WHAT CONVERGES IN THE PLATONIC REPRESENTATION HYPOTHESIS? STRUCTURE OVER GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

The Platonic Representation Hypothesis suggests that increasingly capable models converge toward shared representations. Recent work narrows this claim to shared local neighborhood relationships, finding that capacity-dependent trends in several global similarity measures largely disappear after calibration. We challenge this interpretation by showing that prior local-global comparisons confound structural scale (local versus global) with what is compared: relational structure, defined by which samples are related, versus metric geometry, characterized by quantitative relations such as distances, similarities, or correlations. To disentangle these factors, we construct a controlled 2×2 framework that evaluates both relational structure and metric geometry at local and global scales. We introduce H_0 skeleton overlap as a global counterpart to mutual k-nearest neighbors, together with matched distance-aware variants. Across vision-language models, relational structure exhibits robust representational convergence at both scales after calibration, whereas increasingly stringent distance agreement substantially weakens alignment and progressively flattens the capacity-dependent trend. We further extend the analysis beyond ambient Euclidean geometry by evaluating distance agreement under a Riemannian metric approximation and recover the same structure-geometry pattern. The pattern is also reproduced in video-text representations. Together, these results show that relational convergence extends beyond local neighborhoods to global spanning structure, whereas metric geometry exhibits substantially weaker convergence.

1 INTRODUCTION

Whether independently trained neural networks learn similar internal representations has long been a central question in representation learning (Li et al., 2015; Morcos et al., 2018; Wang et al., 2018; Kornblith et al., 2019). Existing studies show that independently trained neural networks can share similar representational structures, while also exhibiting substantial variations depending on training conditions and similarity measures (Morcos et al., 2018; Bansal et al., 2021; Ding et al., 2021). As modern models become increasingly capable and heterogeneous, questions of representational convergence have gained renewed importance (Sucholutsky et al., 2025; Huh et al., 2024; Tjandrasuwita et al., 2025). Yet the notion of representational convergence remains underspecified: the key question is not only whether representations become more similar, but what exactly converges.

The Platonic Representation Hypothesis (PRH) proposes that increasingly capable models converge toward a shared representation (Huh et al., 2024). Correspondingly, representational alignment tends to increase with model capability across vision and language models (Huh et al., 2024). Recent work, however, shows that model scale and layer-wise aggregation can inflate raw similarity, and that aggregation-aware calibration substantially weakens the apparent convergence of several global measures (Gröger et al., 2026). In contrast, local neighborhood agreement measured by mutual k-nearest neighbors (mKNN) remains robust, motivating the view that increasingly capable models converge in their local neighborhood relationships (Gröger et al., 2026).

However, we challenge this local neighborhood interpretation because the underlying comparisons differ in structural scale and in whether they compare which samples are related or the numerical relations among them. mKNN captures local relational structure by comparing which samples are selected as neighbors, without requiring their distances to agree. In contrast, global similarity mea-

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.

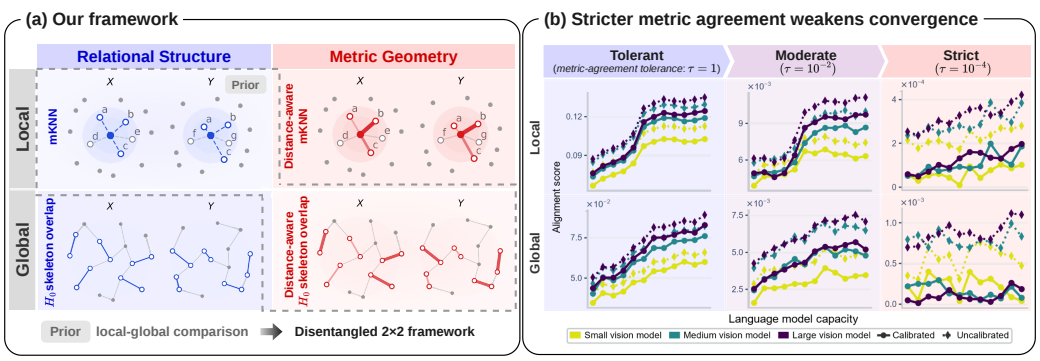


Figure 1: **Disentangling structural scale from what is compared.** (a) Prior local-global comparisons contrast local relational structure with global measures that depend on representation geometry, changing both structural scale and what is compared. Our controlled 2×2 framework disentangles these factors by comparing relational structure and metric geometry independently at local and global scales. (b) As metric-agreement requirements become more stringent, capacity-dependent convergence weakens at both local and global scales. Tolerant, moderate, and strict settings correspond to $\tau = 1$, 10^{-2}, and 10^{-4} in Eq. 3.

sures such as centered kernel alignment (CKA) (Kornblith et al., 2019), CCA-based methods such as SVCCA and PWCCA (Raghu et al., 2017; Morcos et al., 2018), and representational similarity analysis (RSA) (Kriegeskorte et al., 2008) depend on numerical relations such as similarities, correlations, or dissimilarities that encode aspects of representation geometry (Williams et al., 2021). Thus, prior local-global comparisons vary along both dimensions, making it unclear whether the observed difference reflects scale or the distinction between relational structure and metric geometry. We use *metric geometry* as a broad conceptual category for quantitative relations among samples.

To disentangle these factors, we evaluate representational alignment in relational structure and metric geometry at both local and global scales (Figure 1a). While mKNN captures local relational structure through shared neighbors, we construct a global counterpart using zero-dimensional persistent homology (H_0), which captures how disconnected components merge (Edelsbrunner et al., 2002). The corresponding H_0 death edges coincide with the edges of a minimum spanning tree (MST) (Skraba et al., 2020), allowing us to compare which edges form the global spanning structure without requiring their distances to agree. We call this measure H_0 *skeleton overlap*.

To probe metric geometry while holding relational structure fixed, we introduce distance-aware variants of both mKNN and H_0 skeleton overlap. These variants additionally require agreement in the distances associated with shared neighbors or spanning edges, with the strictness of this agreement controlled by a parameter τ in Eq. 3. As τ decreases, distance agreement becomes increasingly stringent. We further extend the analysis beyond ambient Euclidean geometry by evaluating distance agreement on the same underlying relations using a Riemannian metric approximation constructed from local covariance information (Singer & Coifman, 2008; Berry & Sauer, 2016).

Across vision-language models, alignment in relational structure remains robust after calibration and generally strengthens with model capacity at both scales, indicating that representational convergence extends beyond local neighborhoods to global spanning structure. In contrast, as distance-agreement requirements become more stringent, alignment weakens and capacity-dependent trends progressively flatten at both scales, while differences between local and global scales emerge primarily in the most stringent regime. Extending the analysis beyond Euclidean geometry to a Riemannian metric approximation yields the same structure-geometry pattern. The pattern is further reproduced in video-text representations, suggesting that it is not specific to a single multimodal setting.

Taken together, our results challenge a local neighborhood interpretation of representational convergence: relational convergence extends to global spanning structure, whereas increasingly stringent metric agreement substantially weakens convergence.

The contributions of this work are as follows:

- **Revisiting the local neighborhood interpretation of convergence.** We challenge the recent interpretation that calibrated representational convergence is captured by shared local neighbor-

hood relationships. We show that the comparisons underlying this view confound structural scale with what is compared: relational structure or metric geometry.

- **Global relational convergence in a controlled 2×2 framework.** We introduce a 2×2 framework that compares relational structure and metric geometry at both local and global scales. Using H_0 skeleton overlap as a global counterpart to mKNN, we show that relational convergence extends beyond local neighborhoods to global spanning structure.
- **Relational structure versus metric geometry.** We find that this distinction is more pronounced than the local-global distinction, with scale-dependent differences emerging mainly under stringent agreement. Relational structure shows robust convergence at both scales, whereas increasingly stringent metric agreement weakens alignment and progressively flattens the capacity-dependent trend.
- **Robustness beyond Euclidean geometry and across multimodal settings.** We extend the controlled distance-aware analysis beyond ambient Euclidean geometry using a Riemannian metric approximation and recover the same structure-geometry pattern. We further reproduce this pattern in video-text representations.

2 Related Work

2.1 Representational Similarity Measures

Representational similarity measures quantify agreement between neural representations using different numerical relations among representation vectors. Methods based on Canonical Correlation Analysis (CCA), including SVCCA and PWCCA (Raghu et al., 2017; Morcos et al., 2018), compare representation spaces through canonical correlations between linear projections. Centered Kernel Alignment (CKA) (Kornblith et al., 2019) compares kernel-based similarities, while Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) compares pairwise dissimilarities. Although these measures differ in their invariances and constructions, their scores depend on numerical relations such as correlations, similarities, or dissimilarities. Relatedly, Williams et al. (2021) formulated representation comparison through generalized shape metrics, emphasizing how the choice of comparison measure determines the geometric structure and invariances being compared.

Neighborhood-based approaches capture relational information without requiring exact distances to agree. Mutual k-Nearest Neighbors (mKNN) (Huh et al., 2024), for example, compares the identities of neighboring samples rather than the corresponding distances. Other approaches retain ordinal rather than full numerical information. Soares et al. (2026) introduced the Triplet and Quadruplet Similarity Indices (TSI and QSI) to compare relative distance rankings and established a formal connection between TSI and local neighborhood alignment. More broadly, representational similarity measures preserve and compare different aspects of a representation (Klabunde et al., 2025). Consequently, an increase in a particular similarity score does not by itself imply convergence in all aspects of representation geometry; interpreting convergence requires identifying which relations a measure preserves and which geometric information it additionally compares.

2.2 The Platonic Representation Hypothesis and Its Recent Critiques

The Platonic Representation Hypothesis (PRH) (Huh et al., 2024) proposes that increasingly capable models converge toward a shared statistical representation of the underlying world despite differences in architecture, training data, or modality. Empirical results across multimodal models have supported this view by showing increasing representational alignment with model capability.

However, recent work questions whether raw alignment scores provide direct evidence of representational convergence. Gröger et al. (2026) showed that model scale and layer-wise aggregation can inflate similarity scores. Using permutation-based null calibration, they found that scaling trends in CKA, SVCCA, and Procrustes distance largely disappear after calibration, whereas neighborhood-based measures such as mKNN, cycle-kNN, and CKNNA retain clear scaling trends. They further showed that models increasingly agree on local neighbor identities, while the corresponding pairwise distances do not exhibit the same alignment.

This finding establishes an important distinction between relational structure and distance agreement at the local scale, but leaves open how to interpret the contrast between robust local neighborhood

agreement and weakened global similarity. In particular, the corresponding measures differ not only in structural scale, but also in what they compare. Our work addresses this ambiguity by evaluating relational structure and metric geometry at both local and global scales. We pair local mKNN with global H_0 skeleton overlap to compare relational structure and construct matched distance-aware variants to assess metric agreement while preserving the underlying relations. We further extend this controlled analysis beyond ambient Euclidean geometry using a Riemannian metric approximation.

2.3 Topological Comparison and Alignment of Representation Spaces

Persistent homology (PH) has been used to preserve and compare structural information in learned representations, including topology-preserving representation learning (Moor et al., 2020; Kim et al., 2024) and direct representation comparison through RTD (Barannikov et al., 2021). RTD-Lite further uses minimum spanning trees (MSTs) to efficiently compare connectivity structure across weighted graphs (Tulchinskii et al., 2025). The connection between H_0 persistence and MSTs is well established (Kruskal, 1956; Skraba et al., 2020). Our H_0 skeleton overlap discards filtration values and compares only the identities of labeled spanning edges, thereby isolating global relational structure from the distances assigned to those edges.

Topology has also been used as an alignment signal in multimodal representation learning. Homology Consistency, ToMCLIP, and ToMA use PH-derived structure to align or regularize vision-language representations (Zhang et al., 2024; You et al., 2026a;b). In contrast, our goal is diagnostic rather than optimization-based: we use H_0 skeleton overlap as a global counterpart to local mKNN for comparing relational structure, and construct matched distance-aware variants at both scales to probe metric agreement while preserving the underlying relations.

3 A Controlled Framework for Representational Convergence

We develop a controlled framework that separates structural scale from what is compared. At each scale, we first compare relational structure through the identities of relations selected in each representation, and then probe metric geometry by incorporating distance agreement without changing those relations. This construction enables matched comparisons at local and global scales while varying whether alignment is evaluated only in relational structure or additionally in the distances assigned to the selected relations.

3.1 Problem Setup and Comparison Framework

Let $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^n$ denote two representations of the same n samples, with pairwise distance matrices D_X and D_Y, respectively. We distinguish *structural scale*, separating local relations from global spanning structure. At each scale, relational structure is represented by the identities of selected relations, whereas metric geometry is probed through the distances assigned to those same relations. This yields the 2×2 framework in Figure 1a, allowing us to vary structural scale while holding what is compared fixed, and to introduce distance agreement without changing the underlying relations.

3.2 Relational Structure Alignment

We first compare relational structure using only the identities of selected relations, without requiring their distances to agree. We refer to the set of selected relation identities as the structural support. We use mKNN to capture local relational structure and introduce H_0 skeleton overlap as its global counterpart. We denote all alignment scores by $S(\cdot, \cdot)$, with subscripts specifying the corresponding measure.

Local relational structure: mKNN. Following Huh et al. (2024), let $N_k^X(i)$ and $N_k^Y(i)$ denote the sets of k nearest neighbors of sample i in representations X and Y, respectively. We measure local relational alignment by the average overlap between the corresponding neighborhoods:

$$S_{\text{mKNN}}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \frac{|N_k^X(i) \cap N_k^Y(i)|}{k}. \tag{1}$$

The score lies in $[0, 1]$, with larger values indicating greater agreement in neighbor identities. Importantly, mKNN depends only on which samples are selected as neighbors and does not require the corresponding neighbor distances to agree.

Global relational structure: H_0 skeleton overlap. To obtain a matched measure of relational structure at the global scale, we use zero-dimensional persistent homology (H_0) computed from the Vietoris-Rips filtration (Edelsbrunner et al., 2002). For each representation, the filtration begins with all samples as separate connected components and progressively adds edges as the distance threshold increases. Whenever an edge connects two disconnected components, one H_0 component disappears; we refer to the corresponding edge as an H_0 death edge. Let $E^X_{H_0}$ and $E^Y_{H_0}$ denote the sets of sample pairs corresponding to these death edges in representations X and Y, respectively.

Proposition 1. *Under a fixed deterministic tie-breaking rule, let T_X and T_Y denote the minimum spanning tree (MST) edge sets of X and Y, respectively. Then the H_0 death-edge set of the Vietoris-Rips filtration coincides with the edge set of the MST; that is, $E^X_{H_0} = T_X$ and $E^Y_{H_0} = T_Y$.*

This correspondence follows from Kruskal's construction (Kruskal, 1956): an edge produces an H_0 death exactly when it connects two previously disconnected components, which is also the criterion for adding an edge to the MST. The connection between persistent homology and minimum spanning structures has also been established more generally (Skraba et al., 2020). Formal definitions and a proof of the proposition are provided in Appendix C.2.

Since each MST contains $n - 1$ edges, we define H_0 skeleton overlap as

$$S_{H_0}(X, Y) = \frac{|E^X_{H_0} \cap E^Y_{H_0}|}{n-1} = \frac{|T_X \cap T_Y|}{n-1}. \tag{2}$$

This score measures agreement in the identities of sample pairs that form the global spanning structure. We use H_0 skeleton overlap as the global counterpart to mKNN, with this local-global distinction formalized in Appendix C.4. Unlike RTD-Lite, which uses MSTs to quantify multiscale connectivity discrepancies between weighted graphs (Tulchinskii et al., 2025), H_0 skeleton overlap retains only the identities of the spanning edges, discarding the corresponding edge lengths. Together, mKNN and H_0 skeleton overlap provide matched measures of relational structure at local and global scales.

3.3 Distance-Aware Alignment

We next probe metric geometry by incorporating distance agreement without changing the relation identities selected at each scale. This allows us to vary the stringency of distance agreement while holding relational structure fixed.

Distance agreement. Given a relation (i, j) shared by the supports of X and Y, we evaluate how closely the two representations agree on its distance. For each representation, pairwise distances are normalized by the 90th percentile of positive distances, $\tilde{D} = D/Q_{0.9}(\{D_{ij} : D_{ij} > 0\})$. We use the 90th percentile as the default reference scale; the resulting patterns are robust to alternative normalization quantiles (Appendix F.1). We then define the distance-agreement weight

$$w^{(\tau)}_{ij} = \exp\left(-\frac{\left|\log \tilde{D}_X(i,j) - \log \tilde{D}_Y(i,j)\right|}{\tau}\right), \tag{3}$$

where $\tau > 0$ controls the stringency of distance agreement. Smaller τ requires closer agreement for a shared relation to receive a large weight, whereas larger τ is more tolerant of distance mismatch; as $\tau \to \infty$, all shared relations receive unit weight, recovering the relational structure measure.

Distance-aware mKNN. For local alignment, we apply Eq. 3 only to neighbors shared by the two representations. We define

$$S^{(\tau)}_{\mathrm{mKNN}}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{j \in N^X_k(i) \cap N^Y_k(i)} w^{(\tau)}_{ij}. \tag{4}$$

The neighborhood sets remain identical to those used in mKNN. Thus, Eq. 4 adds distance agreement while preserving the local structural support.

Distance-aware H_0 skeleton overlap. For global alignment, we apply the same distance-agreement weighting to edges shared by the two H_0 skeletons, equivalently, to edges in $T_X \cap T_Y$, and define

$$S_{H_0}^{(\tau)}(X,Y) = \frac{1}{n-1} \sum_{(i,j)\in T_X \cap T_Y} w_{ij}^{(\tau)}. \tag{5}$$

Controlling distance agreement. The parameter τ controls the strictness of distance agreement: smaller τ more strongly penalizes a given distance mismatch. Conversely,

$$\lim_{\tau\to\infty} S_{\mathrm{mKNN}}^{(\tau)}(X,Y) = S_{\mathrm{mKNN}}(X,Y), \qquad \lim_{\tau\to\infty} S_{H_0}^{(\tau)}(X,Y) = S_{H_0}(X,Y). \tag{6}$$

Since $w_{ij}^{(\tau)} \to 1$ as $\tau \to \infty$, both distance-aware measures recover their corresponding relational structure measures. To examine whether the analysis depends specifically on distance-based comparison, we construct analogous similarity-aware variants using cosine similarity while keeping the same structural supports fixed (Appendix F.2).

3.4 Aggregation-Aware Permutation Calibration

We evaluate all alignment measures using the aggregation-aware permutation calibration of Gröger et al. (2026). For each model pair, sample correspondences are randomly permuted using the same permutation across all layers. We take the maximum scores across layer pairs, ensuring that the null distribution accounts for layer selection and aggregation.

Let $S(X_\ell, Y_{\ell'})$ denote the alignment score between layers ℓ and ℓ'. The observed aggregate alignment and the corresponding aggregate scores under K random permutations $\{\pi_r\}_{r=1}^K$ are

$$T_{\mathrm{obs}} = \max_{\ell,\ell'} S(X_\ell, Y_{\ell'}), \qquad T^{(r)} = \max_{\ell,\ell'} S(X_\ell, \pi_r(Y_{\ell'})). \tag{7}$$

We define c_α as the $(1-\alpha)$ quantile of $\{T_{\mathrm{obs}}, T^{(1)}, \ldots, T^{(K)}\}$. Since all measures in our framework are bounded above by 1, the calibrated score is

$$T_{\mathrm{cal}} = \max\left\{\frac{T_{\mathrm{obs}} - c_\alpha}{1 - c_\alpha}, 0\right\}. \tag{8}$$

Thus, scores that do not exceed the calibration threshold are mapped to zero, while the maximum possible alignment remains one. We apply the same aggregation and calibration procedure to all alignment measures in our framework. Additional details on the number of permutations, permutation significance tests, and multiple-testing correction are provided in Appendix B.

3.5 Beyond Euclidean Geometry

Our primary distance-aware analysis evaluates metric agreement using ambient Euclidean distances. To extend the analysis beyond ambient Euclidean geometry, we construct a Riemannian metric approximation from local covariance information, motivated by local covariance-based geometric constructions (Singer & Coifman, 2008; Berry & Sauer, 2016). We keep the relation identities used in the Euclidean analysis fixed and change only the distances used to evaluate agreement on those relations. This provides a controlled extension in which relational structure remains unchanged while the geometry used to evaluate metric agreement varies. We treat this construction as one alternative geometric model rather than as a uniquely correct intrinsic geometry of the representation space. Construction details are provided in Appendix G.1.

4 Experiments

Our experiments are designed to answer the following key research questions:

- **Relational structure:** Does relational convergence extend from local neighborhoods to global spanning structure? (Section 4.2)
- **Metric geometry:** How does convergence change with stricter distance agreement? (Section 4.3)
- **Beyond Euclidean geometry:** Does the structure-geometry pattern persist under a Riemannian metric approximation? (Section 4.4)
- **Multimodal generalization:** Does the pattern extend to video-text representations? (Section 4.5)

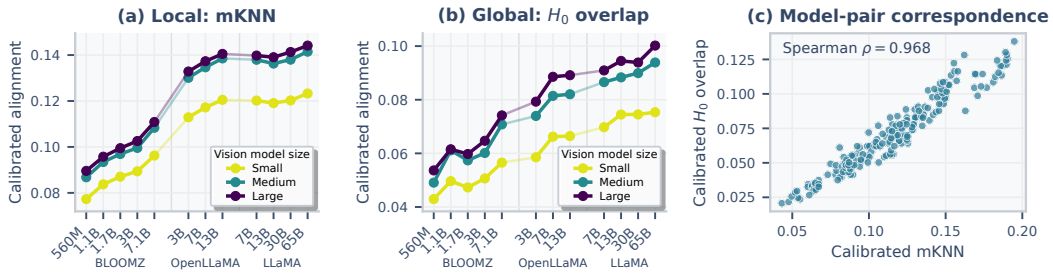


Figure 2: **The relational structure of representations also converges globally.** (a) Following prior work, aggregation-aware calibrated mKNN reproduces the capacity-dependent convergence. (b) Our H_0 skeleton overlap shows that the relational structure of representations also converges at the global scale, extending to global spanning structure. Small, medium, and large denote relative model sizes within each vision-model family, averaged across INet21K, MAE, DINOv2, CLIP, and CLIP fine-tuned on ImageNet-12K. (c) Across all 204 vision-language model pairs, calibrated mKNN and H_0 skeleton overlap are strongly associated (Spearman $\rho = 0.968$).

4.1 Experimental Setup

Vision-language setting. We use the 1,024 paired image-text samples from the WIT subset of the PRH benchmark (Huh et al., 2024; Gröger et al., 2026). Our main evaluation compares 12 language models from the BLOOMZ, OpenLLaMA, and LLaMA families with 17 vision transformers spanning ImageNet-21K supervised models, MAE, DINOv2, and CLIP variants, yielding 204 model pairs. Full model, layer, and representation-extraction details are provided in Appendix A.1.

Alignment and calibration. We use $k = 10$ for mKNN and evaluate the distance-aware measures over $\tau \in \{1, 0.1, 0.01, 0.001, 0.0001\}$. Following Section 3.4, model-pair alignment is obtained by taking the maximum over all layer pairs and calibrated using 500 correspondence permutations at $\alpha = 0.05$. Unless otherwise stated, all reported alignment values are aggregation-aware calibrated scores. We control the false discovery rate across model pairs using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), with full statistical testing details provided in Appendix B.2.

4.2 Relational Structure Converges at Both Local and Global Scales

We first ask whether the convergence in local neighborhood structure reported by prior work extends to global spanning structure after aggregation-aware calibration. All 204 model pairs remain significant for both mKNN and H_0 skeleton overlap after Benjamini-Hochberg correction (Appendix E.1). Full family-wise plots are provided in Appendix D.

Figures 2a and 2b show that the two measures of relational structure exhibit similar scaling behavior. Within language-model families, calibrated alignment generally increases with model capacity, and larger vision models tend to exhibit stronger alignment. Thus, the capacity-dependent pattern observed for local neighborhood relations also extends to global spanning structure.

This correspondence also holds at the level of individual model pairs. Across all 204 vision-language pairs, calibrated mKNN and H_0 skeleton overlap are strongly correlated (Spearman $\rho = 0.968$; Figure 2c). Model pairs with stronger local relational alignment therefore also tend to exhibit stronger global relational alignment. Together, these results indicate that convergence in relational structure is not restricted to local neighborhoods; it extends to global spanning structure even after aggregation-aware calibration.

4.3 Stricter Distance Agreement Weakens Capacity-Dependent Convergence

Having established robust convergence in relational structure at both scales, we next ask how this pattern changes when increasingly stringent distance agreement is required. Starting from the relational structure measures in Section 3.2, we progressively impose distance agreement through the parameter τ defined in Section 3.3. Because decreasing τ assigns smaller weights to a fixed distance mismatch, absolute alignment is expected to decrease. The more substantive question is whether

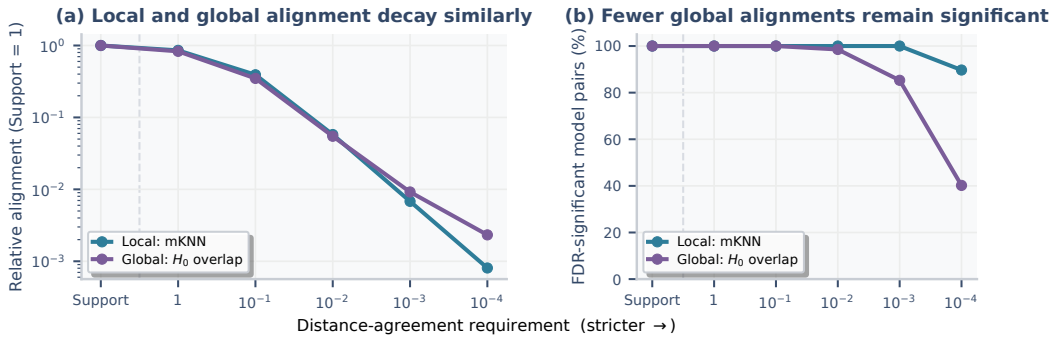


Figure 3: **The metric geometry of representations shows weaker convergence at both local and global scales.** (a) Mean aggregation-aware calibrated alignment across the 204 model pairs, normalized by the corresponding relational structure baseline, which is set to 1. Here, "Support" denotes this baseline without distance weighting. As distance agreement becomes more stringent, alignment weakens at similar rates for both local and global scales. (b) Percentage of the 204 model pairs that remain significant after Benjamini-Hochberg correction at each level of distance agreement. Significance remains similar across scales under weak-to-moderate agreement, while fewer global alignments remain significant in the stringent regime.

the capacity-dependent signature of convergence is preserved as distance agreement becomes more stringent, and whether this behavior differs between local and global scales.

Figure 3a shows that local and global alignment weaken at remarkably similar rates. At $\tau = 1$, the local and global measures retain 85.6% and 83.0% of their relational structure baselines, respectively. These values decrease to 39.2% and 34.9% at $\tau = 10^{-1}$, and to 5.8% and 5.5% at $\tau = 10^{-2}$. Thus, over the weak-to-moderate regime, requiring closer distance agreement produces nearly parallel reductions in alignment at the two structural scales.

More importantly, increasingly stringent distance agreement weakens the capacity-dependent signature observed in relational structure. Across the distance-aware sweep, this pattern becomes less pronounced as the agreement requirement becomes more stringent (Appendix E.1). Exact family-level permutation tests of within-family Spearman associations confirm this weakening for both language and vision capacity at both local and global scales (all Benjamini-Hochberg adjusted $q < 0.05$; Appendix E.3). This indicates that convergence with increasing model capacity is less robust when metric agreement is additionally required.

A scale-dependent difference emerges under the most stringent distance requirements. As shown in Figure 3b, the prevalence of significant local and global alignment remains nearly identical under weak-to-moderate agreement, but diverges as the requirement becomes stringent. All 204 pairs remain significant for local alignment through $\tau = 10^{-3}$, whereas significant global alignment decreases to 98.5% at $\tau = 10^{-2}$, 85.3% at $\tau = 10^{-3}$, and 40.2% at $\tau = 10^{-4}$. At the strictest setting, 89.7% of local alignments remain significant, compared with 40.2% of global alignments.

Together, these results show that increasingly stringent distance agreement weakens convergence at both local and global scales. The dominant contrast is therefore between relational structure and metric geometry, while structural scale becomes more consequential mainly in the stringent regime, where significant global alignment is less prevalent than local alignment. The same qualitative weakening is also observed when agreement is defined using cosine similarity rather than distance (Appendix F.2). Full numerical results and family-wise analyses are provided in Appendix E.

4.4 Extending Beyond the Ambient Euclidean Metric

We next extend the distance-aware analysis beyond the ambient Euclidean metric. Following Section 3.5, we keep the local and global relation identities fixed and evaluate distance agreement using a Riemannian metric approximation constructed from local covariance information. This preserves the underlying relational structure while changing the geometry used to evaluate metric agreement.

As shown in Figure 4a, local and global alignment again decrease at similar rates as τ becomes smaller. At $\tau = 10^{-2}$, only 3.2% and 3.0% of the corresponding relational structure baselines re-

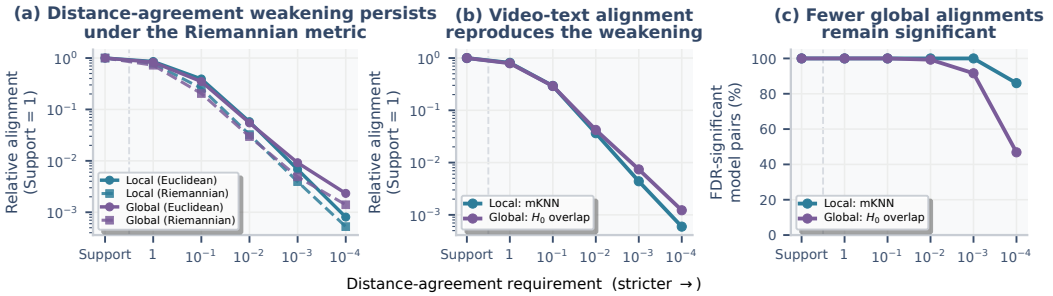


Figure 4: **The structure-geometry pattern extends beyond Euclidean geometry and to video-text representations.** (a) Relative calibrated alignment under the ambient Euclidean metric and a Riemannian metric approximation. The underlying relations are held fixed while only the geometry used to evaluate distance agreement changes. (b) The same weakening under increasingly stringent distance agreement is reproduced in video-text representations at both local and global scales. (c) Percentage of video-text model pairs that remain significant after Benjamini-Hochberg correction at each level of distance agreement.

main for the local and global measures, respectively. More importantly, the full capacity-dependent sweeps show that the increase in alignment with model capacity becomes progressively less pronounced as Riemannian distance agreement becomes more stringent (Appendix G.2).

Thus, the weakening of capacity-dependent convergence under stringent metric agreement is not specific to ambient Euclidean geometry. The same structure-geometry pattern emerges under the Riemannian metric approximation.

4.5 Extension to Video-Text Representations

Finally, we test whether the observed pattern extends to video-text representations. Experimental details are provided in Appendix A.2. Convergence in relational structure is again robust at both scales: all 143 video-text model pairs exhibit significant calibrated alignment for both mKNN and H_0 skeleton overlap. The corresponding family-wise results show similar capacity-dependent patterns across VideoMAE, DINOv2, and CLIP (Appendix H).

Figure 4b shows that increasingly stringent distance agreement again weakens both local and global alignment. The full capacity-dependent sweeps further show that the scaling trend progressively flattens as distance agreement becomes more stringent (Appendix H).

A scale-dependent difference again emerges under stringent distance agreement. At $\tau = 10^{-4}$, 86.0% of local alignments remain significant after Benjamini-Hochberg correction, compared with 46.9% of global alignments (Figure 4c). Thus, the same structure-geometry pattern extends to the video-text setting.

5 Conclusion

Our findings challenge a local neighborhood interpretation of representational convergence. When structural scale is separated from what is compared, relational structure converges at both local and global scales, whereas increasingly stringent metric agreement substantially weakens convergence. This suggests that the key question is not simply whether representations converge locally or globally, but what aspects of their structure are shared across models. These findings also motivate representation alignment methods that consider both local neighborhoods and global spanning structure, rather than focusing on local structure alone.

Our analysis is limited to H_0 spanning structure, a fixed-relation formulation of metric agreement, and a single Riemannian metric approximation. The H_0 skeleton captures only one form of global relational structure, and other global structures may exhibit different convergence patterns. Future work should examine richer topological structures, alternative formulations of metric agreement, and other approximations of intrinsic representation geometry.

AI USE STATEMENT

Generative AI tools were used as assistive tools in several stages of this work. They supported iterative discussion and refinement of the conceptual framing and experimental design, drafting and editing of portions of the manuscript, and the generation, debugging, and visualization of experimental code. They were also used to discuss the interpretation and presentation of experimental results. All final decisions regarding the research questions, methodology, mathematical formulations, experimental analyses, and scientific conclusions were made and critically reviewed by the authors. All AI-assisted code was manually inspected and tested, and all reported results and claims were verified against the underlying experiments. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproduction of all experiments. Sections 3.2-3.5 define the proposed alignment measures, aggregation-aware calibration, and Riemannian robustness analysis, while Section 4.1 summarizes the main experimental protocol. Appendix A provides model, representation-extraction, and preprocessing details; Appendix B specifies permutation calibration, significance testing, and multiple-testing correction; and Appendix C gives the formal H_0 skeleton construction and deterministic tie handling. Appendices D-H report family-wise relational results, full distance-agreement sweeps, normalization and similarity-based robustness analyses, Riemannian experiments, and video-text results. Anonymous supplementary code provides implementations of the proposed measures and calibration procedures, scripts for reproducing the main and robustness experiments, plotting utilities, and the required software dependencies.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. *arXiv preprint arXiv:2201.00058*, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Tyrus Berry and Timothy Sauer. Local kernels and the geometric structure of data. *Applied and Computational Harmonic Analysis*, 40(3):439–469, 2016.

D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. A. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, S. W. Li, P. Dollar, and C. Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

J. H. Cho, A. Madotto, E. Mavroudi, T. Afouras, T. Nagarajan, M. Maaz, Y. Song, T. Ma, S. Hu, S. Jain, M. Martin, H. Wang, H. A. Rasheed, P. Sun, P.-Y. Huang, D. Bolya, N. Ravi, S. Jain, T. Stark, S. Moon, B. Damavandi, V. Lee, A. Westbury, S. Khan, P. Kraehenbuehl, P. Dollar, L. Torresani, K. Grauman, and C. Feichtenhofer. PerceptionLM: Open-access data and models for detailed visual understanding. In *Advances in Neural Information Processing Systems*, 2025.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in neural information processing systems*, 34:1556–1568, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & computational geometry*, 28(4):511–533, 2002.

Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL https://github.com/openlm-research/open_llama.

Phillip Good. *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uz0gAAYydl.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Forty-first International Conference on Machine Learning*, 2024.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9), 2025. doi: 10.1145/3728458.

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. *Frontiers in Systems Neuroscience*, Volume 2 - 2008, 2008. 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.

Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International conference on machine learning*, pp. 7045–7054. PMLR, 2020.

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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pp. 15991–16111, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Robert Clay Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401, 1957.

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. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6076–6085. Curran Associates, Inc., 2017.

Amit Singer and Ronald R Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

Primoz Skraba, Gugan Thoppe, and D Yogeshwaran. Randomly weighted d-complexes: Minimal spanning acycles and persistence diagrams. *The Electronic Journal of Combinatorics*, 27, Apr. 2020. doi: 10.37236/8679. URL https://www.combinatorics.org/ojs/index.php/eljc/article/view/v27i2p11.

Diogo Soares, Pankhil Gawade, Andrea Dittadi, and Ewa Szczurek. Scalable and interpretable representation alignment with ordinal similarity. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=G4D0YzzZEk.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2443–2449, 2021.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=4nPswr1KcP.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Hiq7lUh4Yn. Survey Certification, Expert Certification.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4NJCI4Q3Za.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Eduard Tulchinskii, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov. RTD-lite: Scalable topological analysis for comparing weighted graphs in learning tasks. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=JoDTjOu693.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in neural information processing systems*, 31, 2018.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Junwon You, Kang Dasol, and Jae-Hun Jung. Topological alignment of shared vision-language embedding space. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026a. URL https://openreview.net/forum?id=ecd8cgWZr6.

Junwon You, Mihyun Jang, Sangwoo Mo, and Jae-Hun Jung. Topology-aware representation alignment for semi-supervised vision-language learning. *arXiv preprint arXiv:2604.26370*, 2026b.

Huatian Zhang, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Homology consistency constrained efficient tuning for vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=veMnGKXvTx.

CONTENTS

1 Introduction **1**

2 Related Work **3**

2.1 Representational Similarity Measures 3

2.2 The Platonic Representation Hypothesis and Its Recent Critiques 3

2.3 Topological Comparison and Alignment of Representation Spaces 4

3 A Controlled Framework for Representational Convergence **4**

3.1 Problem Setup and Comparison Framework 4

3.2 Relational Structure Alignment 4

3.3 Distance-Aware Alignment 5

3.4 Aggregation-Aware Permutation Calibration 6

3.5 Beyond Euclidean Geometry 6

4 Experiments **6**

4.1 Experimental Setup 7

4.2 Relational Structure Converges at Both Local and Global Scales 7

4.3 Stricter Distance Agreement Weakens Capacity-Dependent Convergence 7

4.4 Extending Beyond the Ambient Euclidean Metric 8

4.5 Extension to Video-Text Representations 9

5 Conclusion **9**

A Experimental Details **15**

A.1 Vision-Language Models and Representations 15

A.2 Video-Text Experimental Setup 16

B Calibration and Statistical Testing **16**

B.1 Aggregation-Aware Permutation Calibration 16

B.2 Permutation Significance Tests and Multiple-Testing Correction 17

C H_0 Skeleton and Minimum Spanning Trees **18**

C.1 Formal Definition of the H_0 Skeleton 18

C.2 H_0 Death Edges and Minimum Spanning Trees 18

C.3 Tie Handling and Interpretation 19

C.4 Global Nature of the H_0 Skeleton 19

D Additional Relational Structure Results **20**

D.1 Family-Wise Results 20

E Additional Distance-Aware Results **20**

E.1 Full Distance-Agreement Sweep 21

E.2 Family-Wise Distance-Agreement Sweeps 22

E.3 Statistical Validation of the Capacity-Dependent Trend 24

F Robustness to Normalization and Similarity-Based Comparison **26**

F.1 Sensitivity to Distance Normalization 26

F.2 Similarity-Aware Alignment 27

G Robustness under a Locally Estimated Riemannian Metric **28**

G.1 Riemannian Distance Construction 29

G.2 Full Riemannian Results 30

H Additional Video-Text Results **31**

H.1 Family-Wise Relational Structure Results 31

H.2 Full Distance-Agreement Sweep 32

A Experimental Details

A.1 Vision-Language Models and Representations

Our vision-language experiments follow the experimental protocol of Huh et al. (2024), as adopted by Gröger et al. (2026). We use 1,024 paired image-text samples from the WIT (Wikipedia-based Image Text) dataset (Srinivasan et al., 2021). We evaluate language and vision models at multiple scales to examine representational alignment across model capacity.

For language models, we consider 12 models spanning three model families: BLOOMZ (Muennighoff et al., 2023), OpenLLaMA (Geng & Liu, 2023), and LLaMA (Touvron et al., 2023). For vision models, we consider 17 Vision Transformers spanning ImageNet-21K supervised models (Dosovitskiy et al., 2021; Steiner et al., 2022), MAE (He et al., 2022), DINOv2 (Oquab et al., 2024), and CLIP variants (Radford et al., 2021; Cherti et al., 2023). The resulting combination yields 204 vision-language model pairs.

For each language model, we extract representations from all available hidden states and mean-pool over non-padding tokens. For each vision model, we extract representations from every transformer block using the CLS token. Thus, each model is represented by a sequence of layer-wise feature vectors over the same 1,024 samples.

Before computing alignment measures, we apply the same feature preprocessing procedure to all representations. Specifically, for each model representation, we compute the 95th percentile of the absolute feature values for each sample and average these sample-wise quantiles to obtain a single clipping threshold. Feature values are then clipped symmetrically to this threshold, and the resulting layer-wise representations are L2-normalized before constructing neighborhood and spanning supports.

Because the layer at which cross-model alignment occurs is not known a priori, we compute the alignment score for every pair of layers between a vision model and a language model and take the maximum across layer pairs as the model-pair alignment. This aggregation procedure is applied consistently across all alignment measures.

We evaluate alignment in relational structure at both local and global scales. For local relational structure, we use mKNN with $k = 10$, while for global relational structure, we use H_0 skeleton overlap. We then evaluate distance-aware variants of both measures by imposing increasingly stringent distance agreement through $\tau \in \{1, 0.1, 0.01, 0.001, 0.0001\}$.

For the distance-aware analysis, the structural supports are first constructed from the normalized representations and then held fixed while distance agreement is evaluated on the corresponding relations. Our primary analysis uses the ambient Euclidean metric. We additionally repeat the analysis using a locally estimated Riemannian metric while keeping the underlying structural supports fixed, testing whether the observed structure-geometry pattern depends on the choice of metric.

A.2 Video-Text Experimental Setup

We additionally test whether the observed alignment patterns extend beyond vision-language representations by repeating the same analysis in a video-text setting. We use 1,024 video samples from the test split of PVD (Bolya et al., 2025; Cho et al., 2025).

For video-native representations, we use VideoMAE (Tong et al., 2022). We evaluate a scale series consisting of the Base and Large pretrained checkpoints and the Huge checkpoint fine-tuned on Kinetics. This provides video representations spanning multiple model scales and fine-tuning conditions.

As a frame-level baseline, we apply image-based vision models to the middle frame of each video. Specifically, we use DINOv2 (small, base, large, and giant) and CLIP (base, large, huge, and giant) to extract representations from the selected middle frame. These models provide a comparison between video-native representations and representations obtained from a single image frame.

VideoMAE representations are extracted from 16-frame clips using the CLS token at each hidden layer. For sufficiently long videos, we use up to four 16-frame clips obtained from uniformly sampled frames and average the resulting layer-wise representations across clips. DINOv2 and CLIP operate on a single middle frame, with CLS representations extracted from every transformer block.

The language models include the same 12 models used in the vision-language experiments together with Gemma-2-9B-IT (Team et al., 2024), spanning the BLOOMZ, OpenLLaMA, LLaMA, and Gemma families. Language representations are extracted using the same procedure described in Appendix A.1. The resulting combinations yield 143 video-text model pairs.

We evaluate video-text alignment in relational structure at both local and global scales using mKNN and H_0 skeleton overlap, respectively. We further evaluate distance-aware variants of both measures under increasingly stringent distance agreement, with $\tau \in \{1, 0.1, 0.01, 0.001, 0.0001\}$. As in the vision-language experiments, alignment is obtained by taking the maximum across all layer pairs. The same preprocessing, structural-support construction, distance-aware evaluation, and permutation-based calibration procedures are applied to the video-text experiments.

B Calibration and Statistical Testing

B.1 Aggregation-Aware Permutation Calibration

We use permutation-based null calibration (Gröger et al., 2026) to account for alignment that may arise by chance under finite-sample and high-dimensional settings, as well as inflation introduced by layer-wise aggregation. For each pair of models, we first compute the alignment score for every pair of layers and construct the corresponding layer-pair alignment matrix. The observed aggregate alignment is then obtained using the same aggregation operator employed throughout our experiments, namely the maximum over all layer pairs.

Formally, let X_ℓ and $Y_{\ell'}$ denote the representations of layers ℓ and ℓ' from two models. For an alignment measure S, we construct the observed alignment matrix

$$\mathbf{S}_{\text{obs}} = [S(X_\ell, Y_{\ell'})]_{\ell,\ell'} . \tag{9}$$

The observed aggregate statistic is

$$T_{\text{obs}} = \max_{\ell,\ell'} S(X_\ell, Y_{\ell'}). \tag{10}$$

To construct an empirical null distribution, we randomly permute the sample correspondences between the two models. For each permutation π_r, we apply the same permutation to every layer of

the second model and recompute the complete layer-pair alignment matrix:

$$\mathbf{S}^{(r)} = [S(X_\ell, \pi_r(Y_{\ell'}))]_{\ell,\ell'}\,. \tag{11}$$

We then apply the same aggregation operator to obtain the permuted aggregate statistic

$$T^{(r)} = \max_{\ell,\ell'} S(X_\ell, \pi_r(Y_{\ell'})), \qquad r = 1, \ldots, K. \tag{12}$$

Applying a common permutation across all layers preserves sample correspondence across layers within each model while disrupting cross-model correspondence. Moreover, taking the maximum after permutation reproduces the same layer-selection and aggregation procedure used for the observed statistic. The resulting null distribution therefore accounts for inflation induced by selecting the maximum across multiple layer pairs.

We use $K = 500$ random permutations for each model pair and set the significance level to $\alpha = 0.05$ throughout the experiments. Let $\mathcal{T} = \left\{T_{\text{obs}}, T^{(1)}, \ldots, T^{(K)}\right\}$ denote the combined set of observed and permuted aggregate statistics. We sort these values in ascending order as $T_{(1)} \leq \cdots \leq T_{(K+1)}$. The permutation critical value is defined as the empirical right-tail $(1-\alpha)$ quantile:

$$c_\alpha = T_{(\lceil (1-\alpha)(K+1) \rceil)}. \tag{13}$$

This value provides the reference level for calibrating observed alignment relative to the permutation null.

For alignment measures bounded above by one, we transform the observed aggregate statistic into a calibrated score as

$$T_{\text{cal}} = \max\left\{\frac{T_{\text{obs}} - c_\alpha}{1 - c_\alpha}, 0\right\}. \tag{14}$$

This transformation maps scores at or below the calibration threshold to zero while preserving an upper bound of one.

The same permutation scheme, layer-wise aggregation, critical-value construction, and calibration transformation are applied to all alignment measures considered in our experiments. This provides a common calibration procedure across relational structure and distance-aware analyses.

B.2 Permutation Significance Tests and Multiple-Testing Correction

In addition to calibrated alignment scores, we perform permutation-based significance tests to determine whether the observed aggregate alignment exceeds that expected under disrupted cross-model sample correspondence. Using the same permutation aggregate statistics, we compute the one-sided empirical permutation p-value

$$p = \frac{1 + \sum_{r=1}^{K} \mathbb{I}\left[T^{(r)} \geq T_{\text{obs}}\right]}{K+1}, \tag{15}$$

where $\mathbb{I}[\cdot]$ is the indicator function. The add-one correction yields a valid finite-sample permutation p-value and prevents the reported value from being exactly zero.

At the nominal level, a model pair satisfies $p \leq 0.05$ when its observed aggregate alignment is unusually large relative to the permutation distribution. Because multiple model pairs are tested, however, our reported significance decisions are based on false discovery rate control rather than the uncorrected threshold.

We use the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). Suppose that m model-pair comparisons yield permutation p-values $p_1, \ldots, p_m$, sorted as

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

For a target FDR level q, we identify the largest index j satisfying

$$p_{(j)} \leq \frac{j}{m} q. \tag{16}$$

All hypotheses corresponding to $p_{(1)}, \ldots, p_{(j)}$ are then declared significant. We set $q = 0.05$ and apply the Benjamini-Hochberg correction separately across model-pair tests for each alignment measure and each τ setting. The resulting FDR-controlled decisions are used to report statistically significant alignment, while calibrated scores quantify alignment above the permutation-based calibration threshold.

C H_0 SKELETON AND MINIMUM SPANNING TREES

C.1 FORMAL DEFINITION OF THE H_0 SKELETON

Let

$$X = \{x_1, \dots, x_n\}$$

be a representation of n labeled samples, and let D_X denote its pairwise distance matrix.

Consider the complete weighted graph

$$G_X = (V, E, w_X), \qquad V = \{1, \dots, n\}, \qquad w_X(i,j) = D_X(i,j),$$

where each vertex index identifies the same sample across representations. Let

$$e_1 \prec_X e_2 \prec_X \cdots \prec_X e_m, \qquad m = \binom{n}{2},$$

denote a fixed total ordering of the edges that is nondecreasing in w_X. Ties are resolved deterministically as described in Appendix C.3.

For $r = 0, \dots, m$, let

$$G_X^{(r)} = \big(V, \{e_1, \dots, e_r\}\big),$$

with $G_X^{(0)}$ containing only the n isolated vertices. This ordered graph filtration is a tie-refined 1-skeleton filtration of the Vietoris-Rips filtration. An edge $e_r = (i,j)$ produces an H_0 death event precisely when its endpoints belong to distinct connected components of $G_X^{(r-1)}$. We therefore define the labeled H_0 death-edge set as

$$E_{H_0}^X = \left\{ e_r = (i,j) \in E : i \text{ and } j \text{ belong to distinct connected components of } G_X^{(r-1)} \right\}.$$

The corresponding H_0 skeleton is the labeled graph

$$\mathcal{H}_0(X) = \big(V, E_{H_0}^X\big).$$

Each edge in $E_{H_0}^X$ records the sample pair responsible for merging two previously disconnected components. Thus, the H_0 skeleton retains the identities of the relations that form the spanning connectivity structure while discarding their filtration values. In Section 3.2, we compare these labeled edge identities across representations to measure global relational alignment.

C.2 H_0 DEATH EDGES AND MINIMUM SPANNING TREES

We now formalize the correspondence between H_0 death edges and minimum spanning trees. Let $\prec_X$ be any total ordering of the edges that is consistent with their weights, with ties resolved by a fixed deterministic rule. Let T_X denote the edge set returned by Kruskal's algorithm (Kruskal, 1956) under this ordering.

Proposition 1. Under the same total ordering $\prec_X$, the labeled H_0 death-edge set coincides with the minimum spanning tree edge set:

$$E_{H_0}^X = T_X.$$

Proof of Proposition 1. Let $F_X^{(r)}$ denote the forest produced by Kruskal's algorithm after processing the first r edges. We first observe that $F_X^{(r)}$ and $G_X^{(r)}$ have the same connected components for every r. Initially, both contain n isolated vertices. When an edge $e_r = (i,j)$ is processed, either its endpoints are already connected, in which case neither graph changes its connected components, or they lie in distinct components, in which case $G_X^{(r)}$ merges those components and Kruskal's algorithm accepts e_r, producing the same merge.

Therefore,

$$e_r \in E_{H_0}^X \iff i \text{ and } j \text{ are disconnected in } G_X^{(r-1)} \iff e_r \in T_X.$$

Hence $E_{H_0}^X = T_X$. Since the complete weighted graph is connected, the resulting tree contains exactly $n-1$ edges:

$$|E_{H_0}^X| = |T_X| = n - 1.$$

The same argument applies to Y. □

C.3 Tie Handling and Interpretation

When tied edge weights admit multiple valid minimum spanning trees (MSTs), our implementation selects a deterministic representative using a dense implementation of Prim's algorithm (Prim, 1957). The algorithm is initialized at the first sample; among candidates with the same current minimum weight, the lowest-index sample is selected, and an existing parent is updated only when a strictly smaller edge weight is encountered. Since sample indices are fixed across representations, this procedure yields a deterministic labeled MST for each representation.

This implementation is consistent with the correspondence established in Appendix C.2. The deterministic MST selected by Prim's algorithm can be represented as the MST obtained by Kruskal's algorithm under an appropriate deterministic refinement of tied edge weights. Under that refinement, the selected MST therefore coincides with the corresponding labeled H_0 death-edge set. When all pairwise weights are distinct, the MST is unique and no tie refinement is required.

Although individual MST edges may connect nearby samples, the edge set is globally constrained by the requirement of spanning all samples without cycles. The resulting $n-1$ labeled edges therefore encode a connectivity structure over the entire representation. We interpret this edge set as global spanning relational structure, without implying that individual edges are geometrically long-range or that the H_0 skeleton captures the full global geometry.

C.4 Global Nature of the H_0 Skeleton

The distinction between local and global structure in our framework can be understood through connectivity over the point cloud. A k-nearest-neighbor graph is constructed from sample-centered neighborhoods, so its relations are determined from the k nearest samples around individual points. For a fixed k, these local relations need not connect the entire point cloud. For example, when the samples form two well-separated dense clusters, all k nearest neighbors of each sample may lie within the same cluster, leaving the two clusters disconnected.

The H_0 skeleton has a fundamentally different property. By Proposition 1, its edge set coincides with that of an MST. An MST necessarily connects all samples into a single connected structure. Consequently, every pair of samples in the point cloud is connected by a path in the H_0 skeleton, regardless of whether the samples lie in widely separated regions. In the two-cluster example above, the MST must contain an edge crossing the gap between the clusters even when no such relation appears in a fixed-k nearest-neighbor graph. Thus, local neighborhood relations may remain confined within disconnected regions of a representation, whereas the H_0 skeleton necessarily connects these regions into a structure spanning the entire representation.

The following proposition makes this distinction explicit.

Proposition 2. ***(Global connectivity beyond fixed local neighborhoods).*** *Let $k \geq 1$ be fixed, and for every finite point set Z with $|Z| \geq k+1$, let $G_k(Z)$ denote its symmetrized k-nearest-neighbor graph. Then there exists a finite point set $X \subset \mathbb{R}$ such that $G_k(X)$ has exactly two connected components, whereas the H_0 skeleton $H_0(X)$ is connected. Moreover, $H_0(X)$ contains an edge joining the two connected components of $G_k(X)$ that is not contained in $G_k(X)$.*

Proof. Fix $k \geq 1$ and consider

$$A = \{0, 1, \ldots, k\}, \qquad B = \{3k+2, 3k+3, \ldots, 4k+2\},$$

and let $X = A \cup B$.

Each of A and B contains exactly $k+1$ points. Their within-cluster diameters satisfy

$$\operatorname{diam}(A) = \operatorname{diam}(B) = k,$$

whereas the minimum distance between the two clusters is

$$d(A, B) = \min_{a \in A,\, b \in B} |a - b| = (3k+2) - k = 2k+2 > k.$$

Therefore, for every point in either cluster, its k nearest neighbors are precisely the other k points in the same cluster. Hence the subgraphs induced by A and B are both complete, and there is no edge between them. Thus,

$$G_k(X) \cong K_{k+1} \sqcup K_{k+1},$$

where K_{k+1} denotes the complete graph on $k+1$ vertices and $\sqcup$ denotes the disjoint union. In particular, $G_k(X)$ has exactly two connected components, namely A and B.

Now consider the cut (A, B) of the complete weighted graph on X, with Euclidean distance as the edge weight. The unique minimum-weight edge crossing this cut is

$$e^* = \{k,\, 3k+2\}.$$

Since e^* is the unique minimum-weight edge crossing the cut (A, B), the cut property of MSTs implies that e^* belongs to every MST of X (Kruskal, 1956). By Proposition 1, the labeled H_0 death-edge set coincides with the edge set of the MST selected under the corresponding deterministic ordering. Hence $e^* \in E^X_{H_0}$. Since e^* joins A and B, while $G_k(X)$ contains no edge between these two components, $e^* \notin E(G_k(X))$. Therefore, the fixed-k neighborhood graph remains disconnected, whereas the H_0 skeleton necessarily contains an edge connecting its two components and is connected over the entire point set. □

This example captures the sense in which H_0 skeleton overlap provides a global counterpart to mKNN. mKNN compares relations defined within sample-centered local neighborhoods, while the H_0 skeleton compares relations that collectively connect the entire sample set. More generally, under Kruskal's construction, whether an edge belongs to the H_0 skeleton depends on the connectivity induced by the lower-weight edges processed before it: an edge is selected precisely when it joins two previously disconnected components. Thus, the relational support of the H_0 skeleton is determined at the level of the full point cloud rather than independently around individual samples.

Accordingly, we use mKNN to measure local relational structure and H_0 skeleton overlap to measure global relational structure. Here, "global" refers to connectivity across the entire sample set: individual MST edges need not themselves be geometrically long-range, but together they form a connected structure spanning all samples.

D Additional Relational Structure Results

We provide family-wise analyses complementing the aggregate relational structure results in Section 4.2. Figure 2 groups vision models by their relative size within each family and averages the corresponding small, medium, and large models across vision-model families. Here, we disaggregate these averages and report the individual variants within each family. This analysis tests whether the capacity-dependent relational alignment observed in the main results is broadly shared across model families rather than being driven by a particular family.

D.1 Family-Wise Results

Figure 5 reports local mKNN and global H_0 skeleton overlap separately for the ImageNet-21K supervised, MAE, DINOv2, CLIP, and ImageNet-12K fine-tuned CLIP families. Individual vision-model variants are shown within each family, together with both aggregation-aware calibrated and uncalibrated alignment.

Across families, the local and global measures of relational structure exhibit broadly similar capacity-dependent patterns. Alignment generally increases with language-model capacity, and larger vision-model variants tend to exhibit stronger alignment within each family. Aggregation-aware calibration reduces the absolute alignment scores but largely preserves these qualitative relationships. Importantly, the same capacity-dependent pattern is visible in global H_0 spanning structure as in local neighborhood relations. The aggregate convergence in relational structure reported in the main text is therefore not attributable to a particular vision-model family, but is broadly reproduced within the constituent vision-model families.

E Additional Distance-Aware Results

We provide additional results complementing the aggregate distance-aware analysis in Section 4.3. We first present the full capacity-dependent distance-agreement sweep shown compactly in Figure 1b. We then disaggregate the analysis by vision-model family to examine whether the aggregate pattern arises from averaging across heterogeneous representation families. Together, these analyses

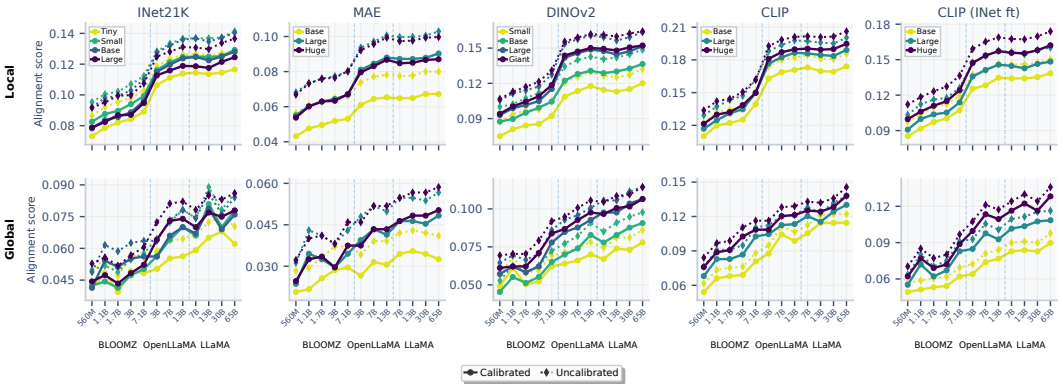


Figure 5: **Family-wise relational structure alignment in vision-language representations.** Top: local relational structure measured by mKNN. Bottom: global spanning structure measured by H_0 skeleton overlap. Columns correspond to the ImageNet-21K supervised, MAE, DINOv2, CLIP, and ImageNet-12K fine-tuned CLIP families, with colors distinguishing individual vision-model variants within each family. Solid lines with circles show aggregation-aware calibrated alignment, and dotted lines with diamonds show uncalibrated alignment. Text-model parameter counts are shown along the horizontal axis. Each panel uses an independently scaled vertical axis to emphasize the within-family capacity-dependent pattern.

show how the capacity-dependent signature of convergence changes as distance agreement becomes increasingly stringent and how consistently this pattern appears across model families.

E.1 Full Distance-Agreement Sweep

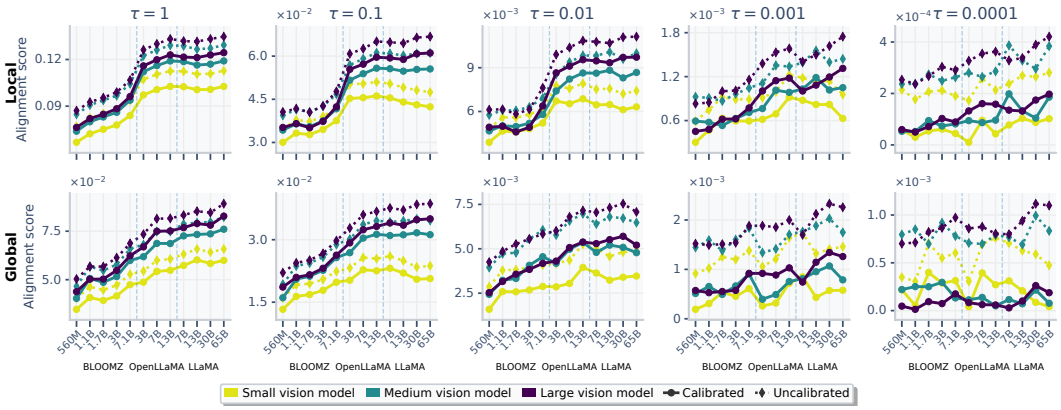


Figure 6: **Full capacity-dependent distance-agreement sweep in vision-language representations.** Top: local distance-aware mKNN. Bottom: global distance-aware H_0 skeleton overlap. Columns progressively strengthen the distance-agreement requirement from $\tau = 1$ to 10^{-4}. Small, medium, and large denote relative model sizes within each vision-model family, averaged across vision-model families. Solid lines with circles show aggregation-aware calibrated alignment, and dotted lines with diamonds show uncalibrated alignment. Each panel uses an independently scaled vertical axis to visualize the within-τ capacity pattern; the cross-τ decrease in alignment magnitude and statistical significance is quantified in Table 1.

Figure 6 presents an enlarged view of the full distance-agreement sweep. As in the relational structure analysis, small, medium, and large denote relative model sizes within each vision-model family, with corresponding size categories averaged across families. Both aggregation-aware calibrated and uncalibrated alignment are shown.

Under relatively tolerant distance agreement, the capacity-dependent pattern observed when comparing relational structure remains clearly visible. At $\tau = 1$ and, to a lesser extent, $\tau = 10^{-1}$,

alignment generally increases with language-model capacity, while larger vision models tend to exhibit stronger alignment for both local distance-aware mKNN and global distance-aware H_0 skeleton overlap. As τ decreases further, however, calibrated alignment weakens substantially and the separation associated with model capacity becomes progressively less stable and less pronounced. Under the most stringent distance-agreement settings, the calibrated trajectories are small in magnitude and no longer exhibit the clear capacity-dependent ordering observed when only relational structure is compared.

Importantly, this transition occurs at both structural scales. The local and global measures differ in their detailed trajectories, particularly under the most stringent settings, but neither retains the robust capacity-dependent pattern observed for relational structure. Thus, the full capacity curves reinforce the aggregate result in Section 4.3: increasingly stringent distance agreement weakens the capacity-dependent signature of representational convergence at both local and global scales.

Table 1: **Aggregate distance-agreement results in vision-language representations.** Relative alignment is the mean aggregation-aware calibrated alignment across 204 model pairs, normalized by the corresponding relational structure mean. Significant pairs are determined after Benjamini-Hochberg correction.

	Relative calibrated alignment		FDR-significant pairs	
Setting	mKNN	H_0	mKNN	H_0
Support	1.0000	1.0000	204/204 (100.0%)	204/204 (100.0%)
$\tau = 1$	0.8560	0.8295	204/204 (100.0%)	204/204 (100.0%)
$\tau = 10^{-1}$	0.3918	0.3487	204/204 (100.0%)	204/204 (100.0%)
$\tau = 10^{-2}$	0.0578	0.0547	204/204 (100.0%)	201/204 (98.5%)
$\tau = 10^{-3}$	0.00681	0.00919	204/204 (100.0%)	174/204 (85.3%)
$\tau = 10^{-4}$	0.000807	0.002327	183/204 (89.7%)	82/204 (40.2%)

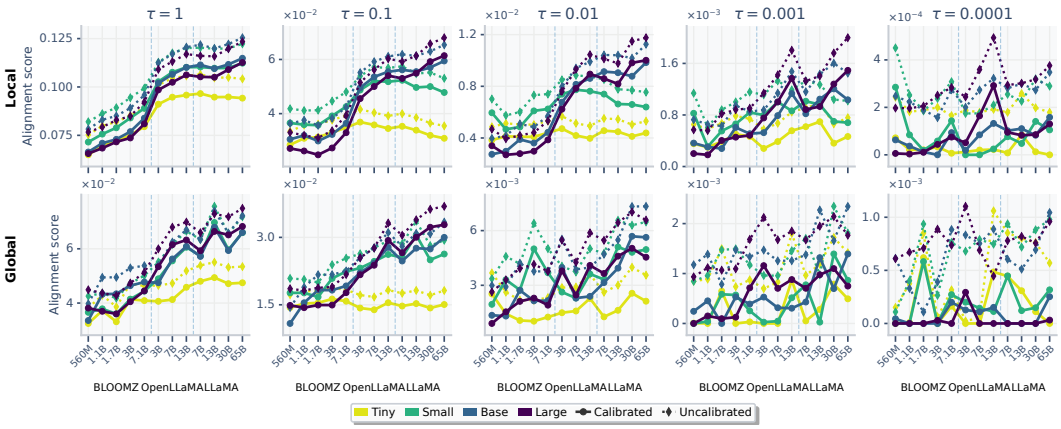
Table 1 quantifies the cross-τ decay that is not directly comparable from Figure 6. Relative calibrated alignment decreases by more than sixfold between $\tau = 10^{-1}$ and 10^{-2} at both structural scales and continues to decline rapidly thereafter. In contrast, a clear local-global difference emerges primarily in statistical significance under stringent agreement: at $\tau = 10^{-4}$, 89.7% of local model pairs remain significant, compared with 40.2% of global pairs. Together, these results show that stricter distance agreement produces the dominant weakening at both scales, while structural scale becomes more consequential mainly in the stringent regime.

E.2 Family-Wise Distance-Agreement Sweeps

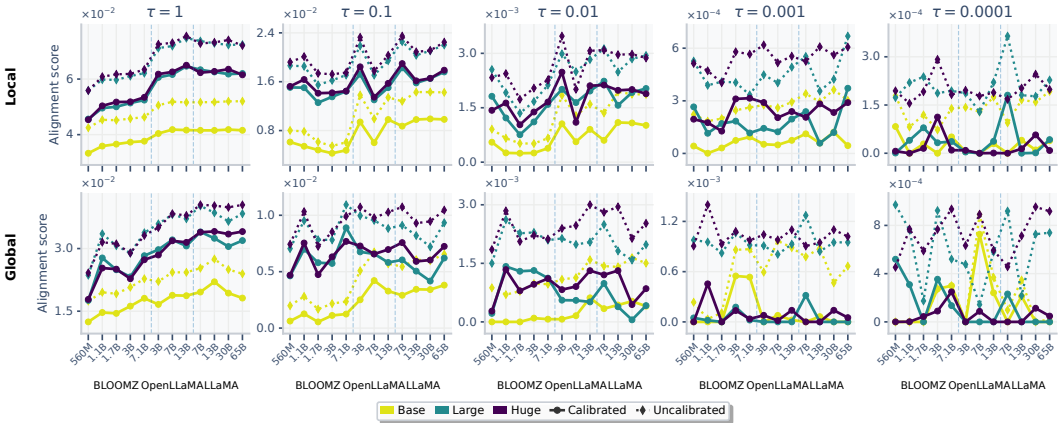
We next disaggregate the distance-agreement sweep by vision-model family. Figure 7 shows the complete τ sweep separately for ImageNet-21K supervised models, MAE, DINOv2, CLIP, and ImageNet-12K fine-tuned CLIP. Unlike Figure 6, which averages models of comparable relative size across families, each subfigure retains the individual vision-model variants within a single family.

The family-wise results broadly reproduce the aggregate transition. At relatively tolerant values of τ, most families retain a recognizable capacity-dependent pattern at both local and global scales, with larger vision-model variants generally exhibiting stronger alignment. As distance agreement becomes more stringent, calibrated alignment decreases substantially and the within-family capacity ordering becomes progressively less pronounced or less stable. The precise transition varies across families: some families retain a clearer capacity-dependent pattern at intermediate values of τ, whereas others become irregular earlier. Under the most stringent settings, however, the calibrated trajectories are weak and substantially less structured across all families.

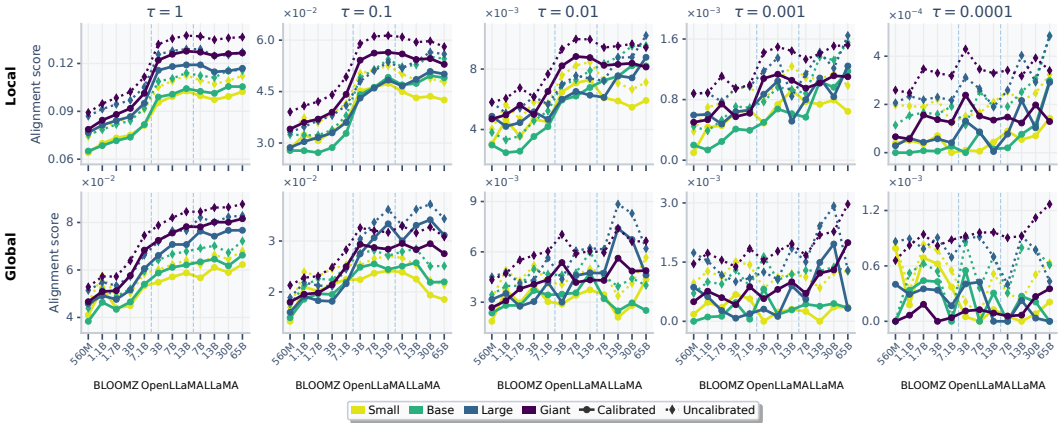
These family-specific differences do not alter the main qualitative result. Both local and global distance-aware alignment weaken as distance agreement becomes more stringent within the different representation families considered here. The aggregate weakening in Section 4.3 therefore cannot be explained simply by averaging across heterogeneous vision-model families. Instead, the loss of robust capacity-dependent alignment under increasingly stringent distance agreement is broadly reproduced within the constituent families themselves.



(a) ImageNet-21K supervised models.

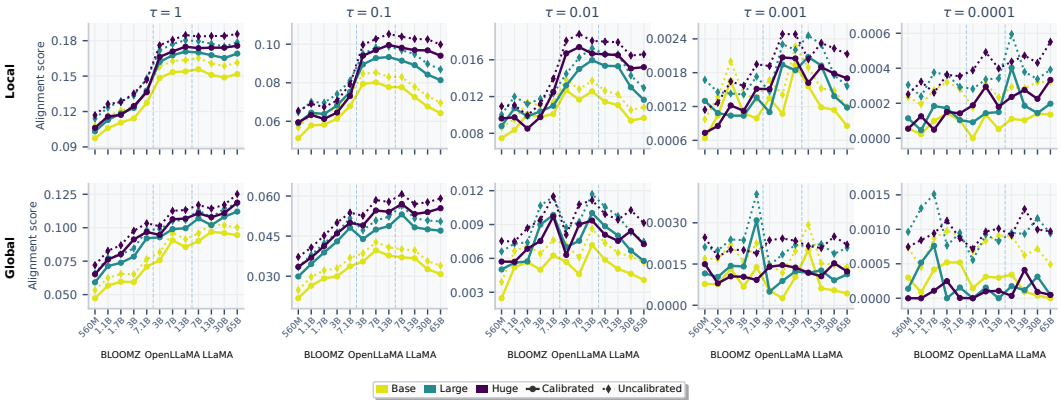


(b) MAE models.

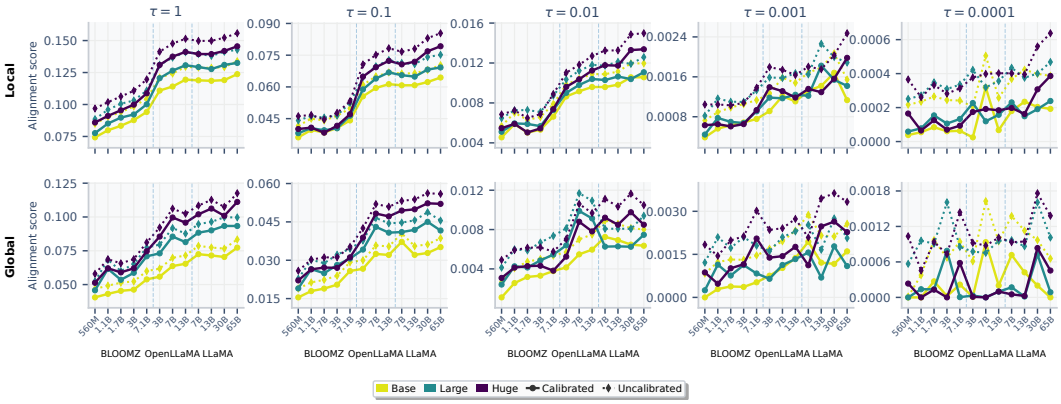


(c) DINOv2 models.

Figure 7: **Family-wise distance-agreement sweeps in vision-language representations.** Top and bottom rows within each subfigure show local distance-aware mKNN and global distance-aware H_0 skeleton overlap, respectively. Columns vary the distance-agreement requirement from $\tau = 1$ to 10^{-4}. Solid lines with circles show aggregation-aware calibrated alignment, while dotted lines with diamonds show uncalibrated alignment. Each panel uses an independently scaled vertical axis.



(d) CLIP models.



(e) ImageNet-12K fine-tuned CLIP models.

Figure 7: **Family-wise distance-agreement sweeps (continued).**

E.3 Statistical Validation of the Capacity-Dependent Trend

The full distance-agreement sweeps in Appendix E.1 show that the capacity-dependent alignment pattern becomes less pronounced as the agreement requirement becomes more stringent. Because the absolute magnitude of the distance-aware score is expected to decrease as τ decreases, we additionally test whether the association between model capacity and alignment itself becomes weaker. We quantify this association using Spearman's rank correlation (Spearman, 1904), which depends only on the ordering of the alignment values and is therefore invariant to a uniform rescaling of scores at a given τ.

Within-family capacity association. We evaluate capacity dependence separately along the language- and vision-model axes. For language capacity, we hold the vision model fixed and compute Spearman's rank correlation between model capacity and alignment within each language-model family (BLOOMZ, OpenLLaMA, and LLaMA). This yields 51 within-family comparisons from 17 fixed vision models and three language-model families. For vision capacity, we hold the language model fixed and compute Spearman's rank correlation between within-family vision-model size and alignment separately for ImageNet-21K supervised models, MAE, DINOv2, CLIP, and ImageNet-12K fine-tuned CLIP. This yields 60 within-family comparisons from 12 fixed language models and five vision-model families. For each structural scale and distance-agreement setting, we summarize the capacity association by the mean Spearman correlation across valid within-family comparisons.

Because the same models occur across multiple fixed-model comparisons, we do not treat these comparisons as independent observations. Instead, we use an exact permutation test that preserves this model-reuse structure by permuting capacity labels at the model-family level (Good, 2005). For

language capacity, capacity labels are permuted independently within BLOOMZ, OpenLLaMA, and LLaMA, and each resulting assignment is applied jointly across all fixed vision models and all values of τ. This gives $5! \times 3! \times 4! = 17{,}280$ possible assignments. For vision capacity, the within-family size labels are permuted jointly across all fixed language models, giving $4! \times 3! \times 4! \times 3! \times 3! = 124{,}416$ possible assignments. We enumerate all assignments in both analyses.

We use two complementary one-sided statistics. First, the endpoint statistic measures the decrease in mean capacity association between the most tolerant and most stringent distance-aware settings,

$$T_{\text{end}} = \bar{\rho}_{\tau=1} - \bar{\rho}_{\tau=10^{-4}}.$$

Second, we test whether capacity association becomes weaker across the complete distance-agreement sweep. Because $\tau \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ is evenly spaced on the log scale, we regress the mean Spearman correlation on the corresponding stringency index $0, 1, 2, 3, 4$ and use the negative slope as the test statistic. Exact permutation p-values are obtained from the complete family-level label-permutation distribution. Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) is applied across the four language/vision $\times$ local/global comparisons separately for the endpoint and full-sweep tests.

Table 2: Statistical validation of the capacity-dependent trend for aggregation-aware calibrated alignment. Entries under Support and τ denote the mean within-family Spearman correlation between model capacity and alignment. Support is shown for reference and is not included in the endpoint or full-sweep tests. The endpoint decrease is defined as $\Delta\rho = \bar{\rho}_{\tau=1} - \bar{\rho}_{\tau=10^{-4}}$. Exact p-values are obtained from family-level capacity-label permutations that preserve model reuse across within-family comparisons. Benjamini-Hochberg adjusted q-values are computed across the four language/vision $\times$ local/global comparisons separately for the endpoint and full-sweep tests.

Capacity axis	Scale	Support	$\tau = 1$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	$\Delta\rho$
Language	Local	.913	.735	.376	.269	.351	.293	.442
Language	Global	.770	.645	.388	.353	.319	.128	.517
Vision	Local	.755	.775	.707	.623	.478	.363	.412
Vision	Global	.889	.850	.722	.590	.414	-.174	1.024

Capacity axis	Scale	Endpoint p	Endpoint q	Full-sweep p	Full-sweep q
Language	Local	.0156	.0156	.0369	.0369
Language	Global	.0150	.0156	.0288	.0369
Vision	Local	.00962	.0156	.00593	.0119
Vision	Global	1.61×10^{-5}	6.43×10^{-5}	1.61×10^{-5}	6.43×10^{-5}

Capacity trends after aggregation-aware calibration. Table 2 shows the within-family capacity associations for aggregation-aware calibrated alignment. For the support-only measures, capacity shows a strong positive association with alignment along both model axes. The mean Spearman correlations are 0.913 locally and 0.770 globally for language capacity, and 0.755 and 0.889, respectively, for vision capacity.

Within the distance-aware sweep, these positive capacity associations become substantially weaker as the agreement requirement becomes more stringent. For language capacity, the mean local correlation decreases from 0.735 at $\tau = 1$ to 0.293 at $\tau = 10^{-4}$. The intermediate trajectory is not strictly decreasing at every value of τ, but the overall decrease is significant under the endpoint test ($\Delta\rho = 0.442$, exact $p = 0.0156$, BH-adjusted $q = 0.0156$) and is also significant across the full sweep ($p = 0.0369$, $q = 0.0369$). For global alignment, the mean language-capacity correlation decreases from 0.645 to 0.128. The endpoint decrease is again significant ($\Delta\rho = 0.517$, $p = 0.0150$, $q = 0.0156$), as is the weakening across the complete sweep ($p = 0.0288$, $q = 0.0369$).

The same pattern appears along the vision-capacity axis. Mean local correlation decreases from 0.775 at $\tau = 1$ to 0.363 at $\tau = 10^{-4}$ ($\Delta\rho = 0.412$, exact $p = 0.00962$, $q = 0.0156$), with a significant decrease across the full sweep ($p = 0.00593$, $q = 0.0119$). For global alignment, the mean correlation decreases from 0.850 to -0.174, giving the largest endpoint decrease among the four comparisons ($\Delta\rho = 1.024$, $p = 1.61 \times 10^{-5}$, $q = 6.43 \times 10^{-5}$). The full-sweep test is also significant ($p = 1.61 \times 10^{-5}$, $q = 6.43 \times 10^{-5}$).

For the vision-capacity/global comparison, the negative mean correlation at the strictest setting should not be interpreted as evidence for a general inverse scaling relationship. Rather, it indicates that the strong positive capacity ordering observed under the support-only measure and tolerant distance agreement is no longer preserved under this stringent agreement requirement.

These results separate the weakening of the capacity-dependent pattern from the mechanically expected decrease in alignment magnitude. Spearman correlation depends only on the ordering of alignment values. Therefore, if decreasing τ merely multiplied all alignment scores by a smaller common factor, the capacity association would remain unchanged. Instead, the rank association between model capacity and alignment itself becomes substantially weaker at both structural scales and along both model-capacity axes.

F Robustness to Normalization and Similarity-Based Comparison

F.1 Sensitivity to Distance Normalization

Our main distance-aware analysis normalizes each pairwise distance matrix by the 90th percentile of its positive distances before evaluating distance agreement. To test whether the observed pattern depends on this particular normalization choice, we repeat the full analysis using the 50th, 75th, and 95th percentiles as alternative reference scales. Specifically, we replace the normalization factor $Q_{0.9}(\{D_{ij} : D_{ij} > 0\})$ with $Q_q(\{D_{ij} : D_{ij} > 0\})$ for $q \in \{0.50, 0.75, 0.95\}$, while keeping the local and global structural supports, the τ sweep, aggregation-aware calibration, permutation testing, and multiple-testing correction unchanged.

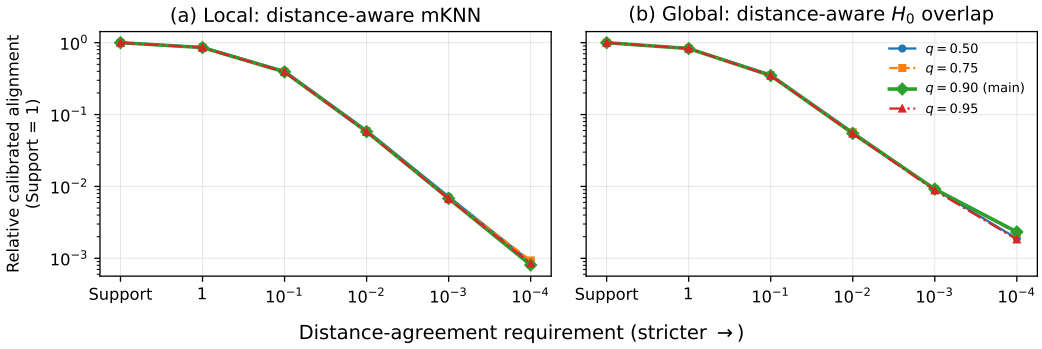


Figure 8: **Sensitivity to distance normalization.** Relative aggregation-aware calibrated alignment under alternative distance-normalization quantiles. The 90th percentile ($q = 0.90$) is the main setting used throughout the paper, while $q \in \{0.50, 0.75, 0.95\}$ provides alternative normalization scales. Across both **(a)** local distance-aware mKNN and **(b)** global distance-aware H_0 skeleton overlap, the alignment trajectories remain nearly unchanged across normalization choices as the distance-agreement requirement becomes more stringent.

Figure 8 shows that the resulting alignment trajectories are highly consistent across normalization quantiles. For local distance-aware mKNN, the alternative normalizations retain approximately 85-87% of the corresponding relational structure baseline at $\tau = 1$ and only about 6% at $\tau = 10^{-2}$. Global distance-aware H_0 skeleton overlap exhibits nearly the same transition, retaining approximately 83-84% at $\tau = 1$ and about 5-6% at $\tau = 10^{-2}$. Under still more stringent agreement, alignment approaches zero for every normalization choice. The trajectories closely track the main 90th-percentile setting throughout the full τ sweep.

Table 3 shows that the significance results preserve the same qualitative pattern. All 204 model pairs remain significant at both structural scales through $\tau = 10^{-1}$ under every normalization choice. At $\tau = 10^{-2}$, all local pairs remain significant, while 201-202 global pairs remain significant. Differences become more pronounced only under stringent agreement: at $\tau = 10^{-4}$, 176-182 local pairs and 58-109 global pairs remain significant across the alternative normalization quantiles. Al-

Table 3: **FDR-significant model pairs under alternative distance-normalization quantiles.** Values denote the number of significant pairs out of 204 after Benjamini-Hochberg correction.

q	Scale	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}
0.50	Local	204	204	204	204	182
	Global	204	204	201	186	84
0.75	Local	204	204	204	202	176
	Global	204	204	202	184	109
0.90 (main)	Local	204	204	204	204	183
	Global	204	204	201	174	82
0.95	Local	204	204	204	201	181
	Global	204	204	201	181	58

though the exact significance counts vary in this regime, global alignment consistently becomes less prevalent than local alignment.

Together, these results show that the weakening under increasingly stringent distance agreement, as well as the greater global fragility in the stringent regime, is robust to the choice of distance-normalization quantile.

F.2 Similarity-Aware Alignment

Our main analysis evaluates agreement using distances assigned to the shared relations. We next examine whether the observed weakening persists under an analogous similarity-based construction by replacing distance agreement with cosine-similarity agreement while keeping the underlying local and global structural supports fixed. For a shared relation (i, j), let

$$s_X(i,j) = \frac{x_i^\top x_j}{\|x_i\|\,\|x_j\|}, \qquad s_Y(i,j) = \frac{y_i^\top y_j}{\|y_i\|\,\|y_j\|}.$$

We replace the distance-agreement weight in Eq. 3 with

$$w_{ij,\mathrm{sim}}^{(\tau)} = \exp\left(-\frac{|s_X(i,j) - s_Y(i,j)|}{\tau}\right), \tag{17}$$

where smaller τ requires closer agreement in cosine similarity. The weight is applied only to relations shared by the two representations, using the same local kNN and global H_0 skeleton supports as in the main analysis. As $\tau \to \infty$, the weights approach one and the corresponding relational structure measures are recovered. We otherwise retain the same τ sweep, aggregation-aware calibration, permutation testing, and multiple-testing correction.

Figure 9 shows the full similarity-based sweep. Under relatively tolerant similarity agreement, both local and global measures retain a clear capacity-dependent structure, with larger vision models generally exhibiting stronger alignment. As the agreement requirement becomes more stringent, calibrated alignment decreases substantially and the capacity-dependent pattern becomes progressively less pronounced at both structural scales. Each panel uses an independently scaled vertical axis to expose the within-τ capacity pattern; the corresponding decrease in alignment magnitude is summarized in Table 4.

The similarity-aware construction closely reproduces the main transition in Table 1. At $\tau = 1$, local and global alignment retain 84.4% and 84.9% of their relational structure baselines, compared with 85.6% and 83.0% under the distance-aware construction. At $\tau = 10^{-2}$, the corresponding similarity-aware values fall to 4.73% and 5.57%, closely paralleling the 5.78% and 5.47% retained under distance agreement. Under still more stringent requirements, alignment approaches zero at both structural scales. Thus, the qualitative weakening persists when agreement on the shared relations is evaluated using cosine similarity rather than distance.

The exact significance trajectories show somewhat greater dependence on the choice of relation value. Similarity-aware alignment remains significant for more model pairs in the stringent regime, particularly at the global scale: 196 and 123 global pairs remain significant at $\tau = 10^{-3}$ and 10^{-4},

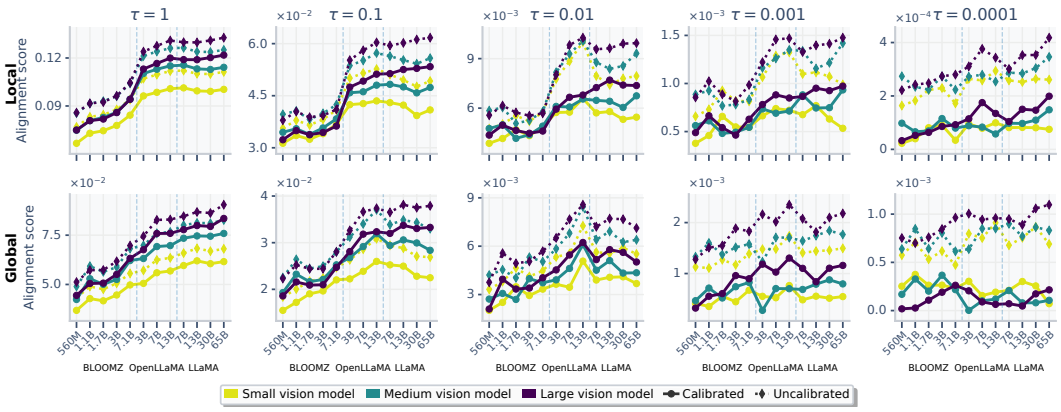


Figure 9: **Full similarity-aware alignment sweep.** Top and bottom rows show local similarity-aware mKNN and global similarity-aware H_0 skeleton overlap, respectively. Columns strengthen the cosine-similarity agreement requirement from $\tau = 1$ to 10^{-4}. Solid-circle and dotted-diamond lines denote calibrated and uncalibrated alignment. As agreement becomes more stringent, calibrated alignment weakens and the capacity-dependent pattern becomes less pronounced at both scales. Each panel uses an independently scaled vertical axis.

Table 4: **Numerical results for similarity-aware alignment.** Relative alignment is the mean aggregation-aware calibrated alignment across 204 model pairs, normalized by the corresponding relational structure mean. Significant pairs are determined after Benjamini-Hochberg correction.

	Relative calibrated alignment		FDR-significant pairs	
Setting	mKNN	H_0	mKNN	H_0
Support	1.0000	1.0000	204/204 (100.0%)	204/204 (100.0%)
$\tau = 1$	0.8437	0.8489	204/204 (100.0%)	204/204 (100.0%)
$\tau = 10^{-1}$	0.3554	0.3522	204/204 (100.0%)	204/204 (100.0%)
$\tau = 10^{-2}$	0.0473	0.0557	204/204 (100.0%)	204/204 (100.0%)
$\tau = 10^{-3}$	0.00571	0.00957	203/204 (99.5%)	196/204 (96.1%)
$\tau = 10^{-4}$	0.000763	0.002628	190/204 (93.1%)	123/204 (60.3%)

compared with 174 and 82 under the distance-aware construction. Nevertheless, the same qualitative pattern remains: alignment weakens at both structural scales as agreement becomes more stringent, while significant global alignment becomes less prevalent than local alignment in the stringent regime. These results show that the observed weakening is not specific to the particular distance-agreement construction used in the main analysis.

G Robustness under a Locally Estimated Riemannian Metric

Our main analysis shows robust convergence in relational structure, whereas requiring increasingly stringent agreement in the associated Euclidean distances progressively weakens alignment. This raises an alternative explanation: the observed weakening may depend specifically on the ambient Euclidean metric. We therefore repeat the distance-aware analysis using a locally estimated Riemannian metric.

Crucially, we keep the local kNN and global MST structural supports fixed and replace only the distances used to evaluate the corresponding relations. Reconstructing the supports under the Riemannian metric would change both relational structure and metric information simultaneously, reintroducing the confound that our framework is designed to avoid. The experiment therefore tests whether the same structure-geometry pattern persists when distance agreement is evaluated under an alternative, locally adaptive metric.

G.1 Riemannian Distance Construction

We construct a locally covariance-adapted Riemannian metric to test whether the observed distance-agreement pattern depends on the ambient Euclidean metric (Singer & Coifman, 2008; Berry & Sauer, 2016). The construction is applied independently to each representation layer. We first ℓ_2-normalize the representation vectors and, for each sample x_i, identify its k_R nearest neighbors in the ambient space. In our experiments, we use $k_R = 30$. The resulting symmetrized k_R-NN graph was connected for all representation layers considered in our experiments.

Because the normalized representations lie on the unit sphere, we map each neighbor x_j to the tangent space at x_i using the spherical logarithm map

$$v_{ij} = \log_{x_i}(x_j) = \frac{\theta_{ij}}{\sin\theta_{ij}} \left(x_j - \cos\theta_{ij}\, x_i\right), \qquad \theta_{ij} = \arccos(x_i^\top x_j).$$

Let V_i denote the matrix whose rows are the tangent vectors $\{v_{ij} : j \in \mathcal{N}_{k_R}(i)\}$. We estimate the local covariance structure as

$$C_i = \frac{1}{k_R} V_i^\top V_i$$

and retain its leading d_R eigenvectors $q_{i1}, \ldots, q_{id_R}$ with corresponding eigenvalues $\lambda_{i1}, \ldots, \lambda_{id_R}$. These eigenvectors define the estimated local tangent subspace

$$\widehat{\mathcal{T}}_i = \mathrm{span}\{q_{i1}, \ldots, q_{id_R}\}.$$

Let P_i denote the orthogonal projection onto $\widehat{\mathcal{T}}_i$. On this subspace, we define the regularized local quadratic form

$$g_i(u,u) = \sum_{r=1}^{d_R} \frac{\langle u, q_{ir}\rangle^2}{\lambda_{ir} + \gamma\bar{\lambda}_i}, \qquad \bar{\lambda}_i = \frac{1}{d_R}\sum_{r=1}^{d_R} \lambda_{ir}, \qquad u \in \widehat{\mathcal{T}}_i,$$

where $\gamma = 10^{-3}$ controls covariance regularization and $d_R = 10$.

For an undirected edge (i,j), we project the endpoint-local tangent displacements onto their respective estimated tangent subspaces and symmetrize the resulting quadratic forms. We define the Riemannian edge length as

$$\ell_R(i,j) = \left[\frac{1}{2}\left(g_i(P_i v_{ij}, P_i v_{ij}) + g_j(P_j v_{ji}, P_j v_{ji})\right)\right]^{1/2}.$$

These lengths define a symmetric weighted graph on the union of the ambient k_R-nearest-neighbor edges.

We use this locally estimated metric differently for the local and global distance-aware measures while keeping their structural supports fixed. For distance-aware mKNN, the original $k = 10$ neighbor identities are retained and only the distances assigned to these relations are replaced by $\ell_R(i,j)$. The edge lengths are normalized by the 90th percentile of the positive Riemannian graph-edge lengths before applying the same log-distance agreement weighting used in the main analysis.

For distance-aware H_0 skeleton overlap, the original MST edge identities remain those obtained from the ambient Euclidean metric. To obtain Riemannian distances for these globally spanning relations, we compute all-pairs shortest-path distances on the symmetric Riemannian-weighted k_R-NN graph,

$$D_R(i,j) = \min_{\pi: i \rightsquigarrow j} \sum_{(u,v)\in\pi} \ell_R(u,v),$$

and normalize the positive entries of D_R by their 90th percentile. The resulting values $D_R(i,j)$ are then evaluated only on the fixed MST support. Thus, the Riemannian experiment changes the metric used to evaluate distance agreement while preserving the local and global relational structures being compared.

G.2 Full Riemannian Results

Figure 10 shows the full capacity-dependent alignment patterns under the locally estimated Riemannian metric. Under relatively tolerant distance agreement, both local and global measures retain the capacity-dependent pattern observed when only relational structure is compared. As the distance-agreement requirement becomes more stringent, however, calibrated alignment weakens and the capacity-dependent pattern becomes progressively less pronounced at both structural scales. This qualitative behavior closely parallels the pattern obtained under the ambient Euclidean metric.

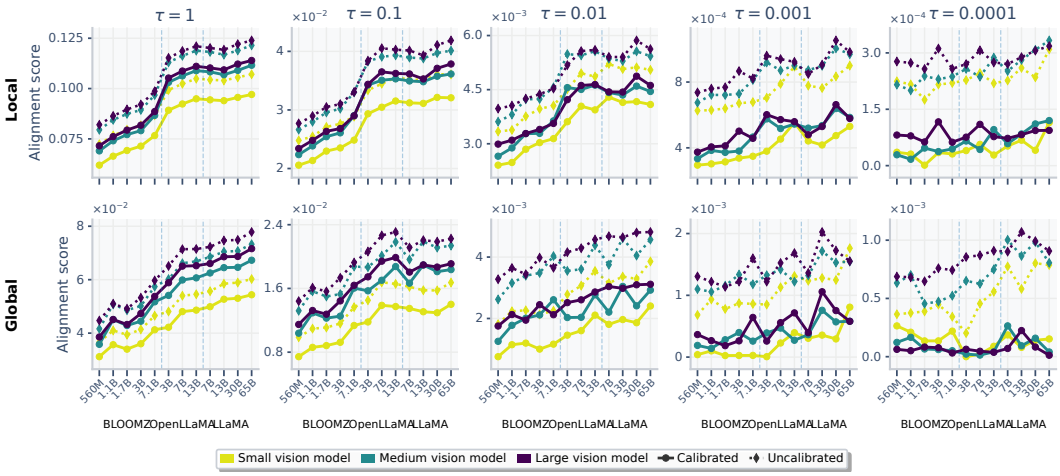


Figure 10: **Full capacity-dependent alignment patterns under a locally estimated Riemannian metric.** Top: local distance-aware mKNN. Bottom: global distance-aware H_0 skeleton overlap. Columns progressively strengthen the distance-agreement requirement from $\tau = 1$ to 10^{-4}. Solid lines show aggregation-aware calibrated alignment, and dotted lines show uncalibrated alignment. As distance agreement becomes more stringent, calibrated alignment weakens and the capacity-dependent pattern becomes less pronounced at both structural scales. Each panel uses an independently scaled vertical axis to visualize the within-τ capacity pattern; the cross-τ decrease in alignment magnitude is summarized in Table 5 and Figure 4a.

Table 5 quantifies this decay. At $\tau = 1$, the Riemannian variants retain 79.0% and 72.4% of the corresponding relational structure baselines for the local and global measures, respectively. These fractions decrease to 25.8% and 20.4% at $\tau = 10^{-1}$ and to only 3.24% and 2.97% at $\tau = 10^{-2}$. Under still stricter requirements, the retained alignment falls below 0.5% at $\tau = 10^{-3}$ and approaches zero at $\tau = 10^{-4}$. Thus, the pronounced decline is not specific to the ambient Euclidean metric.

Table 5: **Distance-aware results under a locally estimated Riemannian metric.** Relative alignment is the mean aggregation-aware calibrated alignment across 204 model pairs normalized by the corresponding relational structure mean. Euclidean relative alignment is included for reference. Significant pairs are reported as counts and percentages after Benjamini-Hochberg correction.

	Local		Global		Riemannian significant pairs	
τ	Euclidean	Riemannian	Euclidean	Riemannian	Local	Global
Support	1.000	1.000	1.000	1.000	204/204 (100%)	204/204 (100%)
1	0.856	0.790	0.830	0.724	204/204 (100%)	204/204 (100%)
10^{-1}	0.392	0.258	0.349	0.204	204/204 (100%)	204/204 (100%)
10^{-2}	0.0578	0.0324	0.0547	0.0297	204/204 (100%)	203/204 (99.5%)
10^{-3}	0.00681	0.00394	0.00919	0.00488	204/204 (100%)	160/204 (78.4%)
10^{-4}	0.000807	0.000525	0.002327	0.001399	175/204 (85.8%)	13/204 (6.4%)

Permutation significance exhibits the same overall progression. All 204 model pairs remain significant at both structural scales through $\tau = 10^{-1}$. At $\tau = 10^{-2}$, all local pairs and 203 of 204 global pairs remain significant. Under stricter agreement, the global measure becomes more fragile:

at $\tau = 10^{-3}$, 204 local pairs but 160 global pairs remain significant, and at $\tau = 10^{-4}$ the counts decrease to 175 and 13, respectively. Thus, increasingly stringent distance agreement weakens alignment at both scales, with an additional scale-dependent difference emerging in the stringent regime.

Taken together, these results show that replacing the ambient Euclidean metric with a locally covariance-adapted Riemannian metric does not restore the robust capacity-dependent convergence observed when only relational structure is compared. The same overall structure-geometry pattern therefore persists under this alternative metric.

H Additional Video-Text Results

We provide additional results for the video-text extension described in Section 4.5. We first examine relational structure alignment across individual visual-model families and then report the full distance-agreement sweep. Together, these analyses show that the qualitative structure-geometry pattern observed in the main vision-language setting extends to video-text representations: alignment in relational structure remains robust at both structural scales, whereas increasingly stringent distance agreement progressively weakens alignment and the capacity-dependent pattern.

For the capacity-dependent plots, we show the 12 language models belonging to the BLOOMZ, OpenLLaMA, and LLaMA families. The singleton Gemma-2-9B-IT model is omitted from these plots because it does not define a within-family capacity trajectory, but it remains included in the aggregate statistics over all 143 video-text model pairs.

H.1 Family-Wise Relational Structure Results

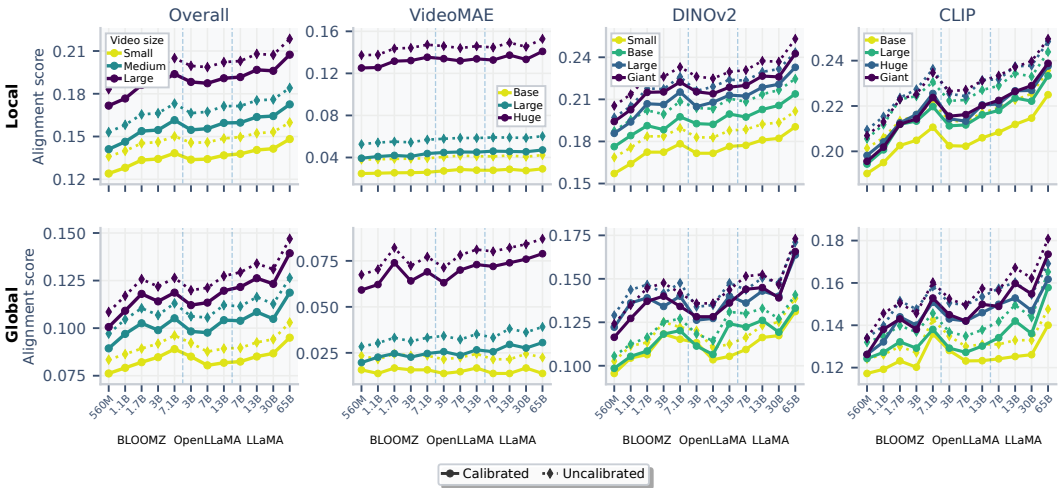


Figure 11: **Family-wise relational structure alignment in video-text representations.** Top: local relational structure measured by mKNN. Bottom: global spanning structure measured by H_0 skeleton overlap. The overall panels summarize small, medium, and large relative size categories across VideoMAE, DINOv2, and CLIP, while the remaining panels show individual variants within each family. Solid lines with circles show aggregation-aware calibrated alignment, and dotted lines with diamonds show uncalibrated alignment. Text-model parameter counts are shown along the horizontal axis. Each panel uses an independently scaled vertical axis to emphasize the within-family capacity-dependent pattern.

Figure 11 expands the relational structure analysis by visual-model family. The overall panels summarize relative small, medium, and large model variants across VideoMAE, DINOv2, and CLIP, while the remaining panels show the individual variants within each family. We report both aggregation-aware calibrated and uncalibrated alignment.

Across the three visual-model families, local mKNN and global H_0 skeleton overlap exhibit broadly similar capacity-dependent patterns. Calibration reduces the absolute alignment scores but preserves the qualitative trends: alignment generally increases with language-model capacity, and larger

visual-model variants tend to exhibit stronger relational alignment. Although the absolute score ranges differ across VideoMAE, DINOv2, and CLIP, the corresponding local and global patterns remain qualitatively consistent. Thus, the robust relational structure alignment observed in the aggregate analysis is not driven by a single visual-model family.

H.2 Full Distance-Agreement Sweep

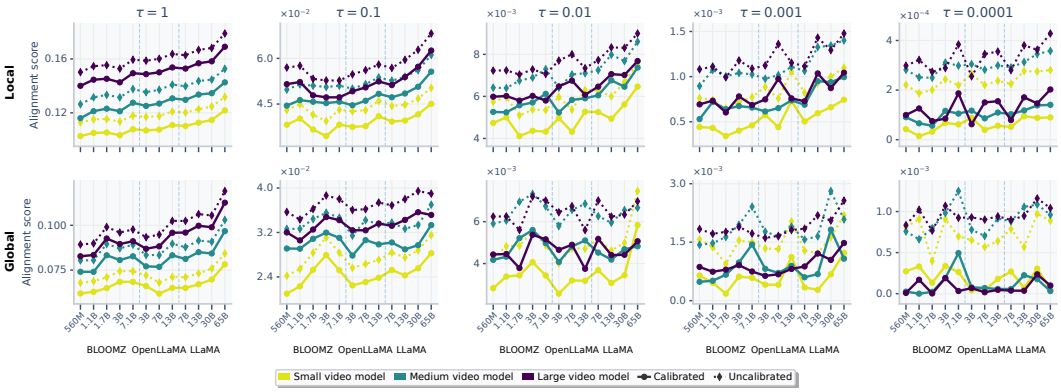


Figure 12: **Full distance-aware alignment sweep in video-text representations.** Top: local distance-aware mKNN. Bottom: global distance-aware H_0 skeleton overlap. Columns progressively strengthen the distance-agreement requirement from $\tau = 1$ to 10^{-4}. Colors denote relative visual-model size, averaged across VideoMAE, DINOv2, and CLIP. Solid lines with circles show aggregation-aware calibrated alignment, and dotted lines with diamonds show uncalibrated alignment. Each panel uses an independently scaled vertical axis to visualize the within-τ capacity pattern; the cross-τ decrease in alignment magnitude is summarized in Table 6 and Figure 4(b,c).

Figure 12 shows the full capacity-dependent video-text results as the distance-agreement requirement is progressively strengthened. At relatively tolerant values of τ, both local distance-aware mKNN and global distance-aware H_0 skeleton overlap retain the capacity-dependent pattern observed when only relational structure is compared. As τ decreases, however, calibrated alignment decreases sharply and the separation associated with model capacity becomes progressively less pronounced at both structural scales.

Importantly, this transition occurs in parallel for the local and global measures. The full capacity curves therefore support the aggregate result in Section 4.5: increasingly stringent distance agreement weakens alignment at both structural scales rather than selectively eliminating global alignment. Differences between the two scales become more apparent only under stringent distance agreement, particularly in the prevalence of model pairs that remain statistically significant.

Table 6: **Full numerical results for video-text representations.** Relative alignment is the mean aggregation-aware calibrated alignment across 143 video-text model pairs, normalized by the corresponding relational structure mean. Significant pairs are reported as counts and percentages after Benjamini-Hochberg correction.

	Relative calibrated alignment		FDR-significant pairs	
Setting	mKNN	H_0	mKNN	H_0
Support	1.0000	1.0000	143/143 (100.0%)	143/143 (100.0%)
$\tau = 1$	0.8109	0.7936	143/143 (100.0%)	143/143 (100.0%)
$\tau = 10^{-1}$	0.2930	0.2914	143/143 (100.0%)	143/143 (100.0%)
$\tau = 10^{-2}$	0.0365	0.0422	143/143 (100.0%)	142/143 (99.3%)
$\tau = 10^{-3}$	0.00439	0.00741	143/143 (100.0%)	131/143 (91.6%)
$\tau = 10^{-4}$	0.000596	0.001224	123/143 (86.0%)	67/143 (46.9%)

Table 6 quantifies the corresponding decay. At $\tau = 1$, local and global alignment retain 81.1% and 79.4% of their respective relational structure baselines. These fractions decrease almost identically to 29.3% and 29.1% at $\tau = 10^{-1}$, and further to 3.65% and 4.22% at $\tau = 10^{-2}$. Thus, the dominant reduction in alignment remains closely matched across structural scales.

A scale-dependent difference emerges more clearly in statistical significance. All 143 local model pairs remain significant through $\tau = 10^{-3}$, whereas the number of significant global pairs decreases from 142/143 at $\tau = 10^{-2}$ to 131/143 at $\tau = 10^{-3}$. At the most stringent setting, $\tau = 10^{-4}$, 123/143 (86.0%) local pairs remain significant, compared with 67/143 (46.9%) global pairs. These results reproduce the same structure-geometry pattern observed in the vision-language analysis: alignment becomes substantially less robust as increasingly stringent distance agreement is required at both scales, while structural scale contributes an additional difference mainly in the stringent regime.