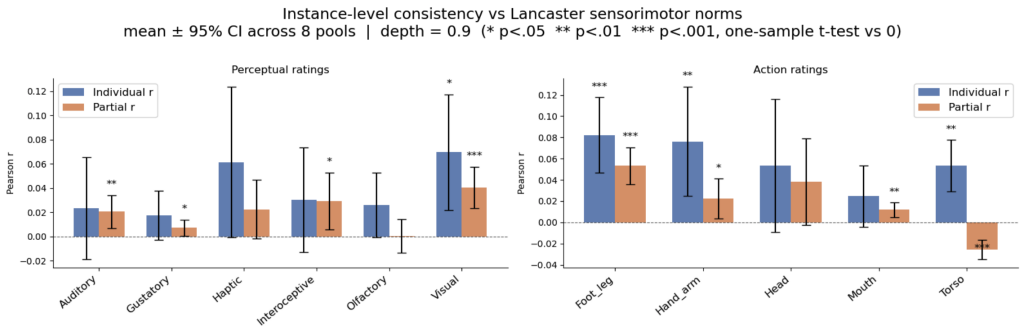


Figure 19: Correlations between instance-level consistency and Lancaster sensorimotor norms across 8 model pools. The left panel shows perceptual dimensions (auditory, gustatory, haptic, interoceptive, olfactory, visual), and the right panel shows action dimensions (foot/leg, hand/arm, head, mouth, torso). Bars show mean ± 95% CI across pools; asterisks indicate significance across pools. The results suggest that consistency is systematically related to grounded semantic content, with especially strong effects for visual and action-related dimensions.

Table 10: Cross-model correlation, RMS score, and stimulus retrieval accuracy for a vision model pool comprising of Resnet and VIT architectures with representations from different depths

Depth	Corr. ↑	RMS ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑
0.25	0.7636	0.0148	0.8445	0.9384	0.9602
0.50	0.6848	0.0189	0.9599	0.9923	0.9964
0.75	0.6001	0.0257	0.9815	0.9969	0.9985
1.0	0.5443	0.0191	0.8891	0.9766	0.9880

A.5.4 Depth analysis

To assess how barycentric alignment depends on representational depth, we repeated the alignment procedure using features extracted from multiple normalized depths within each model. For vision experiments, we considered depths 0.25, 0.5, 0.75, and 1.0 in a mixed ResNet–ViT pool. For language experiments, we evaluated autoregressive language-model pools at normalized depths 0.25, 0.5, 0.75, 0.9, and 1.0. For cross-modal experiments, we varied language and vision depths jointly in the OpenLLaMA–ViT pool. Table 15 and Figures 20–26 summarize how the quality of the learned universal space and the resulting instance-level consistency scores vary with depth.

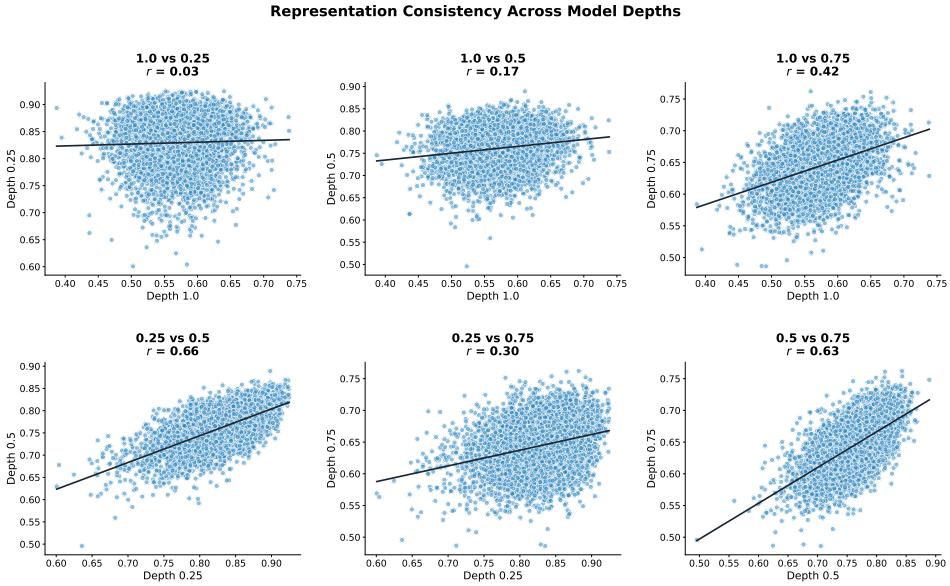


Figure 20: Pairwise correlations of instance-level representational consistency scores across RESNET–ViT model pools at varying depths (0.25, 0.5, 0.75, 1.0). Correlations increase as the difference in model depth decreases, indicating greater agreement between more similar depths.

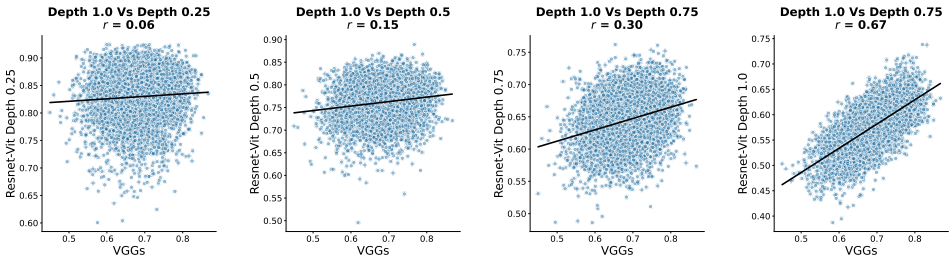


Figure 21: Pairwise correlations of instance-level representational consistency scores between RESNET–ViT model pools at varying depths and VGG model pools using penultimate-layer representations. Correlations are highest when the RESNET–ViT representations are drawn from depths closest to the penultimate layer.

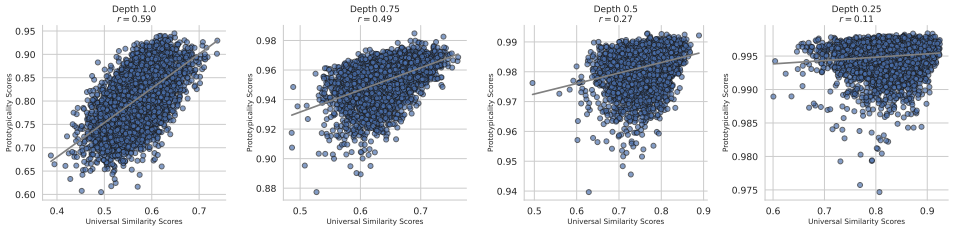


Figure 22: Comparison of instance-level consistency and prototypicality scores in the ResNet–ViT model pool. The correlation between these measures increases with model depth.

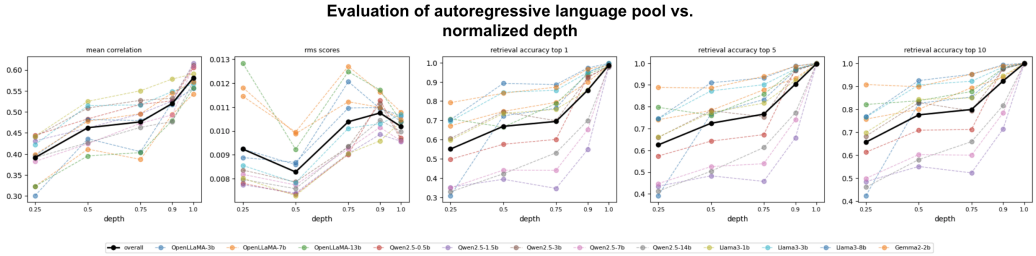


Figure 23: Effect of representation depth on the quality of the universal space in the autoregressive language-model pool. Hyperalignment was trained separately using representations extracted at normalized depths 0.25, 0.5, 0.75, 0.9, and 1.0, and evaluated using mean correlation, RMS, and Top-k retrieval (k=1,5,10). Later depths generally yield stronger universal spaces, with higher mean correlation and retrieval accuracy, indicating that alignment becomes more coherent in deeper layers.

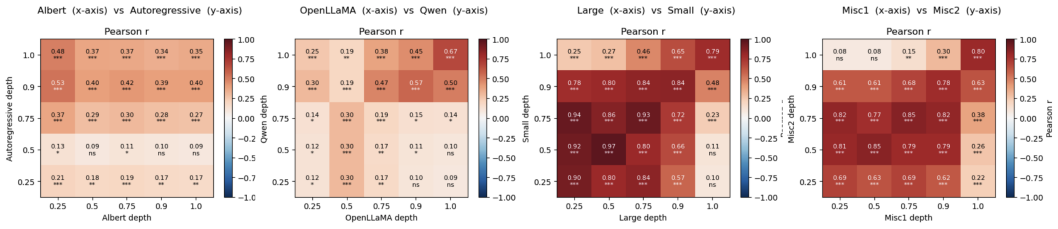


Figure 24: Pairwise correlations of instance-level consistency scores across language-model pools and normalized depths. For pools built from different architectures or model families (e.g., ALBERT vs. autoregressive, OpenLLaMA vs. Qwen), correlations are generally only stronger at later depths. For mixed and random pools, consistency scores are most similar when the compared spaces are constructed from matching depths (diagonal structure). The final layer (depth 1.0) is less correlated with neighboring depths than depth 0.9, consistent with the idea that the last layer is more specialized to the language-model output objective. In the main results from the paper, we used depth 0.9 for language model pools.

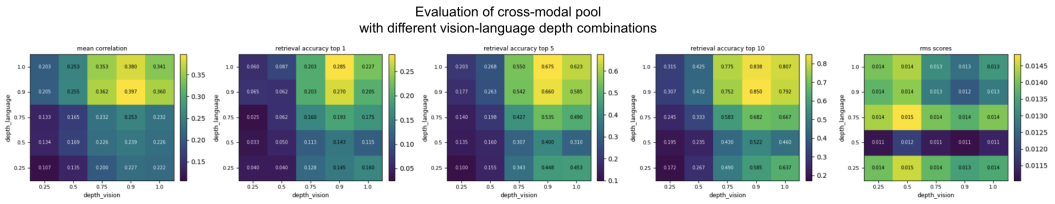


Figure 25: Effect of vision–language representation depth on the quality of the shared cross-modal space in the OpenLLaMA–ViT pool. Evaluation of universal spaces learned from different combinations of normalized language and vision depths, using mean correlation, RMS, and Top-k retrieval accuracy. Deeper depth combinations in both modalities generally yield stronger cross-modal universal spaces, reflected in higher mean correlation and retrieval performance.

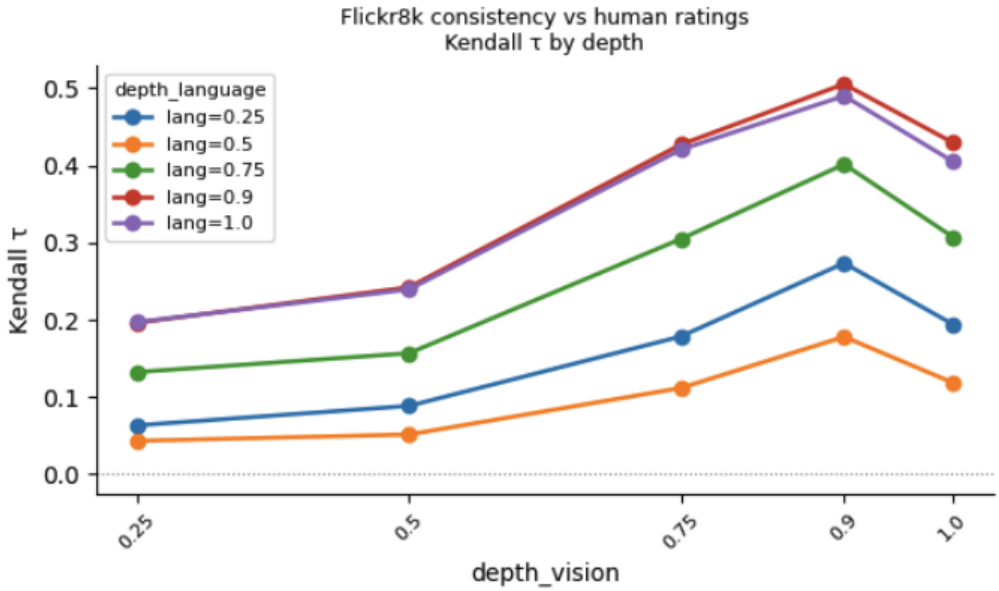


Figure 26: Agreement with Flickr8k-Expert human judgment ratings, measured by Kendall's τ, across the same vision–language depth combinations. Later depth combinations show higher agreement with human judgments, suggesting that deeper representations produce cross-modal consistency scores that are more aligned with human semantic similarity.

A.5.5 Training-data-size analysis

To assess how barycentric alignment depends on the amount of data used to estimate the shared space, we repeated the alignment procedure with varying training-set sizes. For the vision setting, we varied the amount of training data used to construct the universal space for the ResNet–ViT pool and evaluated the resulting shared space using cross-model correlation, RMS, and stimulus retrieval accuracy. For the cross-modal setting, we varied the number of paired MS-COCO training examples used to learn the shared space for the OpenLLaMA–ViT pool, and evaluated both cross-modal retrieval accuracy on a fixed MS-COCO val2017 test set and agreement with Flickr8k-Expert human judgments. Figures 27–29 summarize these results.

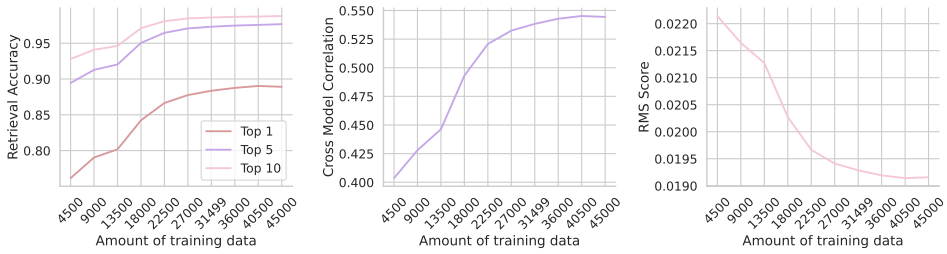


Figure 27: Average cross-model correlation, RMS score, and stimulus retrieval accuracy for the RESNET-VIT model pool, evaluated across all models with varying amounts of training data. Performance improves as the training data increases and begins to plateau when approximately 80% of the data is used.

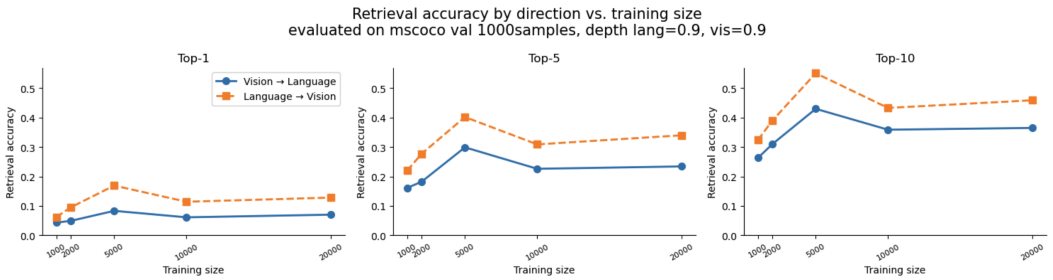


Figure 28: Cross-modal retrieval accuracy as a function of hyperalignment training size for the OpenLLaMA + ViT pool, evaluated on a fixed MS-COCO val2017(n=1000). Retrieval improves from 1k to 5k training pairs, peaks around 5k, and remains relatively stable thereafter, suggesting that the learned cross-modal space saturates with a moderate number of paired examples rather than continuing to improve with larger alignment sets.

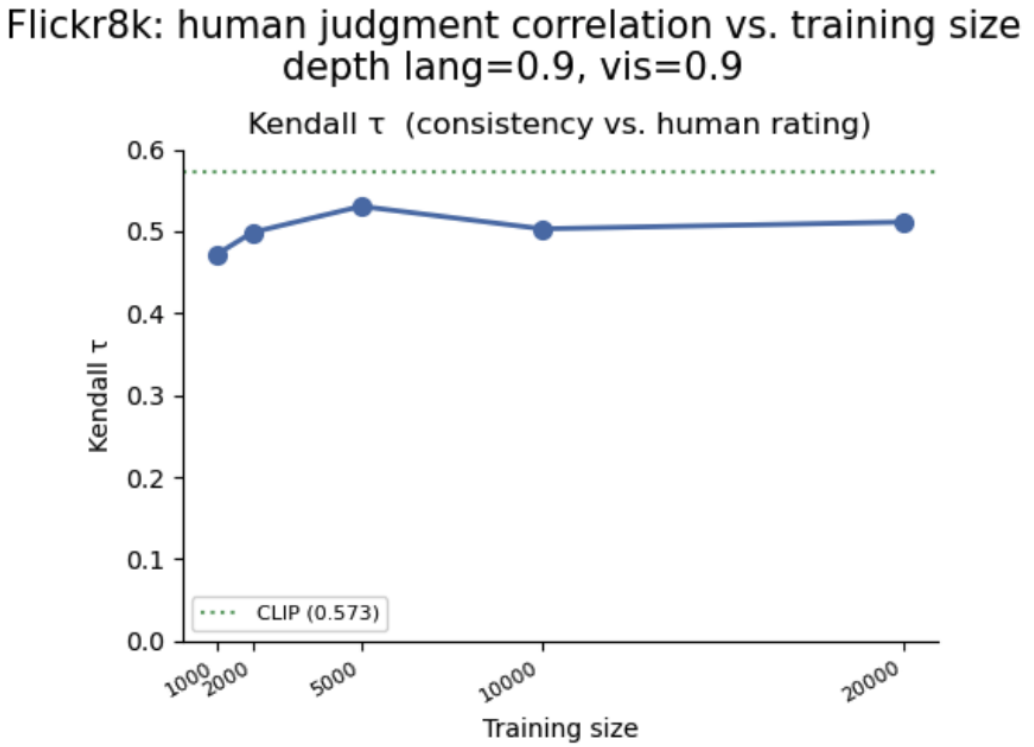


Figure 29: Agreement with Flickr8k-Expert human judgment ratings, measured by Kendall's tau, as a function of hyperalignment training size for the OpenLLaMA + ViT pool.

A.5.6 SANITY CHECKS ON TRAINING DATA COMPOSITION

We investigate the extent to which the composition of the training stimulus set influences the prediction of universal similarity scores. Specifically, we compare universal similarity scores computed for a given category of test images (e.g., birds, animals) under two training conditions: (i) a category-specific training set containing only images from that category, and (ii) a diverse training set spanning multiple categories, reflecting the standard training regime. We find that universal similarity scores for a given category are highly correlated across these two conditions, indicating that category-level universal similarity is largely invariant to the presence of additional, unrelated categories in the training diet (Figure 30, 31).

We further examine a more stringent setting in which a particular category is entirely absent from the training set. In this case, universal similarity scores for that category are largely uncorrelated with scores obtained when the category is included during training. This result indicates that the method does not generalize reliably to strongly out-of-distribution stimuli (Figure 32, 33).

Taken together, these findings suggest that universal similarity scores are robust to training-set diversity within a category's support but degrade substantially when evaluated on categories that are completely excluded from the training distribution.

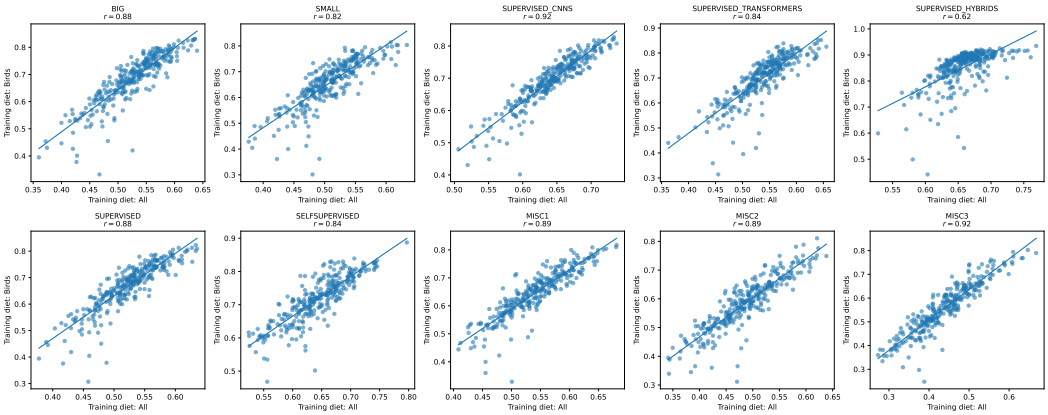


Figure 30: Universality similarity scores for test images belonging to the bird parent category from the ImageNet validation set. When the transformation matrices are trained exclusively on bird images, the resulting universality scores are highly correlated with those obtained when transformation matrices are trained on images from all categories.

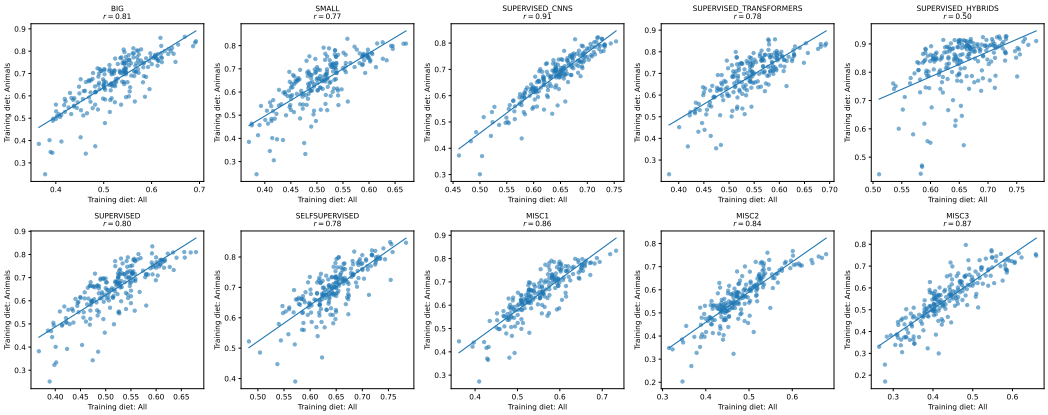


Figure 31: Universality similarity scores for test images belonging to the animal parent category from the ImageNet validation set. When the transformation matrices are trained exclusively on animal images, the resulting universality scores are highly correlated with those obtained when transformation matrices are trained on images from all categories.

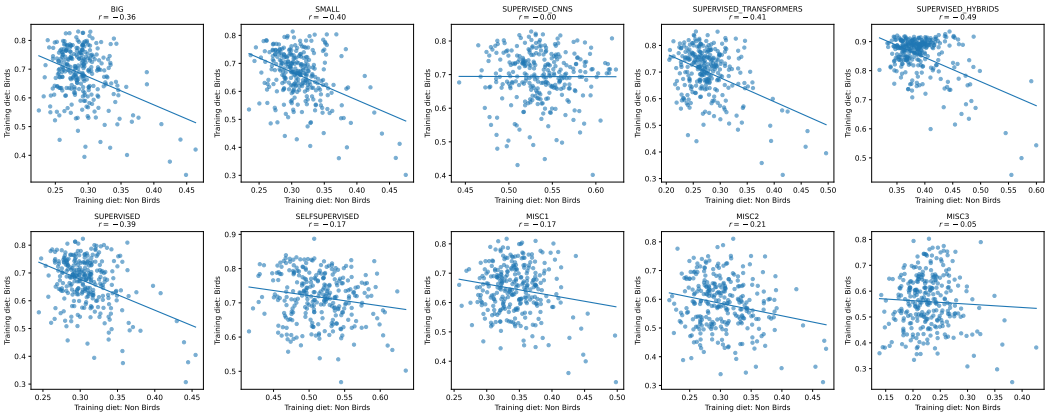


Figure 32: Universality similarity scores for test images belonging to the bird parent category from the ImageNet validation set. When the transformation matrices are trained exclusively on bird images, the resulting universality scores are highly uncorrelated with those obtained when transformation matrices are trained on datasets that exclude bird images.

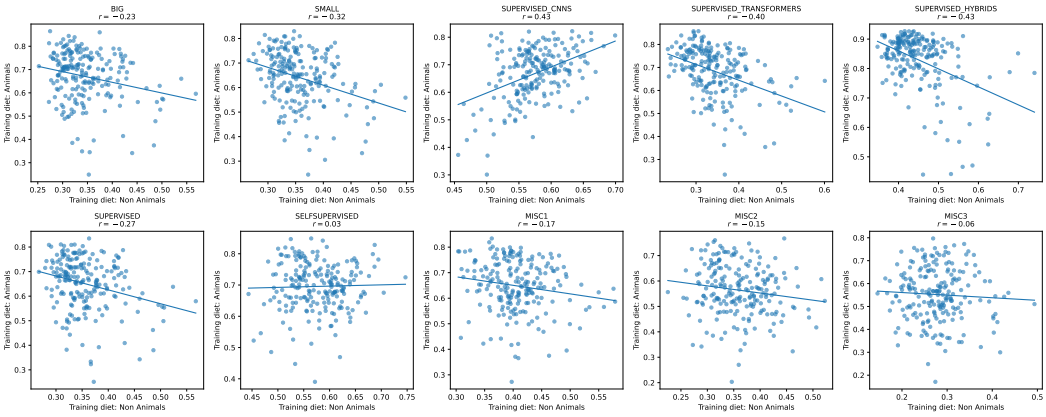


Figure 33: Universality similarity scores for test images belonging to the animal parent category from the ImageNet validation set. When the transformation matrices are trained exclusively on animal images, the resulting universality scores are highly uncorrelated with those obtained when transformation matrices are trained on datasets that exclude animal images.

A.5.7 CROSS-MODAL ROBUSTNESS ACROSS UNIMODAL MODEL POOLS

To further test the robustness of the cross-modal result in Section 3.3, we examine whether post-hoc cross-modal alignment depends on the specific choice of unimodal model pool and how performance varies with model scale and capability. In particular, we compare the original OpenLLaMA-based pool against Qwen-based pools, and evaluate how cross-modal retrieval and agreement with human judgment change as the language and vision models in the pool become stronger. Figures 34–36 show that the cross-modal result is not tied to a single LLM family and generally improves with model capability.

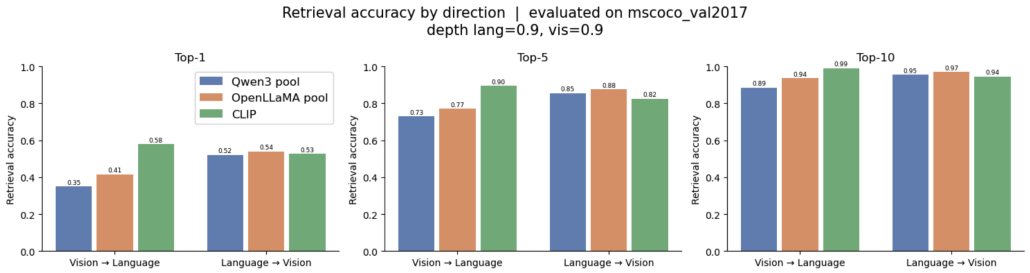


Figure 34: Cross-modal retrieval accuracy on the MS-COCO 2017 validation set for different model pools and retrieval directions. (text→image and image→text; top-k, k $\in 1, 5, 10$). We compare a Qwen-based unimodal pool (Qwen3-8b, Qwen3-14B, ViT-L/14 DINOv2, and ViT-g/14 DINOv2) with the original OpenLLaMA-based unimodal pool and CLIP from the main paper. The Qwen-based pool shows retrieval performance comparable to the original OpenLLaMA-based pool, particularly in the text → image direction, indicating that the cross-modal result is not tied to a single LLM family.

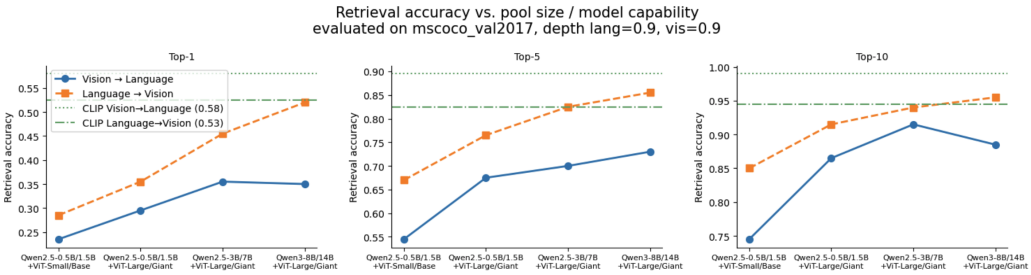


Figure 35: Cross-modal retrieval accuracy with different language/vision pool size and model capability, evaluated on MS-COCO val2017 held-out set. Each point corresponds to a different cross-modal pool constructed from increasingly larger Qwen language models and ViT vision models. Retrieval improves as model size/capability increases, with similarly sized pools yielding comparable performance and larger pools approaching CLIP, especially in the Language → Vision direction.

A.5.8 ADDITIONAL PER-MODEL EVALUATION RESULTS FOR REPRESENTATIVE POOLS

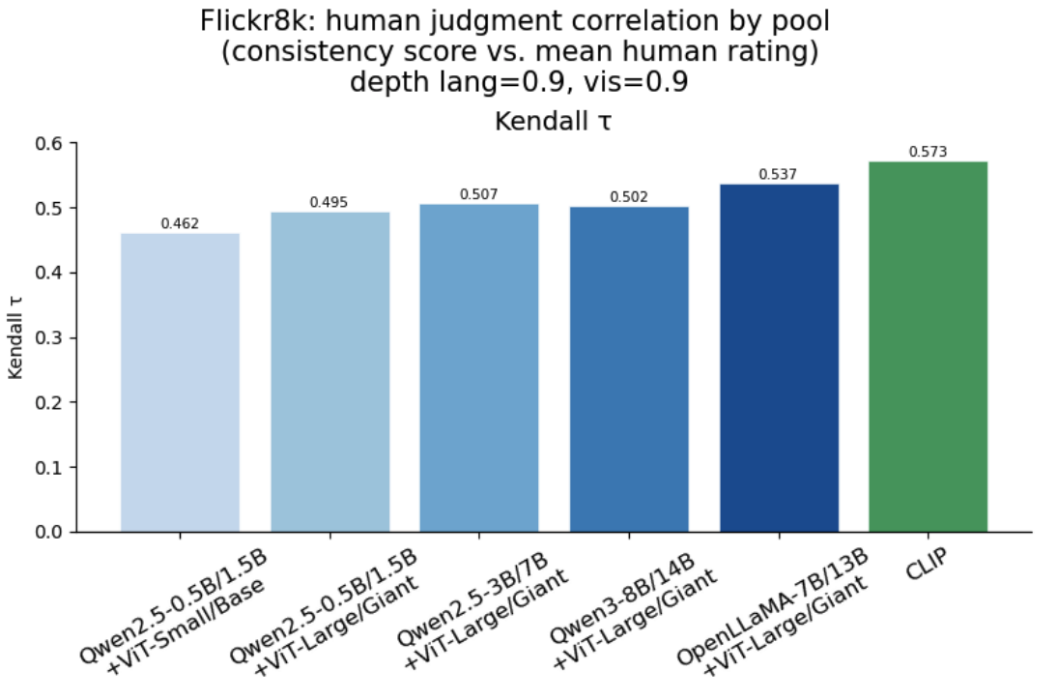


Figure 36: Agreement with Flickr8k-Expert human judgment ratings, measured by Kendall's tau, for cross-modal spaces learned from different language-model pools and compared against CLIP. We evaluate Qwen-based pools of increasing size, the original OpenLLaMA-based pool, and CLIP. Human agreement generally improves with model size, with larger Qwen pools and the OpenLLaMA pool outperforming smaller ones and approaching the performance of CLIP. This pattern is consistent with the retrieval results and suggests that stronger unimodal models yield a shared cross-modal space that is more aligned with human judgments.

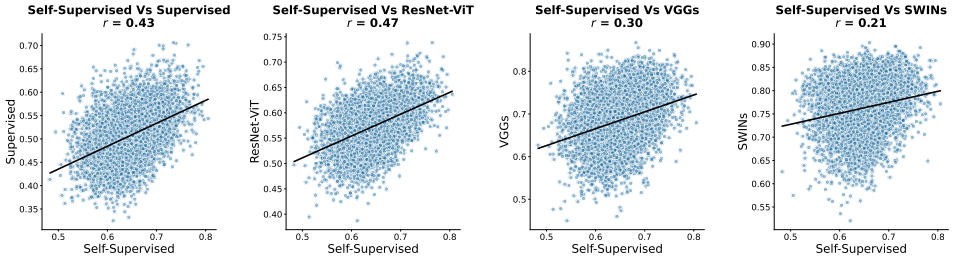


Figure 37: Pairwise correlations of instance-level representational consistency scores between the self-supervised model pool and other model pools (Supervised, ResNet–ViT, VGG, and Swin). The self-supervised pool comprises ResNet and ViT architectures trained using MoCo and DINO, which overlap significantly with the ResNet–ViT pool, resulting in the highest correlations. Correlation with the supervised pool is slightly lower, likely due to the inclusion of a broader range of architectures beyond ResNets and ViTs. In contrast, correlations with the VGG and Swin pools are substantially lower, reflecting the absence of shared architectures.

Table 11: Cross-model correlation, RMS score, and stimulus retrieval accuracy for the supervised vision model pool and the autoregressive language model pool.

Model	**Corr.** ↑	**RMS** ↓	**Top-1** ↑	**Top-5** ↑	**Top-10** ↑
Supervised Vision Model Pool					
ViT-B/16	0.4616	0.0141	0.8112	0.9712	0.9908
ViT-S/16	0.4693	0.0140	0.8357	0.9723	0.9893
ViT-T/16	0.4011	0.0149	0.6762	0.8654	0.9113
ViT-L/16	0.4674	0.0140	0.7175	0.9404	0.9739
ConvNeXt T	0.4398	0.0142	0.7695	0.9516	0.9771
ConvNeXt S	0.4931	0.0135	0.7538	0.9514	0.9781
ConvNeXt B	0.4967	0.0133	0.7291	0.9426	0.9753
ConvNeXt L	0.4933	0.0135	0.6815	0.9267	0.9687
Swin-T	0.5135	0.0134	0.8550	0.9797	0.9932
Swin-B	0.5113	0.0133	0.7214	0.9452	0.9805
Swin-S	0.5047	0.0134	0.7489	0.9556	0.9851
ResNet-18	0.5102	0.0135	0.8719	0.9714	0.9868
ResNet-34	0.5224	0.0133	0.8893	0.9811	0.9921
ResNet-50	0.5534	0.0129	0.9106	0.9871	0.9950
ResNet-101	0.5546	0.0128	0.9068	0.9872	0.9951
ResNet-152	0.5557	0.0128	0.9010	0.9869	0.9950
VGG-11	0.4708	0.0144	0.8677	0.9771	0.9908
VGG-13	0.4675	0.0145	0.8638	0.9784	0.9916
VGG-16	0.4789	0.0143	0.8522	0.9774	0.9917
VGG-19	0.4831	0.0142	0.8392	0.9756	0.9912
AlexNet	0.3670	0.0161	0.7955	0.9399	0.9669
AlexNet	0.3495	0.015 35	0.7576	0.8952	0.9209
Autoregressive Language Model Pool					
OpenLLaMA-3b	0.5231	0.010 94	0.9491	0.9679	0.9761
OpenLLaMA-7b	0.4942	0.011 62	0.8988	0.9309	0.9385
OpenLLaMA-13b	0.4795	0.011 71	0.9406	0.9721	0.9785
Qwen2.5-0.5b	0.5260	0.011 29	0.9261	0.9721	0.9815
Qwen2.5-1.5b	0.5257	0.009 85	0.5479	0.6573	0.7136
Qwen2.5-3b	0.5323	0.011 15	0.9215	0.9648	0.9742
Qwen2.5-7b	0.4929	0.010 15	0.6518	0.7406	0.7836
Qwen2.5-14b	0.4771	0.010 44	0.6991	0.7727	0.8161
Llama3-1b	0.5781	0.009 58	0.8570	0.9209	0.9448
Llama3-3b	0.5483	0.010 30	0.9491	0.9736	0.9833
Llama3-8b	0.5170	0.010 98	0.9727	0.9864	0.9927
Gemma2-2b	0.5450	0.010 98	0.9670	0.9861	0.9885

Table 12: Correlation score, RMS score, and retrieval accuracies across different vision model pools.

Model	Corr. ↑	RMS ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑
Self Supervised Model Pool					
Resnet 50 Dino	0.5726	0.0187	0.9889	0.9985	0.9995
Resnet 50 Moco v2	0.5299	0.0197	0.9760	0.9972	0.9990
Vit B 16 Dino	0.6429	0.0166	0.9669	0.9934	0.9980
Vit S 16 Dino	0.6199	0.0170	0.9732	0.9952	0.9985
Vit B 8 Dino	0.6198	0.0172	0.9647	0.9941	0.9981
ResNet Model Pool					
Resnet 34	0.6959	0.0163	0.9964	0.9999	1.0000
Resnet 50	0.7270	0.0154	0.9957	1.0000	1.0000
Resnet 18	0.6775	0.0167	0.9928	1.0000	1.0000
Resnet 101	0.7252	0.0154	0.9915	0.9999	1.0000
Resnet 152	0.7239	0.0155	0.9880	0.9998	1.0000
ConvNext Model Pool					
Convnext tiny	0.6325	0.0193	0.9513	0.9974	0.9989
Convnext small	0.6947	0.0178	0.9607	0.9973	0.9991
Convnext base	0.7059	0.0173	0.9595	0.9981	0.9997
Convnext large	0.6964	0.0178	0.9518	0.9977	0.9999
Swin Transformers Model Pool					
swin T	0.7403	0.0221	0.9858	0.9985	0.9998
swin B	0.7682	0.0209	0.9675	0.9987	1.0000
swin S	0.7690	0.0208	0.9822	0.9989	1.0000
Vision Transformers Model Pool					
vit B 16	0.4243	0.0309	0.8593	0.9595	0.9768
vit S 16	0.4384	0.0294	0.9147	0.9823	0.9907
vit T 16	0.3360	0.0316	0.7233	0.9024	0.9379
vit L 16	0.4287	0.0298	0.7850	0.9215	0.9481
VGG Model Pool					
VGG 11	0.6768	0.0125	0.9948	1.0000	1.0000
VGG 13	0.6779	0.0125	0.9949	1.0000	1.0000
VGG 16	0.6843	0.0123	0.9893	0.9999	1.0000
VGG 19	0.6790	0.0124	0.9805	0.9998	1.0000
Supervised CNNS Model Pool					
resnet 34	0.5973	0.0130	0.9653	0.9966	0.9992
resnet 50	0.6332	0.0124	0.9623	0.9971	0.9991
resnet 18	0.5960	0.0131	0.9762	0.9985	0.9997
resnet 101	0.6265	0.0125	0.9439	0.9943	0.9982
resnet 152	0.6251	0.0126	0.9334	0.9937	0.9980
VGG 11	0.5894	0.0134	0.9743	0.9990	0.9998
VGG 13	0.5829	0.0135	0.9673	0.9985	0.9997
VGG 16	0.5891	0.0134	0.9543	0.9968	0.9993
VGG 19	0.5885	0.0134	0.9399	0.9958	0.9986
alexnet	0.4686	0.0155	0.9383	0.9938	0.9978

Model	Corr. ↑	RMS ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑
Supervised Transformers Model Pool					
vit B 16	0.5059	0.0298	0.8810	0.9730	0.9876
vit S 16	0.4528	0.0309	0.8858	0.9816	0.9927
vit T 16	0.3257	0.0340	0.6342	0.8538	0.9087
vit L 16	0.4802	0.0301	0.8074	0.9519	0.9728
swin T	0.5505	0.0283	0.8853	0.9726	0.9866
swin B	0.5604	0.0279	0.7905	0.9468	0.9744
swin S	0.5571	0.0279	0.8142	0.9549	0.9787
Big Model Pool					
vit B 16	0.5059	0.0298	0.8810	0.9730	0.9876
vit S 16	0.4528	0.0309	0.8858	0.9816	0.9927
vit T 16	0.3257	0.0340	0.6342	0.8538	0.9087
vit L 16	0.4802	0.0301	0.8074	0.9519	0.9728
swin T	0.5505	0.0283	0.8853	0.9726	0.9866
swin B	0.5604	0.0279	0.7905	0.9468	0.9744
swin S	0.5571	0.0279	0.8142	0.9549	0.9787
Small Model Pool					
resnet34	0.5171	0.0125	0.8954	0.9785	0.9900
resnet50	0.5380	0.0123	0.9110	0.9843	0.9929
resnet18	0.5069	0.0127	0.8805	0.9695	0.9849
vgg11	0.4390	0.0142	0.8687	0.9761	0.9898
vgg13	0.4351	0.0142	0.8649	0.9766	0.9901
swin T	0.4908	0.0129	0.8529	0.9770	0.9909
swin S	0.4743	0.0131	0.7381	0.9527	0.9820
vit S 16	0.4591	0.0132	0.8390	0.9711	0.9878
vit T 16	0.4001	0.0139	0.7027	0.8748	0.9169
convnext tiny	0.4285	0.0134	0.7537	0.9384	0.9670
convnext small	0.4704	0.0129	0.7363	0.9434	0.9713
Misc1 Model Pool					
resnet34	0.4905	0.0197	0.9217	0.9892	0.9965
Resnet 50 Dino	0.4855	0.0204	0.9442	0.9915	0.9971
swin B	0.4380	0.0212	0.7228	0.9673	0.9918
vit B 16 dino	0.5463	0.0186	0.9425	0.9923	0.9982
Misc2 Model Pool					
vgg13	0.4401	0.0148	0.8651	0.9721	0.9873
Resnet 50 Moco v2	0.4199	0.0147	0.8631	0.9671	0.9851
vit S 16 dino	0.4929	0.0135	0.9249	0.9893	0.9973
swin T	0.4234	0.0146	0.7706	0.9539	0.9847
Misc3 Model Pool					
alexnet	0.3465	0.0156	0.7991	0.9412	0.9706
vit S 16 dino	0.4227	0.0137	0.8632	0.9777	0.9928
swin S	0.3654	0.0145	0.6408	0.9059	0.9633

Table 13: Average performance across brain model pools (averaged over subjects 1, 2, 5, and 7).

Subjects	Corr. ↑	RMS ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑	Chance-10
V1v Model Pool						
1	0.2976	0.0475	0.2833	0.5133	0.6433	0.1167
2	0.2974	0.0473	0.3033	0.5233	0.6267	0.1100
5	0.2600	0.0485	0.1800	0.3767	0.5133	0.0967
7	0.2515	0.0489	0.2067	0.4167	0.5333	0.0567
V1d Model Pool						
1	0.2835	0.0425	0.3133	0.4833	0.5967	0.0867
2	0.2890	0.0419	0.3067	0.5133	0.6233	0.1100
5	0.2326	0.0436	0.1433	0.3700	0.4833	0.0833
7	0.2326	0.0436	0.2567	0.4267	0.5333	0.1233
V2v Model Pool						
1	0.2449	0.0413	0.2467	0.5200	0.6600	0.0900
2	0.2420	0.0408	0.2800	0.5133	0.6400	0.1067
5	0.2175	0.0413	0.1367	0.3700	0.4900	0.0833
7	0.2291	0.0411	0.2000	0.4633	0.5700	0.0800
V2d Model Pool						
1	0.2344	0.0496	0.2500	0.4433	0.5500	0.0900
2	0.2216	0.0499	0.1967	0.4133	0.5600	0.1100
5	0.1769	0.0515	0.1167	0.2833	0.4100	0.1233
7	0.2024	0.0504	0.1533	0.3567	0.5000	0.0800
V3v Model Pool						
1	0.2318	0.0482	0.2033	0.4767	0.6000	0.0733
2	0.2244	0.0482	0.2267	0.4700	0.5833	0.1067
5	0.1936	0.0490	0.1133	0.3333	0.4933	0.1167
7	0.2013	0.0487	0.1467	0.3667	0.4833	0.0967
V3d Model Pool						
1	0.2338	0.0526	0.1800	0.4400	0.5633	0.0967
2	0.2308	0.0527	0.2100	0.4067	0.5567	0.0833
5	0.1776	0.0544	0.1100	0.3033	0.4000	0.0800
7	0.1981	0.0537	0.1433	0.3267	0.4733	0.0967
v4 Model Pool						
1	0.2047	0.0476	0.2100	0.4500	0.5567	0.0800
2	0.2133	0.0467	0.2200	0.4500	0.6067	0.0800
5	0.1831	0.0478	0.1733	0.3733	0.4933	0.0867
7	0.1801	0.0478	0.1867	0.3933	0.5200	0.0900
Ventral Visual Stream Model Pool						
1	0.1573	0.0156	0.2433	0.5367	0.6967	0.1000
2	0.1668	0.0156	0.2133	0.5300	0.6733	0.1067
5	0.1622	0.0155	0.1733	0.4833	0.6467	0.1167
7	0.1477	0.0157	0.1833	0.4633	0.6567	0.1067
Dorsal Visual Stream Model Pool						
1	0.1540	0.0225	0.1133	0.3100	0.4533	0.1300
2	0.1481	0.0225	0.1400	0.3600	0.4900	0.0833
5	0.1577	0.0223	0.1200	0.3167	0.4367	0.1133
7	0.1390	0.0226	0.0933	0.2833	0.4033	0.0900
Lateral Visual Stream Model Pool						
1	0.1697	0.0154	0.1467	0.4000	0.5767	0.1033
2	0.1782	0.0152	0.1400	0.3800	0.5733	0.0867
5	0.1746	0.0153	0.1400	0.3833	0.5600	0.0933
7	0.1588	0.0154	0.1300	0.3933	0.5900	0.0833

Table 14: Average performance (Normalised Correlation, Procrustes Residual, Heldout Residual, RMS score, and stimulus retrieval accuracy) across NSD subjects for each region pool (Chance level retrieval scores are 0.01, 0.05 and 0.10 respectively for top-1, top-5 and top-10 respectively.

Region	**Norm Corr.**	**Procrustes Res.**	**Heldout Res.**	**RMS**	**Top-1**	**Top-5**	**Top-10**
V1v	0.465	0.353	0.520	0.048	0.243	0.458	0.579
V1d	0.421	0.344	0.531	0.043	0.255	0.448	0.559
V2v	0.418	0.342	0.533	0.041	0.216	0.467	0.590
V2d	0.378	0.389	0.573	0.050	0.179	0.374	0.505
V3v	0.360	0.367	0.566	0.049	0.173	0.412	0.540
V3d	0.384	0.396	0.578	0.053	0.161	0.369	0.498
V4	0.366	0.373	0.578	0.048	0.198	0.417	0.544
Ventral	0.460	0.136	0.605	0.016	0.203	0.503	0.668
Dorsal	0.361	0.277	0.625	0.023	0.117	0.318	0.446
Lateral	0.436	0.212	0.600	0.015	0.139	0.389	0.575

Table 15: Quantitative evaluation of the learned barycentric embedding as a function of the number of diverse model pools used during training. Different subsets of the supervised model pool are considered.

Model	**Corr.** ↑	**RMS** ↓	**Top-1** ↑	**Top-5** ↑	**Top-10** ↑
Subset 1					
ViT-B/16	0.4587	0.0129	0.8229	0.9811	0.9952
ConvNeXt T	0.4566	0.0127	0.7830	0.9696	0.9914
Swin-T	0.5011	0.0124	0.8668	0.9880	0.9985
ResNet-34	0.4704	0.0129	0.8662	0.9747	0.9892
VGG-11	0.4127	0.0144	0.8451	0.9743	0.9901
Average	0.4599	0.0130	0.8368	0.9775	0.9928
Subset 2					
ViT-B/16	0.4790	0.0134	0.8356	0.9812	0.9958
ViT-S/16	0.4727	0.0135	0.8415	0.9739	0.9901
ConvNeXt T	0.4660	0.0132	0.8008	0.9716	0.9920
ConvNeXt S	0.5144	0.0126	0.7739	0.9683	0.9910
Swin-T	0.5334	0.0126	0.8787	0.9889	0.9982
Swin-B	0.5287	0.0126	0.7460	0.9619	0.9910
ResNet-34	0.5143	0.0131	0.8856	0.9816	0.9924
ResNet-50	0.5404	0.0128	0.9094	0.9889	0.9964
VGG-11	0.4522	0.0144	0.8570	0.9769	0.9915
VGG-13	0.4501	0.0145	0.8539	0.9786	0.9933
Average	0.4951	0.0133	0.8382	0.9772	0.9932
Subset 3					
ViT-B/16	0.4633	0.0135	0.8144	0.9726	0.9911
ViT-S/16	0.4666	0.0135	0.8353	0.9726	0.9894
ViT-T/16	0.3935	0.0144	0.6651	0.8574	0.9053
ConvNeXt T	0.4485	0.0134	0.7721	0.9505	0.9759
ConvNeXt S	0.5027	0.0127	0.7575	0.9520	0.9779
ConvNeXt B	0.5052	0.0126	0.7318	0.9436	0.9749
Swin-T	0.5221	0.0127	0.8620	0.9801	0.9930
Swin-B	0.5219	0.0126	0.7322	0.9476	0.9814
Swin-S	0.5147	0.0127	0.7598	0.9581	0.9858
ResNet-34	0.5134	0.0129	0.8831	0.9786	0.9909
ResNet-50	0.5378	0.0127	0.9048	0.9854	0.9941
ResNet-18	0.5005	0.0131	0.8618	0.9676	0.9850
VGG-11	0.4479	0.0143	0.8540	0.9739	0.9895
VGG-13	0.4453	0.0144	0.8501	0.9751	0.9904
VGG-16	0.4571	0.0142	0.8402	0.9741	0.9903
Average	0.4827	0.0133	0.8083	0.9593	0.9810

Model	**Corr.** ↑	**RMS** ↓	**Top-1** ↑	**Top-5** ↑	**Top-10** ↑
All Models					
ViT-B/16	0.4616	0.0141	0.8112	0.9712	0.9908
ViT-S/16	0.4693	0.0140	0.8357	0.9723	0.9893
ViT-T/16	0.4011	0.0149	0.6762	0.8654	0.9113
ViT-L/16	0.4674	0.0140	0.7175	0.9404	0.9739
ConvNeXt T	0.4398	0.0142	0.7695	0.9516	0.9771
ConvNeXt S	0.4931	0.0135	0.7538	0.9514	0.9781
ConvNeXt B	0.4967	0.0133	0.7291	0.9426	0.9753
ConvNeXt L	0.4933	0.0135	0.6815	0.9267	0.9687
Swin-T	0.5135	0.0134	0.8550	0.9797	0.9932
Swin-B	0.5113	0.0133	0.7214	0.9452	0.9805
Swin-S	0.5047	0.0134	0.7489	0.9556	0.9851
ResNet-18	0.5102	0.0135	0.8719	0.9714	0.9868
ResNet-34	0.5224	0.0133	0.8893	0.9811	0.9921
ResNet-50	0.5534	0.0129	0.9106	0.9871	0.9950
ResNet-101	0.5546	0.0128	0.9068	0.9872	0.9951
ResNet-152	0.5557	0.0128	0.9010	0.9869	0.9950
VGG-11	0.4708	0.0144	0.8677	0.9771	0.9908
VGG-13	0.4675	0.0145	0.8638	0.9784	0.9916
VGG-16	0.4789	0.0143	0.8522	0.9774	0.9917
VGG-19	0.4831	0.0142	0.8392	0.9756	0.9912
AlexNet	0.3670	0.0161	0.7955	0.9399	0.9669
Average	0.4864	0.0138	0.8094	0.9602	0.9819