Revisiting the Platonic Representation Hypothesis: An Aristotelian View

Fabian Gröger [*1 2 3] **Shuo Wen** [*1] **Maria Brbić** [1]

Abstract

The Platonic Representation Hypothesis suggests that representations from neural networks are converging to a common statistical model of reality. We show that the existing metrics used to measure representational similarity are *confounded by network scale*: increasing model depth or width can systematically inflate representational similarity scores. To correct these effects, we introduce a permutation-based null-calibration framework that transforms any representational similarity metric into a calibrated score with statistical guarantees. We revisit the Platonic Representation Hypothesis with our calibration framework, which reveals a nuanced picture: the apparent convergence reported by global spectral measures largely disappears after calibration, while local neighborhood similarity, but not local distances, retains significant agreement across different modalities. Based on these findings, we propose the *Aristotelian Representation Hypothesis*: representations in neural networks are converging to shared local neighborhood relationships.

1. Introduction

Quantifying the similarity between neural network representations is central to understanding the geometry of learned representation spaces (Raghu et al., 2017; Nguyen et al., 2021), guiding transfer learning decisions (Kornblith et al., 2019; Neyshabur et al., 2020; Gröger et al., 2025), and relating artificial representations to neural measurements in neuroscience (Schrimpf et al., 2018). The Platonic Representation Hypothesis (Huh et al., 2024) posits that as neural networks scale, representations across different modalities become increasingly similar, suggesting convergence to a shared statistical model of reality. This hypothesis has motivated a growing literature that uses representational similarity to study whether scaling produces universal structure across models (Huh et al., 2024; Maniparambil et al., 2024; Tjandrasuwita et al., 2025; Zhu et al., 2026). To measure representational similarity across models, different metrics have been proposed, such as Centered Kernel Alignment (Kornblith et al., 2019), Canonical Correlation Analysis (Weenink, 2003), Representational Similarity Analysis (Kriegeskorte et al., 2008), and mutual k-Nearest Neighbors (Huh et al., 2024).

The Aristotelian Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to shared *local neighborhood relationships*.

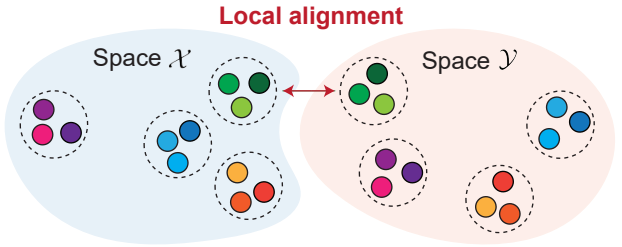


Figure 1. **The Aristotelian Representation Hypothesis:** Local relations ("who is near whom"), rather than distances between data points, are preserved across different representation spaces. Representation learning algorithms will converge to shared *local neighborhood relationships*.

In this work, we identify two pervasive confounders that distort representational similarity measurements. The first is the *model width*: when the embedding dimension increases relative to the sample size, interaction-matrix-based similarity metrics exhibit a systematic positive baseline even when representations are independent. This spurious similarity is a general consequence of dimensionality-driven null inflation: the expected similarity under independence does not vanish but instead depends on both the representation dimensionality and the sample size (Figure 2a). As a result, wider models can appear more aligned simply because their representations live in higher-dimensional spaces. The second confounder is the *model depth*. Many analyses do not compare individual layer pairs, because it is unknown

Project Page: brbiclab.epfl.ch/aristotelian
Code: github.com/mlbio-epfl/aristotelian
[*]Equal contribution [1]EPFL [2]University of Basel [3]HSLU. Correspondence to: Maria Brbić <maria.brbic@epfl.ch>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

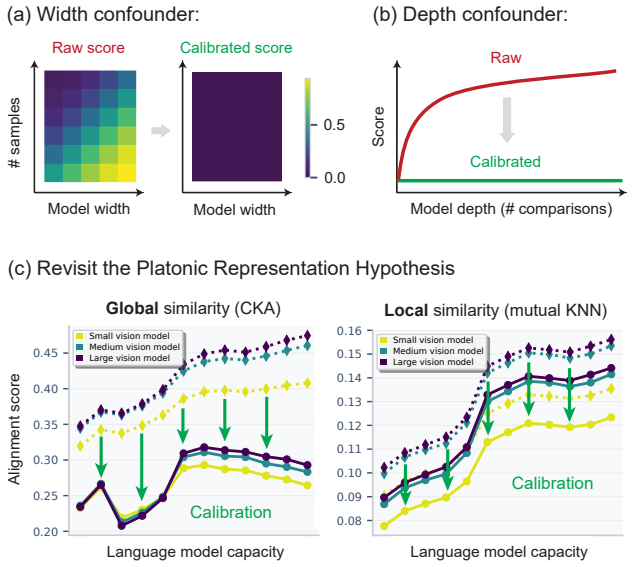


Figure 2. **Null calibration removes width and depth confounders.** (a) *Width* confounder: raw scores exhibit positive null baselines that increase with the ratio of dimension (width) of the spaces and the number of samples; calibration collapses them to zero. (b) *Depth* confounder: selection-based summaries (max over layers) inflate with search space size; aggregation-aware calibration removes this. (c) After calibration, global metrics lose their convergence trend, while local metrics retain significant alignment.

where similarity arises (Schrimpf et al., 2018; Huh et al., 2024). Instead, they search over all pairs and report a summary statistic such as the maximum. This is widespread practice: studies routinely compute similarity over many layer pairs and report an aggregate such as the maximum or best-matching layer (Raghu et al., 2017; Kornblith et al., 2019; Schrimpf et al., 2018; Huh et al., 2024; Kapoor et al., 2025). Taking a maximum over many comparisons inflates the reported score even if there is no similarity, since the expected maximum of independent draws exceeds the mean. This inflation grows with the number of comparisons, so deeper models can appear more aligned simply because more layer pairs are compared (Figure 2b). Together, these confounders undermine the comparative use of representational similarity without calibration.

To address these issues, we introduce the *null-calibration for representational similarity*, a general permutation-based framework that transforms any similarity metric into a calibrated score with a principled null reference, here defined as no relationship (Figure 2). The core idea is to measure how extreme an observed similarity is relative to an empirical null distribution obtained by breaking sample correspondences. For scalar comparisons (*i.e.*, width confounder), we estimate a critical threshold from the null distribution and define a calibrated score that is zero when the observed similarity falls below this threshold and rescaled to preserve the maximum at one. For selection-based summaries (*i.e.*, depth confounder), we apply *aggregation-aware* calibration. We compute the null distribution of the same aggregate statistic that is ultimately reported (*e.g.*, the maximum over all layer pairs), thereby calibrating the selection step itself.

These observations raise a question: *Does the Platonic Representation Hypothesis still hold once similarity is calibrated?* We find that, after calibration, the previously reported convergence in global metrics (Huh et al., 2024; Maniparambil et al., 2024; Tjandrasuwita et al., 2025) largely disappears, suggesting it was driven primarily by width and depth confounders, whereas local neighborhood-based metrics retain significant cross-modal alignment (Figure 2c). However, we also observe that the convergence in local distances is not preserved, suggesting that only local neighborhood relationships are aligned. Motivated by these results, we refine the original Platonic Representation Hypothesis and propose the *Aristotelian Representation Hypothesis*[1]: Neural networks, trained with different objectives on different data and modalities, converge to shared local neighborhood relationships (Figure 1). We name it after the Greek philosopher Aristotle, who was a student of Plato and, in his Categories, established the principles of relatives (Aristotle, ca. 350 B.C.E).

2. Related work

Representational similarity metrics. A long line of work compares representation spaces using a variety of similarity measures. Canonical Correlation Analysis (CCA) (Hotelling, 1992) and variants such as Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017) and Projection Weighted Canonical Correlation Analysis (PWCCA) (Morcos et al., 2018) compare subspaces up to linear transformations, while Procrustes- and shape-based distances compare representations up to restricted alignment classes (Ding et al., 2021; Williams et al., 2021). Centered Kernel Alignment (CKA) (Kornblith et al., 2019) has become a dominant tool for comparing deep representations, with kernelized variants extending to nonlinear similarity. Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), originating in neuroscience, compares representational dissimilarity matrices rather than feature bases. Neighborhood-based approaches, such as mutual k-Nearest Neighbors (mKNN) (Huh et al., 2024), capture local topological consistency rather than global alignment. However, recent evaluations stress that different metrics encode different invariances and can yield qualitatively different conclusions, motivating more robust reporting practices (Klabunde et al., 2025a;b; Ding et al., 2021; Harvey et al., 2024; Bo et al., 2024). This sensitivity

[1]Calling this refinement *Aristotelian*: it emphasizes learned representations converging on relations among instances (who is near whom) rather than the idea of convergence toward a globally matching structure.

is especially consequential in the neuro-AI literature, where the choice of similarity measure can profoundly change conclusions about which models best align with the brain (Soni et al., 2024), which is why scores that are calibrated and comparable across metrics and scales are needed.

Reliability of representational similarity metrics. In finite-sample, high-dimensional regimes, raw similarity scores can be systematically biased. Recent works (Murphy et al., 2024; Chun et al., 2025) propose debiased CKA, but these corrections are *metric-specific*. For neighborhood-based metrics, no analogous debiasing methods exist despite distance concentration effects that inflate random k-NN overlap (Beyer et al., 1999; Aggarwal et al., 2001). Other approaches address confounding from *input population structure*. For instance, Cui et al. (2022) propose regression-style deconfounding to remove effects of shared input statistics on RSA/CKA. Analogous concerns arise for regression-based predictivity in neuroscience, where high-dimensional fits can inflate alignment scores and may fail to identify genuinely brain-like models (Schaeffer et al., 2024). A separate reliability issue arises from layer search, where max or top-k aggregation across many layer pairs introduces multiple-comparison inflation. While resampling-based "maxT" procedures (Westfall & Young, 1993; Nichols & Holmes, 2002) can calibrate such aggregates, this has not yet been applied in representational similarity studies. Our calibration framework addresses both finite-sample bias and selection inflation in a unified, metric-agnostic way.

The Platonic Representation Hypothesis. A growing body of work examines whether neural networks trained under different conditions converge toward similar representations. The Platonic Representation Hypothesis (Huh et al., 2024) posits that as models scale, their representations increasingly converge across architectures and even across modalities such as vision and language, with convergence reported under both global and local similarity measures. Follow-up work has examined factors influencing these trends, including model size, training duration, and data distribution (Raugel et al., 2025), and has explored analogous convergence effects in broader settings such as video models (Zhu et al., 2026) and comparisons to biological vision (Marcos-Manchón & Fuentemilla, 2025). Prior work attributes convergence to shared factors such as training data distribution, scale, and objective functions (Raugel et al., 2025; Huh et al., 2024). Our calibration-based analysis does not contradict these findings but refines their interpretation, showing that the evidence for convergence depends on whether it is measured by global or local similarity. In this work, we revisit the Platonic Representation Hypothesis using our null-calibration framework that controls for width and depth confounders.

3. Problem setup

3.1. Representation spaces and similarity score

Let $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ be two representation spaces, where d_x and d_y are the respective space dimensions. For a set of n input samples, let $\mathbf{X} \in \mathbb{R}^{n \times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$ be the corresponding embeddings in $\mathcal{X}$ and $\mathcal{Y}$. We assume row-wise alignment such that the i-th row of $\mathbf{X}$ and $\mathbf{Y}$ correspond to paired inputs. We use a similarity score $s(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}$ to quantify the agreement between $\mathcal{X}$ and $\mathcal{Y}$. In practice, we compute it from $\mathbf{X}, \mathbf{Y}$ and, by a slight abuse of notation, denote it with $s(\mathbf{X}, \mathbf{Y})$.

We consider three families of metrics: *(i) spectral:* metrics defined on the spectrum of cross-covariance or Gram matrices (*e.g.*, CKA, CCA), *(ii) neighborhood:* metrics measuring local topological overlap (*e.g.*, mKNN), and *(iii) geometric:* second-order isomorphism metrics (*e.g.*, RSA). Appendix B provides definitions of the metrics used in this paper.

3.2. The null hypothesis of independence

We claim that a similarity score $s(\mathbf{X}, \mathbf{Y})$ is uninterpretable without a baseline. To provide this baseline, we define the null hypothesis H_0 as the absence of a relationship between $\mathbf{X}$ and $\mathbf{Y}$ beyond their marginal statistics. We operationalize H_0 via a permutation group Π_n acting on sample indices: draw $\pi \sim \mathrm{Unif}(\Pi_n)$ independently of $(\mathbf{X}, \mathbf{Y})$ and evaluate $s(\mathbf{X}, \pi(\mathbf{Y}))$, where $\pi(\mathbf{Y})$ permutes the rows of $\mathbf{Y}$.

Assumption 3.1 (Exchangeability under the null)**.** Under H_0, the joint distribution of paired samples is invariant to relabeling of correspondences. For any permutation $\pi \in \Pi_n$, $\mathbb{P}_{H_0}(\mathbf{X}, \mathbf{Y}) = \mathbb{P}_{H_0}(\mathbf{X}, \pi(\mathbf{Y}))$.

This assumption implies that if no true relationship exists, the observed pairing is statistically indistinguishable from a random shuffling of the data. It allows us to construct an empirical null distribution by holding $\mathbf{X}$ fixed and shuffling the rows of $\mathbf{Y}$.

3.3. Baseline problem: non-zero null expectations

Ideally, under H_0, we desire $\mathbb{E}_\pi[s(\mathbf{X}, \pi(\mathbf{Y}))] \approx 0$. However, for commonly used raw or *biased* estimators, the expected similarity under the null is *not zero*,

$$\mu_0(n, d_x, d_y) \coloneqq \mathbb{E}_\pi[s(\mathbf{X}, \pi(\mathbf{Y}))]. \tag{1}$$

This baseline μ_0 is metric- and preprocessing-dependent and can deviate from zero in finite samples. It also varies with sample size and dimension, thus acting as a confounding variable in comparative studies.

4. Theoretical motivation: spurious alignment

We motivate and formalize *why* raw representational similarity metrics fail in cross-scale model comparisons. We identify two distinct sources of confounding: *(i) the width confounder* driven by representation dimension, and *(ii) the depth confounder* driven by the number of layers considered when comparing models.

4.1. The width confounder

Many spectral-family similarity metrics, *e.g.*, linear/kernel CKA, the RV coefficient, and CCA-based scores (CCA/SVCCA/PWCCA), can be written as functionals of an *interaction operator* constructed from two representations. One such operator is the (normalized) cross-covariance

$$\widetilde{\mathbf{C}} \;=\; \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{Y}_c \in \mathbb{R}^{d_x \times d_y}, \tag{2}$$

where $\mathbf{X}_c$ and $\mathbf{Y}_c$ denote row-centered representations (Appendix B.1).

A common but misleading intuition is that if $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\widetilde{\mathbf{C}} \approx \mathbf{0}$ and therefore spectral aggregates should be near zero. In high dimension this fails: the null interaction energy is typically non-zero.

Proposition 4.1 (Non-vanishing null interaction energy)**.** *Assume the rows are i.i.d. with* $\mathbb{E}[\mathbf{x}_i] = \mathbb{E}[\mathbf{y}_i] = 0$, $\mathrm{Cov}(\mathbf{x}_i) = \mathbf{I}_{d_x}$, $\mathrm{Cov}(\mathbf{y}_i) = \mathbf{I}_{d_y}$, *and* $\mathbf{x}_i$ *and* $\mathbf{y}_i$ *are independent. Then*

$$\mathbb{E}_{H_0}\left[\|\widetilde{\mathbf{C}}\|_F^2\right] \;=\; \frac{d_x d_y}{n-1}. \tag{3}$$

Proof. See Appendix C.4.

Since CKA is scaled by the normalized self-similarity terms, which each scale as $\mathcal{O}(\sqrt{d})$, the resulting null baseline for the metric is thus $\mathcal{O}(d/n)$. We refer to this $\mathcal{O}(d/n)$ effect as the *width confounder*: at a fixed sample size n it grows with the representation width d, and it persists even when two representations share a genuine signal (Proposition C.5).

This aligns with insights from random matrix theory: in high-dimensional regimes ($d \sim n$), the null singular spectrum of interaction operators (after centering/whitening) concentrates into a non-trivial "noise bulk" whose upper edge depends on d/n and preprocessing, rather than collapsing to zero (Wachter, 1978; Müller, 2002; Livan et al., 2018). Our framework estimates this null baseline directly via permutation, providing a metric- and pipeline-independent alternative to asymptotic formulas.

Neighborhood metrics follow a different regime. While spectral metrics have null baselines scaling as $\mathcal{O}(d/n)$, neighborhood-based metrics such as mutual k-NN exhibit different behavior, as they rely on set comparisons rather than interactions.

Proposition 4.2 (Null baseline for neighborhood metrics)**.** *Assume the rows are i.i.d. with* $\mathbf{x}_i$ *and* $\mathbf{y}_i$ *independent, and that pairwise distances are almost surely distinct (e.g., under absolutely continuous distributions). Then for any* $k < n$*,*

$$\mathbb{E}_{H_0}\Big[\mathrm{mKNN}(\mathbf{X}, \mathbf{Y})\Big] \;=\; \frac{k}{n-1}. \tag{4}$$

Proof. See Appendix C.6.

In particular, neighborhood metrics have null baselines scaling as $\mathcal{O}(k/n)$.

The difference in null baseline between spectral and neighborhood metrics is substantial: *(i)* The neighborhood scale k can be fixed consistently across experiments, whereas the embedding dimension d is determined by the model architecture, making it difficult to control in comparison studies. *(ii)* The neighborhood metrics are much less confounded since $k \ll d$ in typical settings. Geometric metrics (*e.g.*, RSA, Procrustes), similarly to the spectral metrics, are functions of all-pairs distances or inner products, which are the interaction quantities of Proposition 4.1. They thus inherit a width-dependent $\mathcal{O}(d/n)$ baseline rather than the $\mathcal{O}(k/n)$ baseline of rank-based neighborhood metrics. We verify this empirically for RSA and Procrustes in Appendix E.7.

4.2. The depth confounder

A subtle yet pervasive issue is the comparison of *selection-based* alignment summaries across models. Let $S_{\ell,\ell'} := s(\mathbf{X}_\ell^{(A)}, \mathbf{Y}_{\ell'}^{(B)})$ be the similarity between layer ℓ of model A and layer ℓ' of model B. It is common to summarize the similarity between two models by the maximum alignment score $T_{\max} = \max_{\ell,\ell'} S_{\ell,\ell'}$ (Schrimpf et al., 2018; Huh et al., 2024; Raghu et al., 2017; Kornblith et al., 2019). Let $M = L_A L_B$ be the number of layer pairs searched, where L_A and L_B are the depths of models A and B. Even under H_0, taking a maximum over M comparisons inflates the reported score, a "look-elsewhere" effect. This is an instance of the classical multiple comparisons problem (Benjamini & Hochberg, 1995; Bonferroni, 1936): as M increases, the probability that at least one null similarity exceeds any fixed threshold grows, inflating the expected maximum. Consequently, when alignment is summarized via a max or top-k statistic without correction, unrelated representations can exhibit spuriously high reported similarity, as the inflation depends on model depth, making raw summaries non-comparable across architectures.

Characterizing this inflation does not require independence across pairs. It follows from a uniform right-tail bound. Assume there exist a common mean $\mu \in \mathbb{R}$ and $\sigma > 0$ such

that the null fluctuations satisfy, for all (ℓ, ℓ') and all $t \geq 0$,

$$\mathbb{P}(S_{\ell,\ell'} - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{5}$$

For bounded similarities $S_{\ell,\ell'} \in [s_{\min}, s_{\max}]$, Hoeffding's inequality implies a sub-Gaussian right-tail bound of the form Equation (5) with $\sigma \leq (s_{\max} - s_{\min})/2$. This covers many common bounded metrics (*e.g.*, CKA/RSA/mKNN). Crucially, only the right tail is needed for bounding the maximum. Then a union bound gives

$$\mathbb{P}(T_{\max} - \mu \geq t) \leq M \exp\left(-\frac{t^2}{2\sigma^2}\right), \tag{6}$$

and consequently for a constant C

$$\mathbb{E}_{H_0}[T_{\max}] \;\leq\; \mu + C\,\sigma\sqrt{\log M}. \tag{7}$$

Proof. See Appendix C.5.

This creates a depth confounder. Deeper models (larger $M = L_A L_B$) can attain higher raw "max-alignment" scores purely because of a larger search space. Correlations across neighboring layers reduce the *effective* number of comparisons, but the inflation remains monotone in the search space size in typical workflows. Therefore, raw scaling plots of $T_{\max}$ (or top-k summaries) are not comparable across architectures unless the *selection step itself* is calibrated.

The two confounders are not independent but hierarchical: the depth confounder operates on top of the width confounder. The width confounder gives each individual layer pair a positive chance similarity, and deeper models supply a larger pool of such chance similarities, so the selection maximum drifts higher.

5. Representational similarity calibration

To overcome the issues of the width and depth confounders, we introduce the *null-calibration for representational similarity*. The key idea is to compare observed similarity scores against an empirical null distribution obtained by permuting sample correspondences, thereby establishing a principled zero point that accounts for finite-sample, high-dimensional artifacts.

5.1. Null-calibrated similarity

We propose *null-calibrated* similarity measures to correct for width and depth confounders by transforming raw similarity scores into an effect size with a principled zero point.

Given representations $\mathbf{X} \in \mathbb{R}^{n \times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$ aligned by rows, we operationalize the null hypothesis H_0 (no relationship beyond marginal statistics) by permuting sample correspondences. For permutations $\pi_k \in \Pi_n$ drawn i.i.d. uniformly from Π_n and independently of $(\mathbf{X}, \mathbf{Y})$, we form null scores

$$s^{(k)} = s(\mathbf{X}, \pi_k(\mathbf{Y})), \qquad k = 1, \ldots, K. \tag{8}$$

Let $s_{\text{obs}} := s(\mathbf{X}, \mathbf{Y})$ denote the observed score. Let $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(K+1)}$ denote the order statistics of the *combined* multiset $\{s_{\text{obs}}, s^{(1)}, \ldots, s^{(K)}\}$ (with ties allowed). We define a right-tail rank-based critical value:

$$\tau_\alpha \;:=\; s_{(\lceil (1-\alpha)(K+1) \rceil)}, \tag{9}$$

where $\lceil (1-\alpha)(K+1) \rceil$ is the $(1-\alpha)$-quantile of the $(K+1)$-sized multiset and the empirical right-tail p-value:

$$p \;=\; \frac{1 + \#\{k \in \{1, \ldots, K\} : s^{(k)} \geq s_{\text{obs}}\}}{K+1}. \tag{10}$$

The critical value τ_α defines a robust zero point: values below τ_α are typical under H_0 at level α, while p provides an evidence measure that can be combined with multiple-testing correction when many comparisons are performed.

The proposed calibration framework relies on *randomization* (permutation) to construct a null distribution for any similarity statistic. This yields finite-sample guarantees under an exchangeability condition (Assumption 3.1), and it implies useful invariances that make calibrated scores comparable across metrics and implementations.

The permutation p-value in Equation (10) is *super-uniform* under H_0 (*i.e.*, $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$), a standard consequence of randomization inference (Nichols & Holmes, 2002; Phipson & Smyth, 2010; Good, 2005) (see Appendix C.1 for formal definitions and proofs).

Corollary 5.1 (Type-I control for calibrated scores)**.** *Let $s_{\text{obs}} = s(\mathbf{X}, \mathbf{Y})$ and $s^{(k)} = s(\mathbf{X}, \pi_k(\mathbf{Y}))$ for $k = 1, \ldots, K$. Define the add-one permutation p-value p as in Equation (10), and equivalently define the rank-based critical value $\tau_\alpha := s_{(\lceil (1-\alpha)(K+1) \rceil)}$ from the sorted combined set $\{s_{\text{obs}}, s^{(1)}, \ldots, s^{(K)}\}$. Under Assumption 3.1,*

$$\mathbb{P}_{H_0}\big(p \leq \alpha\big) \leq \alpha \quad \textit{and hence} \quad \mathbb{P}_{H_0}\big(s_{\text{obs}} > \tau_\alpha\big) \leq \alpha, \tag{11}$$

so the gating rule "$s_{\text{cal}} > 0$" (where s_{cal} is the calibrated score defined in Equation (12), which implies $p \leq \alpha$) is a finite-sample α-level declaration of similarity above chance.

Proof. Follows directly from Lemma C.2; see Appendix C.1.

Calibrated score (scalar case). While p-values and null percentiles are rank-based and therefore invariant under monotone transformations of the raw score (Proposition C.3; see Appendix C.2), effect sizes serve a complementary purpose: they quantify *how much* similarity exceeds chance on an interpretable scale. The calibrated score achieves this by

rescaling the excess over the null threshold τ_α to the interval $[0, 1]$. This rescaling is not monotone-invariant, by design. A purely rank-based calibration would be equivalent to a score shift and would be unable to correct for the scale-dependent null baselines identified in Section 4. The calibrated score instead adapts to the actual null distribution, providing a meaningful zero point.

For bounded similarity metrics with known maximum $s_{\max}$ (often $s_{\max} = 1$), we define a max-preserving calibrated score

$$s_{\text{cal}} = \max\left(\frac{s_{\text{obs}} - \tau_\alpha}{s_{\max} - \tau_\alpha}, 0\right). \tag{12}$$

This calibrated score depends on the chosen level α through τ_α (Equation (9)). We therefore also report the corresponding permutation p-value and/or null percentile for an α-free summary. This score satisfies $s_{\text{cal}} = 0$ whenever $s_{\text{obs}} \le \tau_\alpha$ (*i.e.*, below the estimated right-tail critical value of the permutation null), and $s_{\text{cal}} = 1$ when $s_{\text{obs}} = s_{\max}$ (*i.e.*, perfect similarity remains 1). When $s_{\max}$ is unknown, or the metric is unbounded, we default to the unnormalized effect size $[s - \tau_\alpha]_+ = \max(s - \tau_\alpha, 0)$.

5.2. Aggregation-aware null-calibration

To analyze the similarity between two models A and B with depths L_A and L_B, a common approach is to compute a layer-by-layer similarity matrix $\mathbf{S} \in \mathbb{R}^{L_A \times L_B}$ by evaluating a similarity score for every pair of layers:

$$S_{\ell,\ell'} = s\left(\mathbf{X}_\ell^{(A)}, \mathbf{Y}_{\ell'}^{(B)}\right), \tag{13}$$

where $\mathbf{X}_\ell^{(A)} \in \mathbb{R}^{n \times d_\ell}$ and $\mathbf{Y}_{\ell'}^{(B)} \in \mathbb{R}^{n \times d_{\ell'}}$ are the representations of models A and B at layers ℓ and ℓ' respectively, evaluated on n samples, and $s(\cdot, \cdot)$ is a similarity metric. A common practice is then to summarize $\mathbf{S}$ by a *selection-based* aggregation operator, such as taking the maximum. These summaries are attractive because they support statements such as "there exists a layer in A that matches some layer in B" or "each layer of A best matches a layer in B". However, selection introduces a statistical effect: even under the null hypothesis of no relationship between representations, selection-based summaries are systematically inflated.

As analyzed in Section 4.2, this inflation grows with the number of layer pairs and makes naïve post-selection p-values anti-conservative. Our aggregation-aware calibration addresses this by calibrating the *reported* statistic directly: the null distribution must match the *entire* analysis pipeline. Let the aggregate score be $T(\mathbf{S})$ (*e.g.*, a maximum), then the appropriate null is the distribution of $T(\mathbf{S})$ under a valid null transformation (*e.g.*, permuting sample correspondences). We therefore define an aggregation-aware permutation null. When T is the maximum over layer pairs, this recovers the classical "maxT" procedure of Westfall & Young (1993). Our formulation generalizes it to arbitrary selection-based aggregates (*e.g.*, top-k or row-max mean) and to any similarity metric.

Consistency of permutations across layers. For each draw $\pi_k \in \Pi_n$, we apply the *same* sample permutation to all layers of model B and define

$$S_{\ell,\ell'}^{(k)} := s\left(\mathbf{X}_\ell^{(A)}, \pi_k\left(\mathbf{Y}_{\ell'}^{(B)}\right)\right), \tag{14}$$
$$\ell = 1, \dots, L_A, \quad \ell' = 1, \dots, L_B,$$

then compute $T^{(k)} := T(\mathbf{S}^{(k)})$. Let $T_{\text{obs}} := T(\mathbf{S})$ denote the observed aggregate. Let $T_{(1)} \le \cdots \le T_{(K+1)}$ denote the order statistics of the combined set $\{T_{\text{obs}}, T^{(1)}, \dots, T^{(K)}\}$ (with ties allowed). We define

$$\tau_\alpha^{\text{agg}} := T_{(\lceil (1-\alpha)(K+1) \rceil)}, \tag{15}$$

where $\lceil (1-\alpha)(K+1) \rceil$ is the $(1-\alpha)$-quantile of the $(K+1)$-sized multiset. We report the right-tail permutation p-value

$$p_{\text{agg}} = \frac{1 + \#\{k \in \{1, \dots, K\} : T^{(k)} \ge T_{\text{obs}}\}}{K+1}, \tag{16}$$

By the same exchangeability argument as for scalar calibration, p_{agg} is super-uniform under H_0 (see Proposition C.4).

Calibrated score (aggregate case). For bounded similarities with maximum $s_{\max}$ (often $s_{\max} = 1$), we report a max-preserving calibrated aggregate

$$T_{\text{cal}} = \max\left(\frac{T_{\text{obs}} - \tau_\alpha^{\text{agg}}}{s_{\max} - \tau_\alpha^{\text{agg}}}, 0\right). \tag{17}$$

This score satisfies $T_{\text{cal}} = 0$ when $T_{\text{obs}} \le \tau_\alpha^{\text{agg}}$ and $T_{\text{cal}} = 1$ when $T_{\text{obs}} = s_{\max}$. As above, T_{cal} depends on α via τ_α^{agg}; we therefore report both T_{cal} (magnitude above null) and p_{agg} (evidence against null), applying multiplicity correction (Holm, 1979; Benjamini & Hochberg, 1995) when many model pairs are evaluated.

5.3. Summary

To compute a calibrated similarity score: *(i)* fix a significance level α (*e.g.*, $\alpha = 0.05$); *(ii)* generate K null scores by permuting sample correspondences; *(iii)* compute critical value τ as the $\lceil (1-\alpha)(K+1) \rceil$-th order statistic of the combined set (observed + null scores); *(iv)* return calibrated score, either s_{cal} or T_{cal}.

Use *scalar calibration* (Section 5.1) when comparing a single pair of representations. Use *aggregation-aware calibration* (Section 5.2) when reporting a summary statistic (*e.g.*, maximum) over multiple layer pairs. Appendix D provides pseudocode for both procedures.

6. Experiments

We quantify the effects of the width and depth confounders in controlled synthetic experiments and show that our calibration framework effectively removes them. We then revisit the Platonic Representation Hypothesis using our calibration framework, assessing which convergence trends remain robust after controlling for these confounding factors.

6.1. Null-calibration removes width confounder

We validate that our calibration eliminates width-related inflation of similarity across metrics, regimes, and noise distributions, without metric-specific derivations.

We design controlled synthetic experiments as follows. Under H_0, we draw $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$ independently from Gaussian and heavy-tailed (Student-t, Laplace) distributions. We vary the number of samples $n \in \{128, 256, 512, 1024, 2048, 4096\}$ and the dimension $d \in \{128, 256, 512, 1024, 2048\}$. Under H_1, we inject a shared low-rank signal component and vary the signal-to-noise ratio. We evaluate representative metrics spanning three families. For spectral similarity, we use linear and RBF CKA, as well as CCA/SVCCA/PWCCA; for neighborhood similarity, we use mKNN (with $k = 10$); and for geometric similarities, we use RSA and Procrustes. Figure 3 reports a subset of these metrics for readability; additional metrics are reported in Appendix E.7. For calibration, we use $K = 200$ permutations with $\alpha = 0.05$.

Under H_0, uncalibrated scores increase with d/n, while our calibrated scores stay at zero across settings (Figure 3). This confirms that the similarity scores of wider models can arise purely from high-dimensional finite-sample effects, and our calibration removes this spurious baseline. Importantly, the magnitude of the null baseline is metric-dependent, consistent with our theory: CKA's baseline scales as $\mathcal{O}(d/n)$ (Proposition 4.1), while mKNN's baseline scales as $\mathcal{O}(k/n)$ (Proposition 4.2). Intuitively, mKNN compares local neighborhood overlap at a fixed k, thus only comparing relationships instead of local distances, making its null baseline insensitive to representation width d, which explains the order-of-magnitude gap observed in raw scores. The same pattern holds for heavy-tailed noise (Appendix E.7).

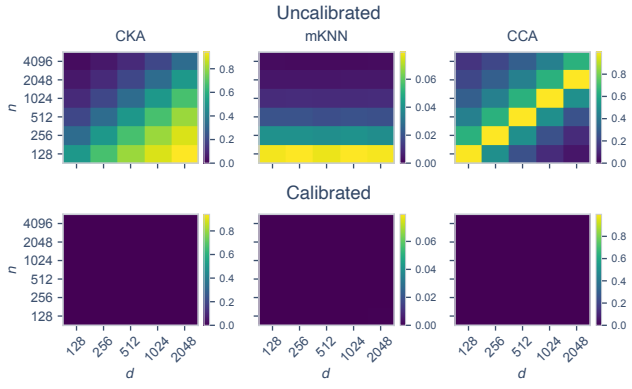


Figure 3. **Calibration eliminates spurious similarity across metrics.** Raw scores (top) drift with d/n; calibrated scores (bottom) collapse to zero. Results for heavy-tailed distributions and additional metrics are in Appendix E.7.

The same width inflation arises under a genuine shared signal, and calibration corrects it there as well, as we show analytically (Proposition C.5) and on real networks at fixed n (Appendix E.5).

Next, we verify the statistical guarantees of our empirical null calibration. For Type-I error control, rejection rates stay at or below the nominal $\alpha = 0.05$ across $(n, d/n)$ configurations (Figure 4a). Crucially, our calibration does not sacrifice sensitivity to real alignment: detection rates increase rapidly with signal strength (Figure 4b). Overall, our calibration preserves signal structure: in the high-signal regime, raw and calibrated scores show the same pattern, while in the low-signal regime, calibration correctly gates scores to zero (Appendix E.2).

Furthermore, we verify that our empirical calibration closely matches existing analytical bias corrections for CKA (Murphy et al., 2024), recovering the width correction without metric-specific derivation (Appendix E.4).

Additionally, we perform ablations on different noise distributions used in the synthetic experiments (Appendix E.1), different calibration approaches (Appendix E.3), and an ablation on the influence of the number of permutations K used for calibration (Appendix E.6).

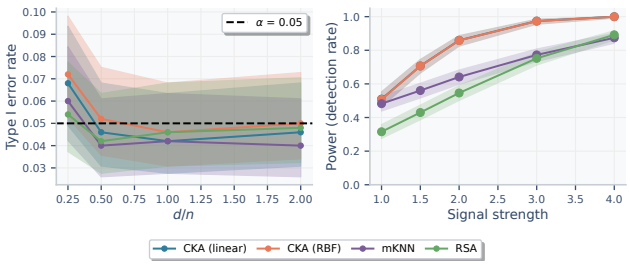
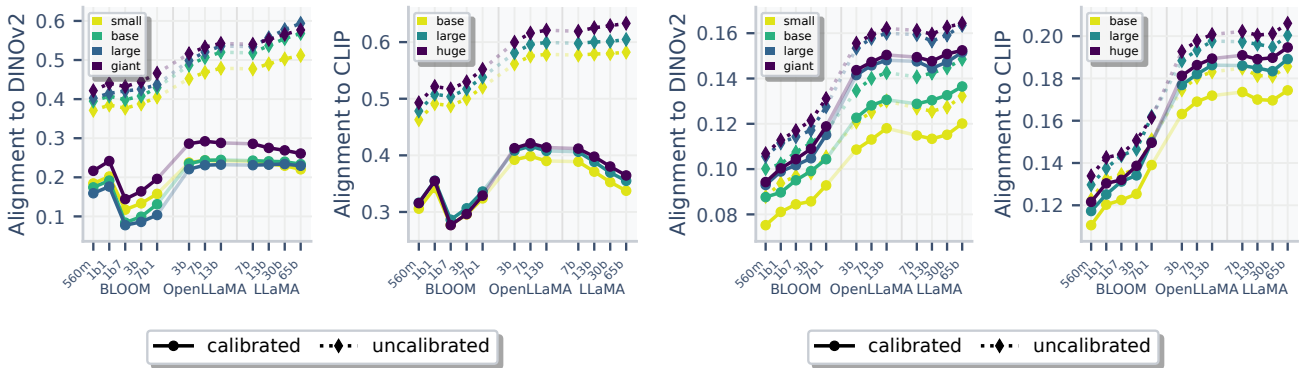


Figure 4. **Statistical guarantees.** (Left) Type-I error stays at or below α across configurations. (Right) Power increases rapidly with signal strength; calibration does not sacrifice sensitivity.

6.2. Null-calibration removes depth confounder

We validate that our aggregation-aware null-calibration eliminates the depth confounder. To build a controlled synthetic setting, we construct two synthetic models, A and B, each with L layers. Under H_0, we sample layer representations $\{\mathbf{X}_\ell\}_{\ell=1}^{L}$ and $\{\mathbf{Y}_{\ell'}\}_{\ell'=1}^{L}$, where each $\mathbf{X}_\ell, \mathbf{Y}_{\ell'} \in \mathbb{R}^{n \times d}$ has i.i.d. $\mathcal{N}(0, 1)$ entries (independent across layers and between models), using $d/n = 8$ to match the upper range of the Platonic Representation Hypothesis setting. We then compute the layerwise similarity matrix $S_{\ell,\ell'} = \mathrm{CKA}_{\mathrm{lin}}(\mathbf{X}_\ell, \mathbf{Y}_{\ell'})$ and summarize it with standard aggrega-



(a) CKA RBF: Global spectral alignment.

(b) mKNN: Local neighborhood overlap.

Figure 5. **Revisiting the Platonic Representation Hypothesis.** Models are ranked according to their language performance (Huh et al., 2024). Solid lines connect the models within the same family, while semi-transparent lines connect the models across different families. (a) Global spectral metrics lose their convergence trend; calibrated scores show no systematic increase with scale. (b) Local neighborhood metrics keep their trend even after calibration. Full results for all vision families and metrics in Appendix E.8.

tion operators.

The uncalibrated max-aggregated scores inflate with layer count even under H_0 (Figure 6): raw max-scores are systematically higher at $L = 128$ than at $L = 1$, despite no true signal. Our aggregation-aware calibration eliminates this bias: calibrated aggregates remain stable regardless of depth. We further show that naively calibrating each scalar comparison still leads to inflation, highlighting the importance of calibrating the final statistic. Furthermore, since deeper models tend to be wider as well, raw comparisons are doubly confounded.

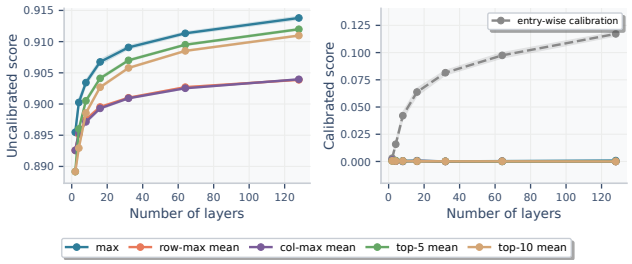


Figure 6. **Aggregation-aware calibration removes depth confounding.** Raw max-aggregates of linear CKA scores inflate with layer count under the null; calibrated aggregates are stable and show that naive entry-wise calibration still leads to inflation.

6.3. Revisiting the Platonic Representation Hypothesis

A central claim behind the Platonic Representation Hypothesis is that, as models become more capable, their representations begin to *converge* across modalities. We revisit this claim through our calibration framework to determine whether the observed alignment reflects genuine shared representation structure or instead arises from width and depth confounders.

We follow the experimental protocol of Huh et al. (2024) using $n = 1024$ image–text pairs (WIT; Srinivasan et al. (2021)) and embeddings from three language model families (BLOOM, OpenLLaMA, LLaMA) and five vision model families (ImageNet-21K, MAE, DINOv2, CLIP, CLIP-finetuned) across multiple scales. As in Huh et al. (2024), we extract per-layer representations using the class token for vision models and mean pooling over tokens for language models. This yields 204 vision–language model pairs spanning $d/n \in [0.75, 8]$. For each pair, we compute layer-wise similarity and report the maximum across layers, as in the original work. We evaluate both global spectral metrics (CKA linear/RBF) and local neighborhood metrics (mKNN, cycle-kNN, CKNNA). Following Huh et al. (2024), we evaluate mKNN, cycle-kNN, and CKNNA with $k = 10$. We further apply Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) to control for multiple comparisons across model pairs.

For the *global* similarity, we find that uncalibrated CKA scores increase with model scale (dotted lines in Figure 5a), reproducing the trend interpreted as evidence of cross-modal convergence (Huh et al., 2024). However, this trend disappears after our calibration (solid lines): calibrated CKA shows no systematic increase with model size. This indicates that global convergence in uncalibrated CKA is largely attributable to width and depth confounders rather than a genuine increase in representational similarity. The same holds for the other global metrics, including shape- and geometry-based ones: SVCCA, the RV coefficient, and Procrustes distance all lose their apparent scaling trend after calibration (Appendix E.8).

In contrast, for the *local* similarity, evidence of cross-modal convergence remains strong for neighborhood-based metrics even under our calibration (Figure 5b). The same qualitative conclusion holds for other neighborhood-based measures (cycle-kNN and CKNNA; Appendix E.8) and different choices of α (Appendix E.11). Further analysis (Appendix E.10) reveals that models converge in local neighbor-

hood structure: models increasingly agree on which points are neighbors, but do not agree on the pairwise distances, since CKA-RBF with a small bandwidth shows no alignment after calibration.

CKA and mKNN capture different invariances, so their raw magnitudes are not directly comparable. We therefore compare, for each metric separately, how strongly its score tracks language-model capability before and after calibration (its Pearson correlation with the model ranking). After calibration this correlation collapses for global metrics (for linear CKA it falls from 0.86 to 0.45, and for the RV coefficient from 0.92 to 0.31) but is essentially unchanged for local ones (mKNN stays near 0.85 and CKNNA near 0.87). Appendix E.8 reports all metrics.

To test whether these findings generalize beyond images and text, we extend our analysis to video–language alignment following Zhu et al. (2026). We compare video encoders (VideoMAE small/base/large/huge, fine-tuned on Kinetics) against the same language model families. Consistent with our previous findings, the global similarity (CKA) shows no trend with model capacity (Figure 7). In contrast, for local similarity (mKNN), a clear scaling trend emerges with VideoMAE-Large/Huge, whereas smaller video encoders appear to act as a bottleneck, limiting alignment regardless of language model size. This confirms that local neighborhood convergence extends to video–language alignment, provided that representations are sufficiently powerful. Appendix E.9 further compares non-finetuned VideoMAE versions and a variety of image models at the frame level on the same dataset, showing the same trend.

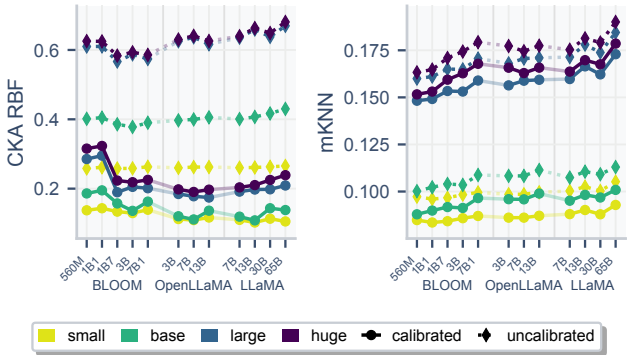


Figure 7. **Video–language alignment.** Extending the Platonic Representation Hypothesis analysis to video encoders (VideoMAE small/base/large/huge) yields the same pattern: calibrated CKA drops substantially while mKNN retains alignment.

Taken together, these results suggest a refined version of the Platonic Representation Hypothesis. After calibration, we find little evidence that representations converge in global spectral structure as models scale, at least under the considered setting. What reliably persists is local geometric alignment: different models preserve similar neighborhood relationships among inputs. We therefore propose the alternative **Aristotelian Representation Hypothesis**: *As models become capable, their representations converge to shared local neighborhood relationships.*

7. Conclusion

Representational similarity metrics are widely used to study learned features, but their interpretation is systematically distorted by two artifacts: width-dependent null baselines and depth-dependent selection inflation. We introduced a unified null-calibration framework that corrects both, turning similarity scores into effect sizes with principled zero points and valid p-values. Applying our framework to the Platonic Representation Hypothesis reveals that previously reported global spectral convergence is largely confounded by width and depth, whereas local neighborhood alignment remains significant, motivating an Aristotelian Representation Hypothesis.

Relationship to the Platonic hypothesis. The Aristotelian hypothesis refines rather than refutes the Platonic one. Global convergence implies local convergence, because matching the full geometry preserves neighborhoods, but the converse does not hold. The Platonic hypothesis therefore implies the Aristotelian one, which is a weaker criterion. Our experiments support this local form and, after calibration, find little evidence for the global one. Prior reports of convergence thus remain compatible with our analysis, and the Aristotelian hypothesis offers a more precise lens on what converges across models and modalities.

Limitations and outlook. Our conclusions come with two caveats. First, representational similarity has no ground-truth scale, so we report the presence or absence of calibrated evidence for convergence rather than proving it. Second, our guarantees assume exchangeability (Assumption 3.1), so grouped or clustered samples require restricted permutations that preserve their dependence structure (Appendix C.1), at only a modest, metric-agnostic cost. Why representations converge in local neighborhoods but not in global geometry remains the key open question.

Acknowledgements

We thank Artyom Gadetsky, Siba Smarak Panigrahi, Debajyoti Dasgupta, David Frühbuss, Shin Matsushima, Rishubh Singh, Adriana Moreno Castan, and Gioele La Manno for their valuable suggestions, which helped improve the manuscript. We are especially grateful to Simone Lionetti for additional input and support. We gratefully acknowledge the support of the Swiss National Science Foundation (SNSF) starting grant TMSGI2_226252/1, SNSF grant IC00I0_231922, and the Swiss AI Initiative. M.B. is a CIFAR Fellow in the Multiscale Human Program.

Impact statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*. Springer, 2001.

Aristotle. Categories, ca. 350 B.C.E.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. When is "nearest neighbor" meaningful? In *International Conference on Database Theory*. Springer, 1999.

Bo, Y., Soni, A., Srivastava, S., and Khosla, M. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv preprint arXiv:2411.14633*, 2024.

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

Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936.

Cai, M. B., Schuck, N. W., Pillow, J. W., and Niv, Y. Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLoS Computational Biology*, 2019.

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

Chun, C., Canatar, A., Chung, S., and Lee, D. D. Estimating Neural Representation Alignment from Sparsely Sampled Inputs and Features. *arXiv preprint arXiv:2502.15104*, 2025.

Cramér, H. *Mathematical methods of statistics*. 1999.

Cui, T., Kumar, Y., Marttinen, P., and Kaski, S. Deconfounded representation similarity for comparison of neural networks. *Advances in Neural Information Processing Systems*, 2022.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., and Kriegeskorte, N. Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neurons, Behavior, Data Analysis, and Theory*, 2021.

Ding, F., Denain, J.-S., and Steinhardt, J. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 2021.

Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling Extremal Events: for Insurance and Finance*. Stochastic Modelling and Applied Probability. 2013.

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

Gröger, F., Wen, S., Le, H., and Brbic, M. With limited data for multimodal alignment, let the STRUCTURE guide you. *Advances in Neural Information Processing Systems*, 2025.

Harvey, S. E., Lipshutz, D., and Williams, A. H. What Representational Similarity Measures Imply about Decodable Information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*. PMLR, 2024.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979.

Hotelling, H. Relations between two sets of variates. In *Breakthroughs in Statistics: Methodology and Distribution*. 1992.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Kapoor, C., Srivastava, S., and Khosla, M. Bridging Critical Gaps in Convergent Learning: How Representational Alignment Evolves Across Layers, Training, and Distribution Shifts. *Advances in Neural Information Processing Systems*, 2025.

Klabunde, M., Schumacher, T., Strohmaier, M., and Lemmerich, F. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 2025a.

Klabunde, M., Wald, T., Schumacher, T., Maier-Hein, K., Strohmaier, M., and Lemmerich, F. ReSi: A Comprehensive Benchmark for Representational Similarity Measures. In *International Conference on Learning Representations*, 2025b.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. 2005.

Livan, G., Novaes, M., and Vivo, P. *Introduction to Random Matrices: Theory and Practice*. SpringerBriefs in Mathematical Physics. 2018.

Maniparambil, M., Akshulakov, R., Djilali, Y. A. D., El Amine Seddik, M., Narayan, S., Mangalam, K., and O'Connor, N. E. Do Vision and Language Encoders Represent the World Similarly? In *Conference on Computer Vision and Pattern Recognition*, 2024.

Marcos-Manchón, P. and Fuentemilla, L. Convergent transformations of visual representation in brains and models. *arXiv preprint arXiv:2507.13941*, 2025.

Morcos, A., Raghu, M., and Bengio, S. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 2018.

Müller, R. R. A random matrix model of communication via antenna arrays. *IEEE Transactions on information theory*, 2002.

Murphy, A., Zylberberg, J., and Fyshe, A. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *arXiv preprint arXiv:2405.01012*, 2024.

Neyshabur, B., Sedghi, H., and Zhang, C. What is being transferred in transfer learning? *Advances in Neural Information Processing Systems*, 2020.

Nguyen, T., Raghu, M., and Kornblith, S. Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth. In *International Conference on Learning Representations*, 2021.

Nichols, T. E. and Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 2002.

Phipson, B. and Smyth, G. K. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics & Molecular Biology*, 2010.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In *Advances in Neural Information Processing Systems*, 2017.

Raugel, J., Szafraniec, M., Vo, H. V., Couprie, C., Labatut, P., Bojanowski, P., Wyart, V., and King, J.-R. Disentangling the factors of convergence between brains and computer vision models. *arXiv preprint arXiv:2508.18226*, 2025.

Robert, P. and Escoufier, Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1976.

Schaeffer, R., Khona, M., Chandra, S., Ostrow, M., Miranda, B., and Koyejo, S. Position: Maximizing Neural Regression Scores May Not Identify Good Models of the Brain. In *NeurIPS 2024 Workshop on Unifying Representations in Neural Models (UniReps)*, 2024.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 2018.

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C., and Van Erk, M. Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics*, 2009.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 2012.

Soni, A., Srivastava, S., Kording, K., and Khosla, M. Conclusions about Neural Network to Brain Alignment are Profoundly Impacted by the Similarity Measure. *bioRxiv*, 2024.

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

Tjandrasuwita, M., Ekbote, C., Ziyin, L., and Liang, P. P. Understanding the Emergence of Multimodal Representation Alignment. In *International Conference on Machine Learning*, 2025.

Tong, Z., Song, Y., Wang, J., and Wang, L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 2022.

Wachter, K. W. The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 1978.

Weenink, D. Canonical correlation analysis. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 2003.

Westfall, P. H. and Young, S. S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. 1993.

Williams, A. H., Kunz, E., Kornblith, S., and Linderman, S. W. Generalized Shape Metrics on Neural Representations. In *Advances in Neural Information Processing Systems*, 2021.

Zhu, T., Han, T., Guibas, L., Pătrăucean, V., and Ovsjanikov, M. Dynamic Reflections: Probing Video Representations with Text Alignment. In *International Conference on Learning Representations*, 2026.

A. Existing calibration approaches for representational similarity metrics

Table 1. Comparison of prior works. **Y**=yes, **N**=no, **P**=partial/indirect. "Debias" indicates an explicit null correction of the reported similarity. "Bounded" indicates whether the corrected score preserves an interpretable upper bound (*e.g.*, 1 for perfect alignment). "Agg-aware" indicates calibration of selection-based aggregates (*e.g.*, max over layer pairs).

Ref	Metric(s)	Debias?	Bounded?	Agg-aware?
Murphy et al. (2024)	CKA	Y	N	N
Chun et al. (2025)	CKA	Y	N	N
Cui et al. (2022)	RSA/CKA	P	N	N
Diedrichsen et al. (2021)	RSA (cv/WUC)	Y	P	N
Cai et al. (2019)	RSA (Bayes)	P	N	N
Smilde et al. (2009)	RV / adj. RV	Y	N	N
Ours	**Any bounded metric**	**Y**	**Y**	**Y**

B. Metrics and score definitions

This appendix gives the definitions of the similarity metrics $s(\mathbf{X}, \mathbf{Y})$ used throughout the paper. The main text focuses on the calibration procedure (Sections 5.1 and 5.2). Here we provide concrete instantiations of the metrics referenced in Section 3 and Section 6.

B.1. Preprocessing and basic notation

Let $\mathbf{X} \in \mathbb{R}^{n \times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$ denote row-aligned representations evaluated on the same n inputs. We use the centering matrix

$$\mathbf{H} \;=\; \mathbf{I}_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top, \tag{18}$$

where $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ is the identity matrix and $\mathbb{1}_n \in \mathbb{R}^n$ is the all-ones vector. We define row-centered representations $\mathbf{X}_c = \mathbf{H}\mathbf{X}$ and $\mathbf{Y}_c = \mathbf{H}\mathbf{Y}$. Unless stated otherwise, similarities are computed on centered representations.

B.2. Raw similarity metrics

This section provides formal definitions of the similarity metrics used throughout the paper. In the main text, we primarily use CKA (linear and RBF kernel) (Kornblith et al., 2019), RSA (Kriegeskorte et al., 2008), and mutual k-NN (Huh et al., 2024) as representative metrics from the spectral, geometric, and neighborhood families, respectively. Additional metrics (SVCCA, PWCCA, cycle-kNN, CKNNA, RV coefficient, Procrustes) are included for completeness and used in supplementary experiments.

B.2.1. Spectral metrics

Linear Centered Kernel Alignment (CKA). Linear CKA (Kornblith et al., 2019) can be written as a normalized Frobenius energy of the *sample cross-covariance* operator. With $\mathbf{X}_c, \mathbf{Y}_c$ as above, define the sample (cross-)covariances

$$\widetilde{\mathbf{\Sigma}}_{XX} := \frac{1}{n-1}\mathbf{X}_c^\top\mathbf{X}_c, \qquad \widetilde{\mathbf{\Sigma}}_{YY} := \frac{1}{n-1}\mathbf{Y}_c^\top\mathbf{Y}_c, \qquad \widetilde{\mathbf{C}} := \widetilde{\mathbf{\Sigma}}_{XY} := \frac{1}{n-1}\mathbf{X}_c^\top\mathbf{Y}_c. \tag{19}$$

The biased linear Hilbert-Schmidt Independence Criterion (HSIC) energy equals $\|\widetilde{\mathbf{C}}\|_F^2$. The commonly used linear CKA normalization can be written as

$$\mathrm{CKA}_{\mathrm{lin}}(\mathbf{X}, \mathbf{Y}) \;=\; \frac{\|\widetilde{\mathbf{C}}\|_F^2}{\|\widetilde{\mathbf{\Sigma}}_{XX}\|_F\,\|\widetilde{\mathbf{\Sigma}}_{YY}\|_F} \;=\; \frac{\|\mathbf{X}_c^\top\mathbf{Y}_c\|_F^2}{\|\mathbf{X}_c^\top\mathbf{X}_c\|_F\,\|\mathbf{Y}_c^\top\mathbf{Y}_c\|_F} \;\in\; [0, 1], \tag{20}$$

where the second equality follows by cancellation of common $1/(n-1)$ factors.
What it measures: linear CKA treats two representations as similar when pairs of points with high inner product in one space also have high inner product in the other, *i.e.*, when their global second-order geometry agrees. It is invariant to rotations and isotropic rescaling, but not to general invertible linear maps.

Kernel Centered Kernel Alignment. Kernel CKA (Kornblith et al., 2019) generalizes linear CKA by replacing dot products with kernel functions. Let $k_X : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$ and $k_Y : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be positive semidefinite kernel functions (*e.g.*, RBF kernel $k_X(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2)$). Let $\mathbf{K}_X \in \mathbb{R}^{n\times n}$ and $\mathbf{K}_Y \in \mathbb{R}^{n\times n}$ be Gram matrices with entries $(\mathbf{K}_X)_{ij} = k_X(\mathbf{x}_i, \mathbf{x}_j)$ and $(\mathbf{K}_Y)_{ij} = k_Y(\mathbf{y}_i, \mathbf{y}_j)$. Let $\widetilde{\mathbf{K}}_X = \mathbf{H}\mathbf{K}_X\mathbf{H}$ and $\widetilde{\mathbf{K}}_Y = \mathbf{H}\mathbf{K}_Y\mathbf{H}$ denote centered Gram matrices. Kernel CKA is defined as:

$$\mathrm{CKA}_{k_X,k_Y}(\mathbf{X}, \mathbf{Y}) \;=\; \frac{\langle \widetilde{\mathbf{K}}_X, \widetilde{\mathbf{K}}_Y \rangle_F}{\|\widetilde{\mathbf{K}}_X\|_F \, \|\widetilde{\mathbf{K}}_Y\|_F}. \tag{21}$$

where $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$. With positive semidefinite kernels and the biased HSIC estimator, the numerator is nonnegative, and kernel CKA typically lies in $[0, 1]$.
What it measures: kernel CKA treats representations as similar when their kernel-induced similarity patterns over points agree. With an RBF kernel of small bandwidth, this emphasizes local, nonlinear similarity structure rather than global linear geometry.

Unbiased Centered Kernel Alignment. The biased HSIC estimator can yield inflated similarity scores at finite sample sizes. Song et al. (2012) derived an unbiased HSIC estimator by recognizing that HSIC can be formulated as a U-statistic. Following Kornblith et al. (2019), we substitute the unbiased estimator into the CKA formula. Let $\widetilde{\mathbf{K}}_X = \mathbf{H}\mathbf{K}_X\mathbf{H}$ be the centered Gram matrix with diagonal set to zero. The unbiased HSIC estimator is:

$$\mathrm{HSIC}_u(\mathbf{K}_X, \mathbf{K}_Y) = \frac{1}{n(n-3)} \left(\mathrm{tr}(\widetilde{\mathbf{K}}_X\widetilde{\mathbf{K}}_Y) + \frac{\mathbf{1}^\top\widetilde{\mathbf{K}}_X\mathbf{1}\cdot\mathbf{1}^\top\widetilde{\mathbf{K}}_Y\mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2}\mathbf{1}^\top\widetilde{\mathbf{K}}_X\widetilde{\mathbf{K}}_Y\mathbf{1} \right). \tag{22}$$

Unbiased CKA replaces both numerator and denominator of Equation (21) with this estimator. Unlike the biased version, unbiased CKA can take small negative values at finite n.
What it measures: the same agreement of similarity patterns as kernel CKA, but with the finite-sample inflation of HSIC removed, so a value near zero indicates no more agreement than expected by chance.

Canonical Correlation Analysis (CCA)-based similarity. CCA (Weenink, 2003) measures linear subspace alignment. The sample canonical correlations $\{\rho_i\}_{i=1}^r$ (with $r = \mathrm{rank}(\widetilde{\mathbf{\Sigma}}_{XY})$) are the singular values of the whitened cross-covariance operator

$$\widetilde{\mathbf{T}}_{\mathrm{CCA}} \;=\; \widetilde{\mathbf{\Sigma}}_{XX}^{-\frac{1}{2}} \, \widetilde{\mathbf{\Sigma}}_{XY} \, \widetilde{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}}. \tag{23}$$

Common scalar summaries include the mean canonical correlation $\frac{1}{r}\sum_{i=1}^r \rho_i$ or a weighted average as used in SVCCA (Raghu et al., 2017) and PWCCA (Morcos et al., 2018).
What it measures: CCA treats two representations as similar when one can be mapped onto the other by an invertible linear transformation, *i.e.*, when they span the same linear subspace. It ignores rotations, scaling, and other invertible linear changes within each space.

Singular Vector Canonical Correlation Analysis (SVCCA). SVCCA (Raghu et al., 2017) combines dimensionality reduction via singular value decomposition (SVD) with CCA. First, truncated SVD is applied to each representation to retain the top principal components, yielding $\mathbf{X}' \in \mathbb{R}^{n\times p}$ and $\mathbf{Y}' \in \mathbb{R}^{n\times q}$. Then CCA is applied to the reduced representations, yielding canonical correlations $\{\rho_i\}_{i=1}^r$. The SVCCA similarity is the mean canonical correlation:

$$\mathrm{SVCCA}(\mathbf{X}, \mathbf{Y}) = \frac{1}{r}\sum_{i=1}^{r} \rho_i. \tag{24}$$

What it measures: SVCCA treats representations as similar when their high-variance subspaces are linearly aligned, discarding low-variance (noise) directions before measuring linear correspondence.

Projection Weighted Canonical Correlation Analysis (PWCCA). PWCCA (Morcos et al., 2018) improves upon SVCCA by weighting canonical correlations according to their importance in explaining the original representations. Let $\mathbf{h}_i^X$ and $\mathbf{h}_i^Y$ denote the i-th canonical variables (projections onto canonical directions). The weight for the i-th canonical

correlation is proportional to how much variance it explains:

$$\alpha_i = \sum_{j=1}^{d_x} |\langle \mathbf{h}_i^X, \mathbf{X}_{:,j} \rangle|, \tag{25}$$

where $\mathbf{X}_{:,j}$ is the j-th column of $\mathbf{X}$. The PWCCA similarity is the weighted mean:

$$\text{PWCCA}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{r} \alpha_i \rho_i}{\sum_{i=1}^{r} \alpha_i}. \tag{26}$$

This weighting ensures that canonical correlations corresponding to principal directions receive higher weight than those corresponding to noise dimensions.
What it measures: like CCA, PWCCA treats representations as similar when they linearly correspond, but it counts agreement along high-variance directions more, so two representations are similar when their dominant directions align.

RV coefficient. The RV ("Relation between two sets of Variables") coefficient (Robert & Escoufier, 1976; Smilde et al., 2009) is a multivariate generalization of the squared Pearson correlation. It measures the similarity between two configuration matrices via their inner-product (Gram) matrices. Let $\mathbf{W}_X = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{W}_Y = \mathbf{Y}\mathbf{Y}^\top$ be the sample inner-product matrices. The RV coefficient is:

$$\text{RV}(\mathbf{X}, \mathbf{Y}) = \frac{\text{tr}(\mathbf{W}_X \mathbf{W}_Y)}{\sqrt{\text{tr}(\mathbf{W}_X^2)\,\text{tr}(\mathbf{W}_Y^2)}} \in [0, 1]. \tag{27}$$

What it measures: the RV coefficient treats representations as similar when their point-by-point inner-product (Gram) matrices match. Like linear CKA, it captures global second-order geometry and is invariant to rotation but sensitive to scaling.

B.2.2. Geometric metrics

Representational Similarity Analysis (RSA) via Spearman correlation of dissimilarity matrices. RSA (Kriegeskorte et al., 2008) compares the geometry induced by pairwise dissimilarities. Let $\delta(\cdot, \cdot)$ be a dissimilarity on representation vectors (*e.g.*, correlation distance $\delta(\mathbf{u}, \mathbf{v}) = 1 - \text{corr}(\mathbf{u}, \mathbf{v})$, cosine distance). Define Representational Dissimilarity Matrices (RDMs)

$$(\mathbf{D}_X)_{ij} = \delta(\mathbf{x}_i, \mathbf{x}_j), \qquad (\mathbf{D}_Y)_{ij} = \delta(\mathbf{y}_i, \mathbf{y}_j), \tag{28}$$

and let $\text{vec}_\triangle(\mathbf{D}) \in \mathbb{R}^{n(n-1)/2}$ denote vectorization of the strict upper triangle. RSA is then computed as a rank correlation between the two RDM vectors:

$$\text{RSA}(\mathbf{X}, \mathbf{Y}) \;=\; \rho_S(\text{vec}_\triangle(\mathbf{D}_X),\; \text{vec}_\triangle(\mathbf{D}_Y))\,, \tag{29}$$

where Spearman's ρ can be expressed as Pearson correlation of ranks,

$$\rho_S(\mathbf{u}, \mathbf{v}) \;=\; \text{corr}(\text{rank}(\mathbf{u}),\; \text{rank}(\mathbf{v}))\,. \tag{30}$$

What it measures: RSA treats two representations as similar when they rank pairs of points in the same order of (dis)similarity. It compares the relational geometry of distances rather than the coordinates, and depends only on the pattern of pairwise dissimilarities.

Procrustes distance. The orthogonal Procrustes distance (Williams et al., 2021) measures the minimal Euclidean distance between two representations after optimal orthogonal alignment. Assuming $d_x = d_y = d$, the optimal orthogonal matrix $\mathbf{Q}^* \in \mathcal{O}(d)$ is:

$$\mathbf{Q}^* = \underset{\mathbf{Q} \in \mathcal{O}(d)}{\text{argmin}} \|\mathbf{X} - \mathbf{Y}\mathbf{Q}\|_F^2, \tag{31}$$

which has the closed-form solution $\mathbf{Q}^* = \mathbf{V}\mathbf{U}^\top$ where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{X}^\top\mathbf{Y}$ is the SVD. The Procrustes distance is:

$$d_{\text{Proc}}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\mathbf{Q}^*\|_F. \tag{32}$$

To convert to a similarity in $[0, 1]$, one can use $1 - d_{\text{Proc}}^2 / (\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2)$ after appropriate normalization.
What it measures: Procrustes treats two representations as similar when one can be rotated and reflected onto the other with small residual. It compares absolute geometry up to rigid motion and, unlike neighborhood metrics, is sensitive to the actual pairwise distances.

B.2.3. NEIGHBORHOOD METRICS

Mutual k-Nearest Neighbors (mKNN). mKNN (Huh et al., 2024) focuses on local topology. For each anchor sample i, define the set of its k nearest neighbors according to a distance measure $\mathrm{dist}(\cdot,\cdot)$ in $\mathbf{X}$ and $\mathbf{Y}$,

$$N_{\mathbf{X}}(i) = \mathrm{KNN}_k(i;\mathbf{X}), \qquad N_{\mathbf{Y}}(i) = \mathrm{KNN}_k(i;\mathbf{Y}), \tag{33}$$

where $\mathrm{KNN}_k(i;\mathbf{X})$ denotes the indices of the k samples (excluding i) that minimize $\mathrm{dist}(\mathbf{x}_i,\mathbf{x}_j)$. mKNN is then defined as the average fraction of shared neighbors:

$$\mathrm{mKNN}_k(\mathbf{X},\mathbf{Y}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{|N_{\mathbf{X}}(i)\cap N_{\mathbf{Y}}(i)|}{k} \;\in\; [0,1]. \tag{34}$$

What it measures: mKNN treats two representations as similar when each point keeps the same set of nearest neighbors in both spaces. It captures local topology, depends only on the rank order of distances, and is therefore invariant to any transformation that preserves k-nearest-neighbor sets, ignoring exact distances and global geometry.

Cycle-kNN (bidirectional k-NN). While mKNN measures one-directional neighborhood overlap, cycle-kNN enforces bidirectional consistency (Huh et al., 2024). A pair (i,j) forms a *cycle* if $j \in N_{\mathbf{X}}(i)$ and $i \in N_{\mathbf{X}}(j)$ (mutual neighbors in $\mathbf{X}$), and similarly for $\mathbf{Y}$. Define the set of bidirectional neighbors:

$$C_{\mathbf{X}}(i) = \{j : j \in N_{\mathbf{X}}(i) \text{ and } i \in N_{\mathbf{X}}(j)\}. \tag{35}$$

Cycle-kNN measures the overlap of these symmetric neighborhoods:

$$\text{cycle-kNN}_k(\mathbf{X},\mathbf{Y}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{|C_{\mathbf{X}}(i)\cap C_{\mathbf{Y}}(i)|}{\max(|C_{\mathbf{X}}(i)|,1)} \;\in\; [0,1]. \tag{36}$$

This metric is stricter than mKNN, requiring that shared neighbors be mutually recognized in both representation spaces. *What it measures:* cycle-kNN treats points as similar only when they are *mutual* nearest neighbors in both spaces, a stricter local-topology criterion than mKNN.

CKA with Neighborhood Alignment (CKNNA). CKNNA (Huh et al., 2024) combines the kernel-based formulation of CKA with local neighborhood structure. Instead of computing CKA on full Gram matrices, CKNNA restricts interaction to k-nearest neighbor graphs. Let $\mathbf{A}_X \in \{0,1\}^{n\times n}$ be the adjacency matrix of the k-NN graph on $\mathbf{X}$, with $(\mathbf{A}_X)_{ij} = 1$ if $j \in N_{\mathbf{X}}(i)$ or $i \in N_{\mathbf{X}}(j)$. CKNNA applies the CKA formula (Equation (21)) to the centered adjacency matrices:

$$\mathrm{CKNNA}_k(\mathbf{X},\mathbf{Y}) \;=\; \frac{\langle \mathbf{H}\mathbf{A}_X\mathbf{H}, \mathbf{H}\mathbf{A}_Y\mathbf{H}\rangle_F}{\|\mathbf{H}\mathbf{A}_X\mathbf{H}\|_F\,\|\mathbf{H}\mathbf{A}_Y\mathbf{H}\|_F}. \tag{37}$$

What it measures: CKNNA treats representations as similar when their k-nearest-neighbor graphs agree, combining CKA's normalization with a purely local notion of neighborhood structure.

C. Theoretical Derivations

In this section, we provide the theoretical justification for the confounding factors identified in Section 4.

C.1. Permutation validity, super-uniformity, and gating

This section formalizes the finite-sample validity of permutation calibration.

Definition C.1 (Super-uniformity)**.** A p-value p is *super-uniform* under H_0 if for all $t \in [0,1]$,

$$\mathbb{P}_{H_0}(p \le t) \le t. \tag{38}$$

Equivalently, p-values under H_0 are stochastically larger than $\mathrm{Unif}(0,1)$, which is sufficient for valid Type-I error control.

Lemma C.2 (Permutation p-values are super-uniform)**.** *Under Assumption 3.1, the permutation p-value in Equation (10) satisfies super-uniformity: $\mathbb{P}_{H_0}(p \le \alpha) \le \alpha$ for all $\alpha \in [0,1]$ (finite-sample validity).*

Proof of Lemma C.2. Let $s_{\text{obs}} = s(\mathbf{X}, \mathbf{Y})$ be the observed statistic and let $s^{(k)} = s(\mathbf{X}, \pi_k(\mathbf{Y}))$ for $k = 1, \dots, K$ be the statistics computed on permuted pairings. Under Assumption 3.1, the vector $(s_{\text{obs}}, s^{(1)}, \dots, s^{(K)})$ is *exchangeable*: its joint distribution is invariant to permutations of the indices. Consider the (upper) rank

$$R \;=\; 1 + \#\{k \in \{1, \dots, K\} : s^{(k)} \geq s_{\text{obs}}\} \;\in\; \{1, \dots, K+1\}. \tag{39}$$

If the scores are almost surely distinct, exchangeability implies that the rank of s_{obs} among $\{s_{\text{obs}}, s^{(1)}, \dots, s^{(K)}\}$ is uniform on $\{1, \dots, K+1\}$. With possible ties, the add-one p-value of Phipson & Smyth (2010),

$$p = \frac{R}{K+1}, \tag{40}$$

is conservative, implying $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$. □

Proof of Corollary 5.1. Let $s_{\text{obs}} = s(\mathbf{X}, \mathbf{Y})$ and $s^{(k)} = s(\mathbf{X}, \pi_k(\mathbf{Y}))$ for $k = 1, \dots, K$. Under Assumption 3.1, the vector $(s_{\text{obs}}, s^{(1)}, \dots, s^{(K)})$ is exchangeable. Let

$$\tau_\alpha := s_{(\lceil (1-\alpha)(K+1) \rceil)}$$

be the $(1-\alpha)$-quantile defined via the order statistic of the *combined* multiset $\{s_{\text{obs}}, s^{(1)}, \dots, s^{(K)}\}$. Define the (upper) rank

$$R \;=\; 1 + \#\{k \in \{1, \dots, K\} : s^{(k)} \geq s_{\text{obs}}\} \;\in\; \{1, \dots, K+1\},$$

and the corresponding add-one p-value $p = R/(K+1)$. By construction of τ_α, the rejection event $\{s_{\text{obs}} > \tau_\alpha\}$ implies that s_{obs} lies among the largest $\lfloor \alpha(K+1) \rfloor$ values of $\{s_{\text{obs}}, s^{(1)}, \dots, s^{(K)}\}$, hence $R \leq \alpha(K+1)$ and therefore $p \leq \alpha$. By Lemma C.2, $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$, which yields

$$\mathbb{P}_{H_0}(s_{\text{obs}} > \tau_\alpha) \leq \mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha.$$

□

Restricted permutations under dependence. Assumption 3.1 treats the n row pairs as exchangeable units. In practice, exchangeability can be violated even without sequential structure (*e.g.*, grouped or clustered samples). Validity is then recovered by using *restricted* permutations that preserve the dependence structure (*e.g.*, permuting within blocks or permuting block labels) and re-running under each restricted permutation.

C.2. Monotone invariance of rank-based calibration

The following proposition is a standard result in randomization inference; we state it here for completeness and to clarify its role in justifying the calibrated score design.

Proposition C.3 (Monotone invariance of rank-based calibration (Lehmann & Romano, 2005))**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing. Define p_g by applying Equation (10) to the transformed statistic $g \circ s$ using the* same *permutations. Then $p_g = p$, and likewise the null percentile (the rank of s_{obs} among the combined set) is invariant under g.*

Proof. Let g be strictly increasing. For any two real numbers a, b, we have $a \geq b$ if and only if $g(a) \geq g(b)$. Therefore, for each permutation draw k,

$$\mathbb{1}\{s^{(k)} \geq s_{\text{obs}}\} = \mathbb{1}\{g(s^{(k)}) \geq g(s_{\text{obs}})\}. \tag{41}$$

Summing over k shows that the permutation rank R (and thus the add-one p-value) is unchanged by applying g to both the observed and permuted statistics. The same argument applies to the null percentile, since the ordering of samples is preserved under g. □

C.3. Post-selection inflation and aggregation-aware validity

Proposition C.4 (Validity for aggregation-aware calibration)**.** *Let T be any measurable aggregation operator applied to a layer-wise similarity matrix $\mathbf{S}$ (e.g., max, row-max, top-k). If $T_{\text{obs}} = T(\mathbf{S})$ is calibrated against the permutation null $\{T(\mathbf{S}^{(k)})\}_{k=1}^{K}$ as in Equation (16), then the resulting p_{agg} is super-uniform under H_0.*

Proof of Proposition C.4. Let T be any measurable functional of the full data (representations across all layers), producing the scalar report T_{obs}. Under Assumption 3.1 and consistent layer-wise permutation of sample correspondences, the vector $(T_{\text{obs}}, T^{(1)}, \ldots, T^{(K)})$ is exchangeable. Applying the same rank argument as in Appendix C.1 yields super-uniformity for the add-one p-value in Equation (16). □

C.4. The width confounder

This appendix provides concrete calculations that justify the width confounder using Random Matrix Theory (RMT): even under independence, interaction operators have non-trivial magnitude and spectrum when d is not negligible relative to n.

Proof of Proposition 4.1. Let $\mathbf{X} \in \mathbb{R}^{n \times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$ have i.i.d. rows with mean 0, identity covariance, and $\mathbf{x}_i$ and $\mathbf{y}_i$ independent. Let $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$ be the centering matrix, so $\mathbf{X}_c = \mathbf{H}\mathbf{X}$ and $\mathbf{Y}_c = \mathbf{H}\mathbf{Y}$. Since $\mathbf{H}$ is symmetric and idempotent ($\mathbf{H}^2 = \mathbf{H}$), the sample cross-covariance is

$$\widetilde{\mathbf{C}} = \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{Y}_c = \frac{1}{n-1}\mathbf{X}^\top \mathbf{H}\mathbf{Y}. \tag{42}$$

Denote entry (a, b) as $\widetilde{C}_{ab}$. Expanding via $H_{ij} = \delta_{ij} - \frac{1}{n}$:

$$\widetilde{C}_{ab} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_{ia}Y_{ib} - \frac{1}{n}\Big(\sum_{i=1}^{n} X_{ia}\Big)\Big(\sum_{j=1}^{n} Y_{jb}\Big)\right). \tag{43}$$

We compute $\mathbb{E}[\widetilde{C}_{ab}^2]$ using independence of $\mathbf{x}_i$ and $\mathbf{y}_j$ for all i, j, zero means, and identity covariance.

Term 1: $\mathbb{E}\big[(\sum_i X_{ia}Y_{ib})^2\big] = \sum_{i,j}\mathbb{E}[X_{ia}X_{ja}]\mathbb{E}[Y_{ib}Y_{jb}]$. For $i \neq j$, independence across rows and zero mean give $\mathbb{E}[X_{ia}X_{ja}] = \mathbb{E}[X_{ia}]\mathbb{E}[X_{ja}] = 0$. For $i = j$, we have $\mathbb{E}[X_{ia}^2]\mathbb{E}[Y_{ib}^2] = 1$. Thus $\mathbb{E}\big[(\sum_i X_{ia}Y_{ib})^2\big] = n$.

Term 2: $\mathbb{E}\big[(\sum_i X_{ia}Y_{ib})(\sum_j X_{ja})(\sum_k Y_{kb})\big] = \sum_{i,j,k}\mathbb{E}[X_{ia}X_{ja}]\mathbb{E}[Y_{ib}Y_{kb}]$. This is nonzero only when $i = j$ and $i = k$, yielding $\sum_i 1 \cdot 1 = n$.

Term 3: $\mathbb{E}\big[(\sum_i X_{ia})^2(\sum_j Y_{jb})^2\big] = \mathbb{E}[(\sum_i X_{ia})^2]\mathbb{E}[(\sum_j Y_{jb})^2] = n \cdot n = n^2$.

Combining:

$$\mathbb{E}\left[\widetilde{C}_{ab}^2\right] = \frac{1}{(n-1)^2}\left(n - \frac{2}{n}\cdot n + \frac{n^2}{n^2}\right) = \frac{1}{(n-1)^2}(n - 2 + 1) = \frac{1}{n-1}. \tag{44}$$

Summing over all entries:

$$\mathbb{E}\left[\|\widetilde{\mathbf{C}}\|_F^2\right] = \sum_{a=1}^{d_x}\sum_{b=1}^{d_y}\mathbb{E}[\widetilde{C}_{ab}^2] = \frac{d_x d_y}{n-1}. \tag{45}$$

□

Interpretation. The null interaction energy is $\mathcal{O}(d_x d_y/n)$. In the common regime $d_x, d_y \asymp n$, the null energy is $\mathcal{O}(n)$ and therefore *does not vanish*. Since many spectral similarity metrics aggregate singular values (*e.g.*, via $\|\widetilde{\mathbf{C}}\|_F^2 = \sum_i \sigma_i^2(\widetilde{\mathbf{C}})$), this already explains a positive baseline under H_0 and its dependence on (n, d_x, d_y).

Width inflation persists under genuine signal (H_1). The width confounder is not specific to the null hypothesis. Even when two representations share a fixed, genuine signal, finite-sample CKA is inflated by width, because increasing d grows the diagonal and off-diagonal Gram entries at different rates and the CKA denominator does not cancel the resulting width-dependent self-similarity.

Proposition C.5 (Width inflation under genuine signal)**.** *Fix $n \geq 2$ and a target alignment $\rho \in (0,1)$. For each width d, draw $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ independently across $i = 1, \ldots, n$, and set*

$$\mathbf{x}_i = \sqrt{\rho}\,\mathbf{u}_i + \sqrt{1-\rho}\,\mathbf{v}_i, \qquad \mathbf{y}_i = \sqrt{\rho}\,\mathbf{u}_i + \sqrt{1-\rho}\,\mathbf{w}_i, \tag{46}$$

so that $\mathbf{x}_i$ and $\mathbf{y}_i$ share the latent signal $\mathbf{u}_i$. Then the population linear CKA equals $\rho^2 < 1$, whereas for fixed *n the sample centered linear CKA satisfies $\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) \to 1$ almost surely as $d \to \infty$. Hence the finite-sample estimate is inflated by width, with $\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) - \rho^2 \to 1 - \rho^2 > 0$.*

Proof. By independence of $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i$, the marginal covariances are $\Sigma_x = \Sigma_y = \mathbf{I}_d$ and the cross-covariance is $\Sigma_{xy} = \mathrm{Cov}(\mathbf{x}_i, \mathbf{y}_i) = \rho\,\mathbf{I}_d$. The population linear CKA is therefore

$$\mathrm{CKA}_{\mathrm{pop}} = \frac{\|\Sigma_{xy}\|_F^2}{\|\Sigma_x\|_F\,\|\Sigma_y\|_F} = \frac{\rho^2 d}{\sqrt{d}\,\sqrt{d}} = \rho^2. \tag{47}$$

For the sample statistic at fixed n, write the columns of $\mathbf{X}$ as $\mathbf{g}_1, \ldots, \mathbf{g}_d \in \mathbb{R}^n$. Each entry of $\mathbf{X}$ is $\mathcal{N}(0,1)$ (since $\rho + (1-\rho) = 1$), and the columns are i.i.d. across j with $\mathbf{g}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. By the strong law of large numbers,

$$\frac{1}{d}\mathbf{X}\mathbf{X}^\top = \frac{1}{d}\sum_{j=1}^{d} \mathbf{g}_j \mathbf{g}_j^\top \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbf{g}_1 \mathbf{g}_1^\top] = \mathbf{I}_n. \tag{48}$$

Since n is fixed, this convergence also holds in Frobenius norm, so $\frac{1}{d}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} \to \mathbf{H}$ a.s.; the identical argument gives $\frac{1}{d}\mathbf{H}\mathbf{Y}\mathbf{Y}^\top\mathbf{H} \to \mathbf{H}$. By scale invariance of CKA,

$$\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\langle \frac{1}{d}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H},\ \frac{1}{d}\mathbf{H}\mathbf{Y}\mathbf{Y}^\top\mathbf{H}\rangle_F}{\|\frac{1}{d}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H}\|_F\,\|\frac{1}{d}\mathbf{H}\mathbf{Y}\mathbf{Y}^\top\mathbf{H}\|_F} \xrightarrow{\text{a.s.}} \frac{\langle \mathbf{H}, \mathbf{H}\rangle_F}{\|\mathbf{H}\|_F^2} = 1. \tag{49}$$

Subtracting the population value gives $\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) - \rho^2 \to 1 - \rho^2 > 0$. □

Thus, at any fixed sample size n, increasing the width drives the estimated CKA toward 1 regardless of the true alignment ρ^2, inflating it by up to $1 - \rho^2$. This width-driven gap is distinct from the $\mathcal{O}(1/n)$ sample-size effect that vanishes as n grows, and it is exactly what permutation calibration removes. Appendix E.5 verifies this on real pretrained networks with nonzero true similarity.

Why we use permutation rather than closed forms. Closed-form bulk edges are ensemble- and normalization-specific and are brittle to the preprocessing used in practice (*e.g.*, centering, whitening, kernelization). Moreover, finite-n corrections can be non-negligible. We therefore estimate the relevant right-tail behavior nonparametrically via permutation. This yields a conservative, implementation-faithful estimate of chance fluctuations without relying on fragile analytical formulas.

C.5. The depth confounder

Here we formalize why selection-based summaries (*e.g.*, maximum similarity over layer pairs) inflate with the size of the search space using Extreme Value Theory (EVT).

Let $\mathcal{S} = \{S_{\ell,\ell'} : 1 \leq \ell \leq L_A,\ 1 \leq \ell' \leq L_B\}$ denote the collection of null similarity fluctuations under H_0, and let $M = L_A L_B$.

Assumption C.6 (Uniform sub-Gaussian right tails and integrability)**.** There exist $\mu \in \mathbb{R}$ and $\sigma > 0$ such that for all (ℓ, ℓ') and all $t \geq 0$,

$$\mathbb{P}(S_{\ell,\ell'} - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{50}$$

Moreover, each $S_{\ell,\ell'}$ is integrable: $\mathbb{E}|S_{\ell,\ell'}| < \infty$ for all (ℓ, ℓ').

Proposition C.7 (Maximal inequality, no independence required)**.** *Under Assumption C.6 and for $M \geq 2$,*

$$\mathbb{E}\Big[\max_{\ell,\ell'} S_{\ell,\ell'}\Big] \ \leq\ \mu + C\,\sigma\sqrt{\log M}, \tag{51}$$

where $C > 0$ is a constant (e.g., one can take $C = 3$).

Proof. Let $Z := \max_{\ell,\ell'} S_{\ell,\ell'} - \mu$. Since $M < \infty$ and $\mathbb{E}|S_{\ell,\ell'}| < \infty$ for all (ℓ, ℓ'), we have

$$\mathbb{E}|Z| \le \mathbb{E}\Big[\max_{\ell,\ell'} |S_{\ell,\ell'}|\Big] + |\mu| \le \sum_{\ell,\ell'} \mathbb{E}|S_{\ell,\ell'}| + |\mu| < \infty, \tag{52}$$

so Z is integrable, and the tail-integration formula applies. By the union bound and Assumption C.6,

$$\mathbb{P}(Z \ge t) \le M \exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad \text{for all } t \ge 0. \tag{53}$$

Using the tail-integration formula for an integrable real-valued random variable Z,

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge t)\, dt - \int_0^\infty \mathbb{P}(Z \le -t)\, dt \le \int_0^\infty \mathbb{P}(Z \ge t)\, dt, \tag{54}$$

and the bound $\mathbb{P}(Z \ge t) \le 1$, we obtain

$$\mathbb{E}[Z] \;\le\; \int_0^\infty \min\left\{1,\; M \exp\left(-\frac{t^2}{2\sigma^2}\right)\right\} dt. \tag{55}$$

Let $t_0 = \sigma\sqrt{2\log M}$. This value of t_0 is the solution of $M \exp\big(-t_0^2/2\sigma^2\big) = 1$, *i.e.*, the crossover where the bound $\min\{1, \cdot\}$ switches. Splitting the integral at t_0 yields

$$\mathbb{E}[Z] \;\le\; t_0 \;+\; M \int_{t_0}^\infty \exp\left(-\frac{t^2}{2\sigma^2}\right) dt. \tag{56}$$

Applying the standard Gaussian tail bound $\int_{t_0}^\infty e^{-t^2/(2\sigma^2)} dt \le (\sigma^2/t_0) e^{-t_0^2/(2\sigma^2)}$ gives

$$\mathbb{E}[Z] \;\le\; \sigma\sqrt{2\log M} \;+\; \frac{\sigma}{\sqrt{2\log M}}. \tag{57}$$

For $M \ge 2$, the right-hand side is at most $3\sigma\sqrt{\log M}$, proving the claim with $C = 3$. □

Remark. When the $S_{\ell,\ell'}$ are i.i.d. (or weakly dependent), classical Extreme Value Theory yields sharper asymptotics. For example, if $S_{\ell,\ell'} \sim \mathcal{N}(\mu_0, \sigma_0^2)$ i.i.d., the centered maximum converges to a Gumbel distribution and

$$\mathbb{E}[T_{\max}] \approx \mu_0 + \sigma_0 \left(\sqrt{2\ln M} - \frac{\ln\ln M + \ln 4\pi}{2\sqrt{2\ln M}}\right), \tag{58}$$

as stated in standard references (Cramér, 1999; Embrechts et al., 2013). Real layer-wise similarities are dependent, so the approximation above should be treated as heuristic; Proposition C.7 provides a dependence-robust upper bound.

C.6. Null Baselines for Neighborhood Metrics

The preceding analysis focused on spectral metrics whose null baselines scale with d/n. Neighborhood-based metrics such as mutual k-NN follow a fundamentally different regime, which we now characterize.

Definition C.8 (Mutual k-NN overlap)**.** For representations $\mathbf{X} \in \mathbb{R}^{n \times d_x}$, $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$ and neighborhood size $k < n$, let $N_{\mathbf{X}}(i) \subseteq \{1, \dots, n\} \setminus \{i\}$ denote the indices of the k nearest neighbors of sample i in $\mathbf{X}$ (*e.g.*, Euclidean or cosine), and similarly for $N_{\mathbf{Y}}(i)$. The mutual k-NN overlap is

$$\mathrm{mKNN}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n \frac{|N_{\mathbf{X}}(i) \cap N_{\mathbf{Y}}(i)|}{k}. \tag{59}$$

Proposition C.9 (Uniformity of k-NN index sets under i.i.d. sampling)**.** *Fix an anchor index $i \in \{1, \dots, n\}$. Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ be i.i.d. and define the k-NN set $N_{\mathbf{X}}(i) \subseteq \{1, \dots, n\} \setminus \{i\}$ using a fixed distance $\mathrm{dist}(\cdot, \cdot)$. Assume either (i) $\{\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)\}_{j \ne i}$ are almost surely distinct, or (ii) ties are broken by selecting a uniformly random k-subset among the set of minimizers. Then $N_{\mathbf{X}}(i)$ is uniformly distributed over the $\binom{n-1}{k}$ k-subsets of $\{1, \dots, n\} \setminus \{i\}$.*

Proof. Let $\mathcal{I} := \{1, \dots, n\} \setminus \{i\}$ be the candidate-neighbor index set. For any permutation π of $\mathcal{I}$, i.i.d. sampling implies

$$(\mathbf{x}_j)_{j\in\mathcal{I}} \overset{d}{=} (\mathbf{x}_{\pi(j)})_{j\in\mathcal{I}}.$$

The k-NN selection rule depends on the candidate points only through their distances to $\mathbf{x}_i$, so permuting the candidate indices permutes the resulting neighbor set. Under either the no-ties assumption or the stated uniform tie-break rule, for any two k-subsets $S, S' \subseteq \mathcal{I}$ there exists a permutation π with $\pi(S) = S'$ and hence

$$\mathbb{P}\big(N_{\mathbf{X}}(i) = S\big) = \mathbb{P}\big(N_{\mathbf{X}}(i) = S'\big).$$

Since the events $\{N_{\mathbf{X}}(i) = S\}$ over all $|S| = k$ partition the sample space, each has probability $\binom{n-1}{k}^{-1}$. □

Theorem C.10 (Null baseline for mutual k-NN)**.** *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times d}$ have i.i.d. rows, with $\mathbf{X}$ independent of $\mathbf{Y}$. Define $N_{\mathbf{X}}(i)$ and $N_{\mathbf{Y}}(i)$ as in Definition C.8, using either almost sure absence of distance ties or uniform random tie-breaking. Then*

$$\mathbb{E}_{H_0}\Big[\mathrm{mKNN}(\mathbf{X}, \mathbf{Y})\Big] \;=\; \frac{k}{n-1}.$$

Proof. Fix an anchor i. By Proposition C.9, $N_{\mathbf{X}}(i)$ and $N_{\mathbf{Y}}(i)$ are each uniform random k-subsets of the $(n-1)$-element set $\{1, \dots, n\} \setminus \{i\}$. Moreover, since $\mathbf{X}$ and $\mathbf{Y}$ are independent and $N_{\mathbf{X}}(i)$ (resp. $N_{\mathbf{Y}}(i)$) is a measurable function of $\mathbf{X}$ (resp. $\mathbf{Y}$), the sets $N_{\mathbf{X}}(i)$ and $N_{\mathbf{Y}}(i)$ are independent.

Therefore $|N_{\mathbf{X}}(i) \cap N_{\mathbf{Y}}(i)|$ has a hypergeometric distribution with population size $n-1$, number of "successes" k, and draws k, giving

$$\mathbb{E}_{H_0}\Big[|N_{\mathbf{X}}(i) \cap N_{\mathbf{Y}}(i)|\Big] = \frac{k^2}{n-1}.$$

Substituting into the definition of mKNN,

$$\mathbb{E}_{H_0}\left[\mathrm{mKNN}(\mathbf{X}, \mathbf{Y})\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{H_0}\left[\frac{|N_{\mathbf{X}}(i) \cap N_{\mathbf{Y}}(i)|}{k}\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{k}{n-1} = \frac{k}{n-1}.$$

□

Proposition C.11 (Per-anchor variance and generic bounds for mKNN under the null)**.** *Under the assumptions of Theorem C.10, for each anchor i the intersection size $H_i := |N_{\mathbf{X}}(i) \cap N_{\mathbf{Y}}(i)|$ is hypergeometric with mean $k^2/(n-1)$ and variance*

$$\mathrm{Var}[H_i] \;=\; \frac{k^2(n-1-k)^2}{(n-1)^2(n-2)}.$$

Moreover, since $\mathrm{mKNN}(\mathbf{X}, \mathbf{Y}) \in [0, 1]$ deterministically, we have the fully general bound

$$\mathrm{Var}[\mathrm{mKNN}(\mathbf{X}, \mathbf{Y})] \le \frac{1}{4}.$$

If one additionally assumes *that the per-anchor terms $\{|N_{\mathbf{X}}(i) \cap N_{\mathbf{Y}}(i)|/k\}_{i=1}^{n}$ are independent (this is a modeling assumption, not a consequence of H_0), then* $\mathrm{Var}[\mathrm{mKNN}(\mathbf{X}, \mathbf{Y})] = \mathcal{O}(1/n)$.

Proof. The hypergeometric variance formula gives

$$\mathrm{Var}[H_i] = k \cdot \frac{k}{n-1}\left(1 - \frac{k}{n-1}\right) \cdot \frac{(n-1)-k}{(n-1)-1} = \frac{k^2(n-1-k)^2}{(n-1)^2(n-2)}.$$

The bound $\mathrm{Var}[\mathrm{mKNN}] \le 1/4$ follows from $\mathrm{mKNN} \in [0, 1]$. Under the stated additional independence assumption across anchors,

$$\mathrm{Var}\Big[\mathrm{mKNN}(\mathbf{X}, \mathbf{Y})\Big] = \frac{1}{n}\,\mathrm{Var}\left(\frac{H_1}{k}\right) = \frac{1}{nk^2}\,\mathrm{Var}[H_1],$$

which is $\mathcal{O}(1/n)$ for fixed k. □

D. Implementation

A key advantage of null calibration is its simplicity: the framework can be applied to *any* similarity metric with minimal code changes. This section provides pseudocode for the two main calibration procedures described in the paper.

Scalar null calibration. Algorithm 1 shows the complete procedure for calibrating a single similarity comparison. The only requirement is a function `similarity(X,Y)` that computes the raw metric. The algorithm returns both a permutation p-value and a calibrated score with a principled zero point.

Algorithm 1 Scalar Null Calibration

Require: Representations $\mathbf{X} \in \mathbb{R}^{n \times d_x}$, $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$
Require: Similarity function $\texttt{sim}(\cdot, \cdot)$, permutations K, significance level α
Ensure: Calibrated score s_{cal}, p-value p

$s_{\text{obs}} \leftarrow \texttt{sim}(\mathbf{X}, \mathbf{Y})$ {Observed similarity}
$\texttt{null_scores} \leftarrow []$
for $k = 1$ to K **do**
  $\pi \leftarrow \texttt{random_permutation}(n)$ {Permute sample indices}
  $\mathbf{Y}_\pi \leftarrow \mathbf{Y}[\pi, :]$ {Permute rows of $\mathbf{Y}$}
  $\texttt{null_scores}[k] \leftarrow \texttt{sim}(\mathbf{X}, \mathbf{Y}_\pi)$
end for
$\texttt{combined} \leftarrow [s_{\text{obs}}] \cup \texttt{null_scores}$ {Combined set}
$\tau_\alpha \leftarrow \texttt{quantile}(\texttt{combined}, 1 - \alpha)$ {Critical threshold from combined set}
$p \leftarrow \frac{1 + \sum_{k=1}^{K} \mathbb{1}[\texttt{null_scores}[k] \geq s_{\text{obs}}]}{K+1}$ {Permutation p-value}
$s_{\text{cal}} \leftarrow \max\left(\frac{s_{\text{obs}} - \tau_\alpha}{s_{\max} - \tau_\alpha}, 0\right)$ {Calibrated score (use $s_{\max} = 1$ for bounded metrics)}
return s_{cal}, p

Aggregation-aware calibration for layer-wise comparisons. When comparing models with multiple layers and reporting a summary statistic (*e.g.*, maximum similarity across layer pairs), the aggregation step must also be calibrated. Algorithm 2 shows how to extend scalar calibration to this setting. The key insight is that the *same* sample permutation must be applied consistently across all layers.

Computational cost. Scalar calibration requires K additional similarity computations. Aggregation-aware calibration requires $K \times L_A \times L_B$ computations, which can be parallelized across permutations. In practice, $K = 200$–500 permutations suffice for stable p-values and threshold estimation. On a single 1080 Ti GPU, 200-permutation calibration adds roughly 20 ms per layer pair for CKA and 52 ms for mKNN, so a full layer-wise model comparison completes in seconds.

E. Additional Experimental Results

This appendix provides additional analyses that support the main text claims.

E.1. Phase diagrams across different noise distributions

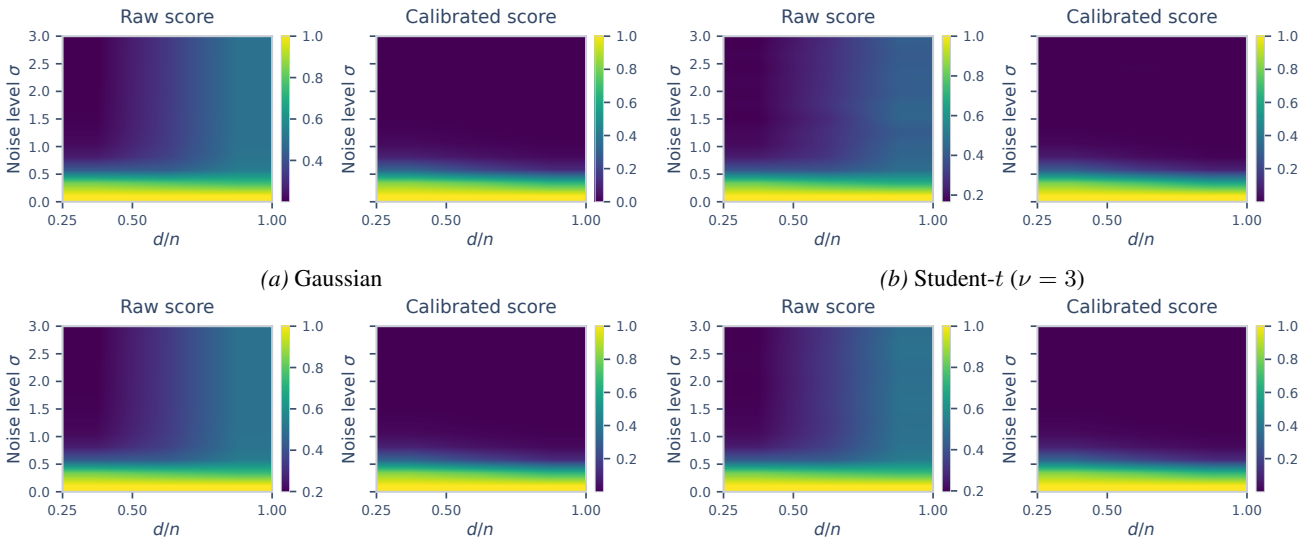
The theoretical analysis in Section 4 assumes Gaussian entries for tractability, but real neural network activations rarely follow Gaussian distributions. Instead, they often exhibit heavy tails, sparsity, or multimodality. A critical question is whether our calibration, which makes no distributional assumptions, remains effective under such deviations.

Figure 8 shows phase diagrams under different noise distributions: Gaussian, Student-t ($\nu = 3$), Laplace, and Gaussian mixtures. Each panel shows raw scores (left) and calibrated scores (right) across the $(d/n, \sigma)$ grid, where σ controls the noise level added to a fixed shared signal. At low σ, the signal dominates and both raw and calibrated scores correctly indicate high similarity. At high σ, noise overwhelms the signal, and similarity should approach zero. The key finding is that raw scores remain elevated (around 0.4–0.6) even at high noise levels where no detectable signal remains, while calibrated scores correctly collapse to near-zero. This pattern holds across all noise distributions tested, confirming that permutation-based calibration adapts to the data-generating process without requiring explicit distributional modeling.

Algorithm 2 Aggregation-Aware Null Calibration

Require: Layer representations $\{\mathbf{X}^{(\ell)}\}_{\ell=1}^{L_A}$, $\{\mathbf{Y}^{(\ell')}\}_{\ell'=1}^{L_B}$ (all n samples)
Require: Similarity function $\texttt{sim}(\cdot,\cdot)$, aggregator T (*e.g.*, max), permutations K, level α
Ensure: Calibrated aggregate T_{cal}, p-value p_{agg}

{Compute observed similarity matrix}
for $\ell = 1$ to L_A **do**
  for $\ell' = 1$ to L_B **do**
    $\mathbf{S}[\ell, \ell'] \leftarrow \texttt{sim}(\mathbf{X}^{(\ell)}, \mathbf{Y}^{(\ell')})$
  end for
end for
$T_{\text{obs}} \leftarrow T(\mathbf{S})$ {*e.g.*, $\max_{\ell,\ell'} \mathbf{S}[\ell, \ell']$}
$\texttt{null_aggregates} \leftarrow []$
for $k = 1$ to K **do**
  $\pi \leftarrow \texttt{random_permutation}(n)$ {Single permutation for all layers}
  for $\ell = 1$ to L_A **do**
    for $\ell' = 1$ to L_B **do**
      $\mathbf{S}^{(k)}[\ell, \ell'] \leftarrow \texttt{sim}(\mathbf{X}^{(\ell)}, \mathbf{Y}^{(\ell')}[\pi, :])$ {Same π for all ℓ'}
    end for
  end for
  $\texttt{null_aggregates}[k] \leftarrow T(\mathbf{S}^{(k)})$ {Aggregate under null}
end for
$\texttt{combined} \leftarrow [T_{\text{obs}}] \cup \texttt{null_aggregates}$ {Combined set}
$\tau_\alpha^{\text{agg}} \leftarrow \texttt{quantile}(\texttt{combined}, 1-\alpha)$ {Critical threshold from combined set}
$p_{\text{agg}} \leftarrow \frac{1+\sum_{k=1}^{K} \mathbb{1}[\texttt{null_aggregates}[k] \geq T_{\text{obs}}]}{K+1}$
$T_{\text{cal}} \leftarrow \max\left(\frac{T_{\text{obs}} - \tau_\alpha^{\text{agg}}}{s_{\max} - \tau_\alpha^{\text{agg}}}, 0\right)$
return $T_{\text{cal}}, p_{\text{agg}}$



(a) Gaussian *(b)* Student-t ($\nu = 3$)

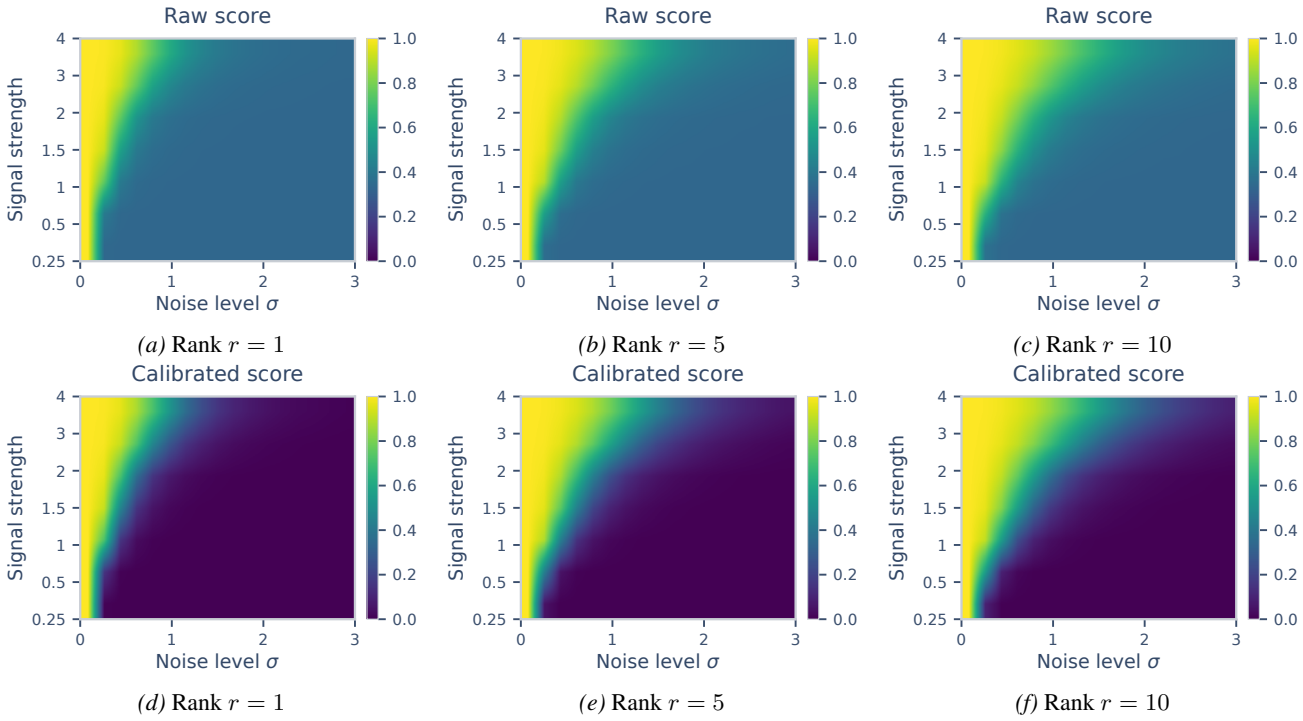
(c) Laplace *(d)* Gaussian mixture

Figure 8. **Phase diagrams under different noise types.** Calibrated scores (right) collapse to near-zero at high noise levels across the $(d/n, \sigma)$ grid, while raw scores (left) exhibit systematic positive bias. Calibration remains effective regardless of tail behavior.

E.2. SNR heatmaps

The experiments of the main paper (Figure 4) demonstrated that calibration eliminates false positives under H_0 while preserving sensitivity to fixed signals. This section extends the analysis by characterizing how calibrated similarity varies jointly with signal strength, noise level, and dimensionality ratio, thereby delineating the regimes in which similarity estimation remains reliable.

Figure 9 presents heatmaps of raw scores (top row) and calibrated scores (bottom row) across the (Noise level, Signal strength) grid for three signal ranks ($r \in \{1, 5, 10\}$). The results reveal a clear phase transition structure. Raw scores (top) show uniformly high values across most of the grid, obscuring the true detection boundary. Calibrated scores (bottom) reveal the underlying signal: high scores concentrate in the low-noise, high-signal corner (bottom-left), while scores correctly collapse to zero as noise increases (moving right) or signal weakens (moving down). The detection boundary shifts rightward (tolerating higher noise) as signal rank increases. This phase structure is meaningful: it delineates when similarity measurements carry information about shared structure versus when they reflect only finite-sample artifacts.



(a) Rank $r = 1$ *(b)* Rank $r = 5$ *(c)* Rank $r = 10$

(d) Rank $r = 1$ *(e)* Rank $r = 5$ *(f)* Rank $r = 10$

Figure 9. **SNR heatmaps (calibrated scores).** Higher-rank signals are detected at higher noise levels. The clear gradient confirms calibration preserves sensitivity to genuine structure.

Figure 10 provides a complementary view by collapsing the 2D heatmaps into 1D curves, plotting calibrated score against noise level for different signal strengths s. As expected, calibrated scores decrease monotonically with noise level: at low noise, scores are high (reflecting the detectable shared signal), while at high noise, scores collapse to zero (reflecting that the signal is buried). Stronger signals (larger s) maintain elevated scores across a wider range of noise levels before eventually succumbing. Higher-rank signals ($r = 5, 10$) show more gradual decay compared to $r = 1$, consistent with their greater statistical detectability. All curves converge to zero at high noise, confirming that the null floor is correctly calibrated regardless of signal strength or rank.

E.3. Comparing calibration approaches

A natural question is whether the choice of calibration summary affects the correction. We consider several approaches: *(i) gated score*, which thresholds at a significance level and rescales ($\alpha \in \{0.05, 0.1\}$); *(ii) null-centered*, subtracting the null mean; *(iii) z-score*, standardizing by null mean and standard deviation; and *(iv) ARI-style*, applying the chance-correction formula $(s - \mathbb{E}[s])/(s_{\max} - \mathbb{E}[s])$. Figure 11 evaluates these variants across metrics as d/n increases.

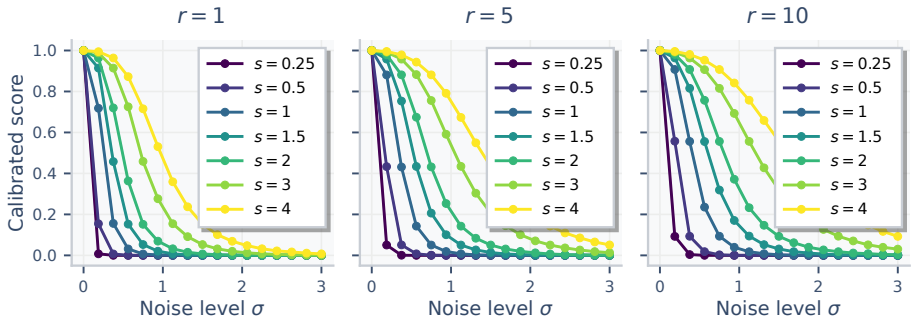


Figure 10. **Calibrated scores decay with noise level.** Each curve shows calibrated score versus noise level for a fixed signal strength s. Stronger signals maintain elevated scores across wider noise ranges; all curves converge to zero at high noise.

The results demonstrate that the gated score, null-centered, and ARI-style corrections all successfully collapse to appropriate null baselines across all metrics, regardless of whether the raw metric exhibits severe inflation (CKA, approaching 0.8) or mild inflation (RSA and mKNN, below 0.1). The z-score calibration, while correcting the mean, can exhibit artifacts when the null distribution is skewed, as occurs for bounded metrics like CKA at high d/n, making it less suitable as a universal correction.

E.4. Comparison with analytical debiasing

We validate our empirical null calibration by comparing it to existing analytical bias corrections for CKA. Figure 12 shows the difference between our calibrated CKA and two existing estimators: the debiased CKA of Murphy et al. (2024) and the dep-cols CKA of Chun et al. (2025).

Our calibrated CKA closely matches the debiased CKA estimator, indicating that our calibration automatically corrects the dominant width-induced bias without requiring a metric-specific derivation. In contrast, dep-cols CKA is designed to correct column dependence, which is not a confound in our experimental setup (where columns are independent by construction), and as a result, it attenuates the true signal under H_1. Our calibration targets the same source of finite-sample bias that the unbiased HSIC estimator of Song et al. (2012) subtracts analytically: the self-similarity (diagonal) terms. This is consistent with the close agreement between the two corrections in Figure 12. Unlike such analytical estimators, our permutation calibration requires no metric-specific derivation and applies to metrics for which no debiasing exists.

E.5. Width confounder under genuine signal on real networks

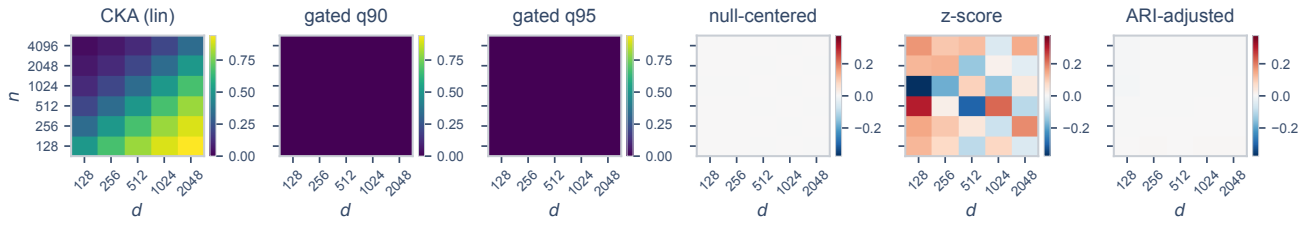
We now verify on real pretrained networks that calibration corrects the width confounder when a genuine signal is present, complementing the analytical result of Proposition C.5. Because the bias scales as $\mathcal{O}(d/n)$ (Proposition 4.1), width and sample size can be separated only by varying one while holding the other fixed, which this experiment does.

We extract last-layer features from the DINOv2 and AugReg ViT families. Within a family the models share training objective, data, and architecture and differ only in width d (*e.g.*, ViT-S/B/L/g for DINOv2), so a within-family pair isolates the effect of width. All pairs use the same $n = 1024$ WIT images as the PRH setting, and for each pair we compute raw and calibrated linear CKA. At fixed $n = 1024$, the permutation threshold τ (the magnitude of the width correction) increases with the total representation width $d_X + d_Y$, with Pearson correlation $r = 0.88$ (Figure 13). Since n is held constant, this dependence is attributable to width alone: calibration adapts its correction to representation width rather than applying a fixed offset. Together with Proposition C.5, this confirms that calibration corrects the width confounder on real networks with genuine similarity.

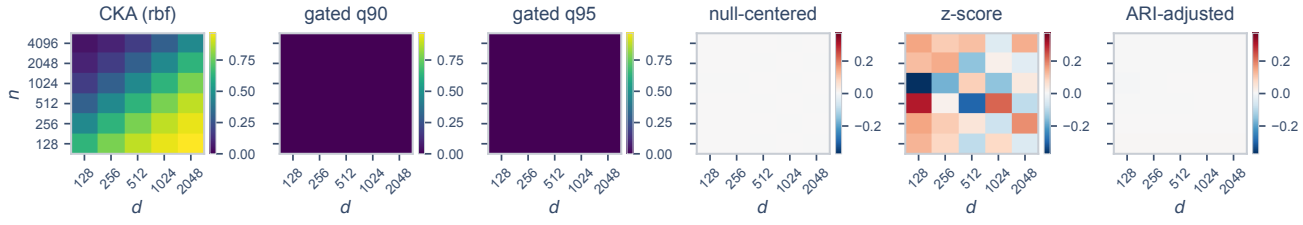
E.6. Permutation budget analysis

Permutation-based calibration introduces a computational-statistical tradeoff: more permutations yield more stable threshold estimates but increase runtime. Practitioners need guidance on the minimum budget required for reliable inference.

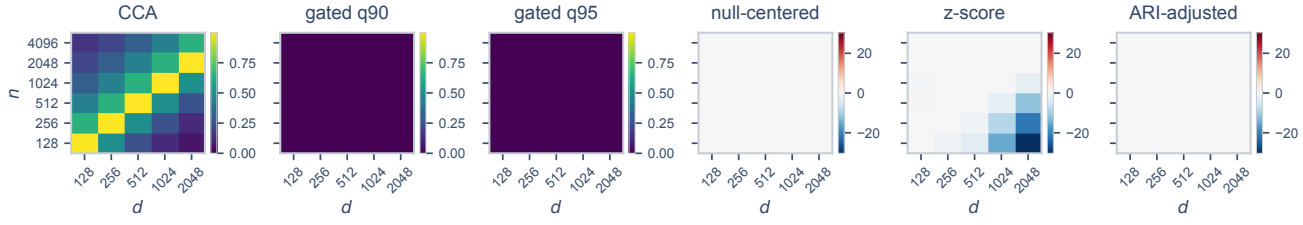
We analyze the stability of threshold estimates τ_α and calibrated scores as a function of the permutation budget K across 50 random seeds. Figure 14 shows two panels: the left panel displays threshold estimates, while the right panel shows calibrated scores under H_0. Threshold estimates (left) stabilize rapidly, reaching stable values by approximately $K = 50$ for



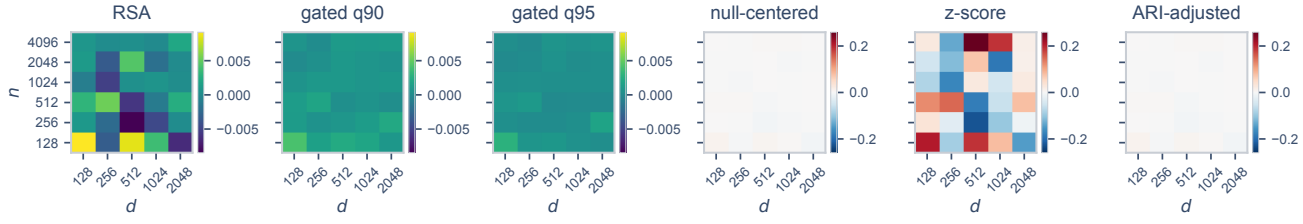
(a) CKA linear



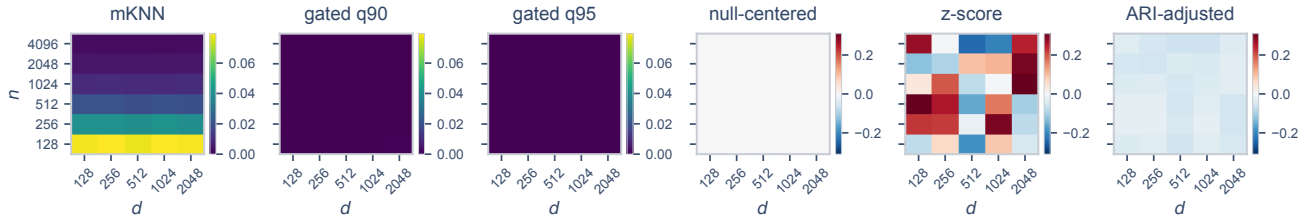
(b) CKA RBF



(c) CCA



(d) RSA (Spearman)



(e) Mutual k-NN

Figure 11. **Comparing calibration approaches across metrics.** Each panel shows raw scores alongside four calibration variants (gated score, null-centered, z-score, ARI-style) as d/n increases. Gated score, null-centered, and ARI-style corrections collapse to appropriate baselines; z-score exhibits artifacts for skewed null distributions.

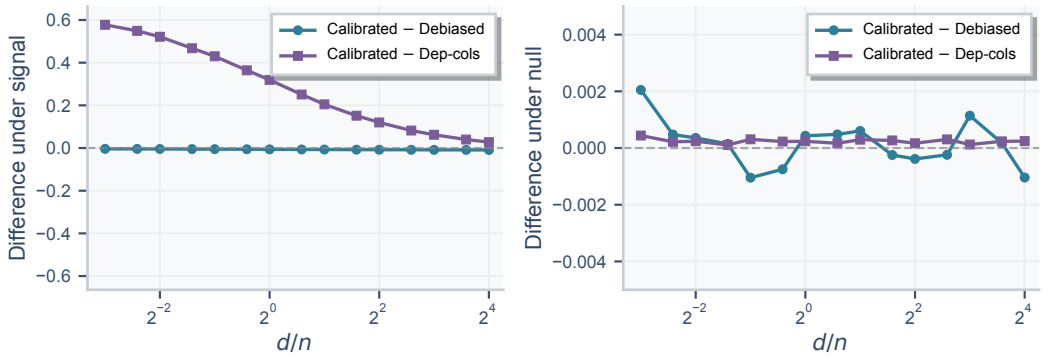


Figure 12. **Calibration recovers analytical debiasing.** Difference between calibrated CKA and existing estimators ($n = 1024$, d/n swept). (Left) Under signal. (Right) Under null.

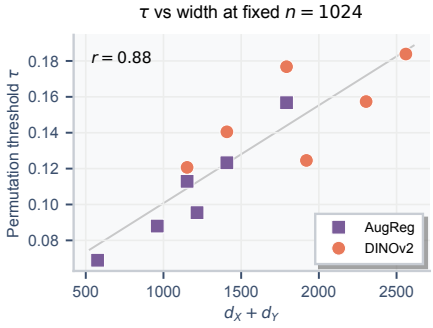


Figure 13. **The width correction scales with representation width (real networks).** For DINOv2 and AugReg ViT pairs on the same $n = 1024$ WIT images, the permutation threshold τ (the width correction) grows with total width $d_X + d_Y$ (Pearson $r = 0.88$). With n fixed, the increasing correction is driven by width alone.

all metrics tested. Calibrated scores (right) exhibit more variability at very low budgets ($K < 50$), with occasional spikes due to unstable threshold estimation, but converge to near-zero by $K \approx 100$–200.

Based on these results, we recommend $K \geq 200$. The computational cost scales linearly with K, so this recommendation represents a favorable tradeoff between precision and efficiency.

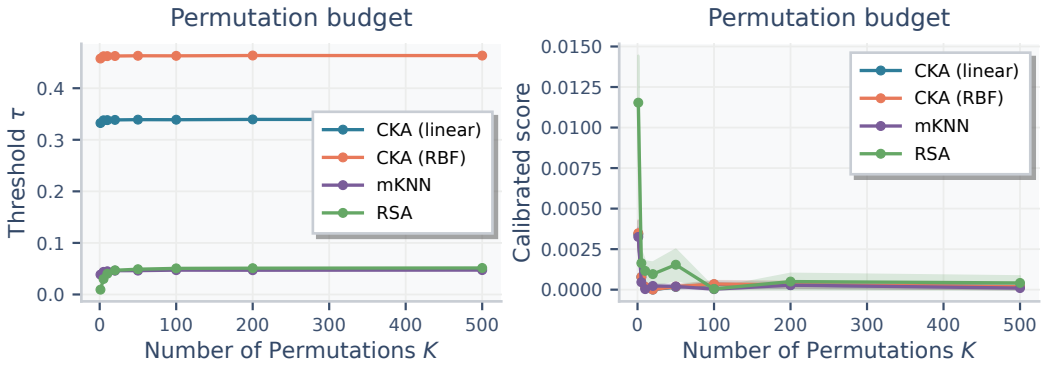


Figure 14. **Permutation budget analysis.** Left: threshold τ_α stabilizes by $K \approx 50$. Right: calibrated scores under H_0 converge to near-zero by $K \approx 100$–200. Shaded regions show variability across random seeds.

E.7. Full null drift results

The main text presents null drift results for a representative subset of metrics under Gaussian noise. Here, we present additional results across all metrics evaluated in this work, including RSA, the RV coefficient, and Procrustes distance, as well as results under heavy-tailed noise distributions.

Figure 15 presents results under Gaussian noise for all metrics. The severity of null drift varies substantially across metric families: CKA variants exhibit the most severe inflation, followed by RV coefficient and CCA-variants, with neighborhood metrics showing the mildest drift. This reflects the structural sensitivity of the metrics to high-dimensional spurious correlations. Critically, calibration eliminates drift across all metrics, collapsing scores to zero regardless of the raw bias magnitude.

Figure 16 extends these results to heavy-tailed noise (Student-t, $\nu = 3$). The qualitative pattern is preserved: all metrics exhibit positive drift under the null, and calibration eliminates this drift. The magnitude of raw bias is generally higher under heavy-tailed noise, consistent with increased finite-sample variability, yet calibration adapts automatically without requiring distributional knowledge.

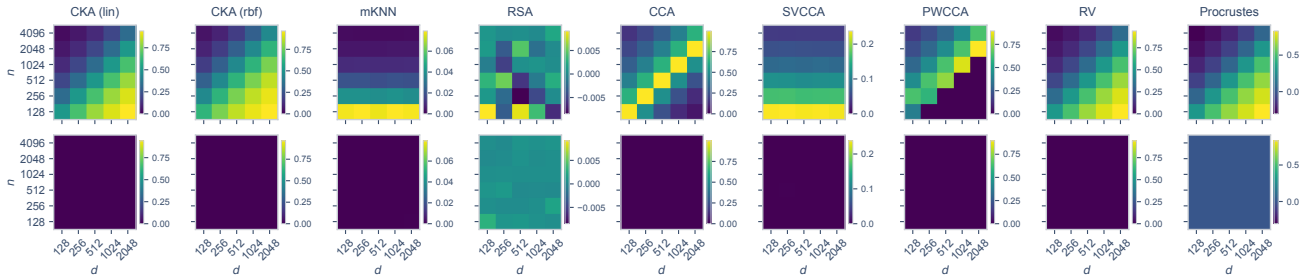


Figure 15. **Full null drift results (Gaussian).** Raw scores (top) exhibit systematic positive bias; calibrated scores (bottom) collapse to zero.

Robustness to the generative process. The experiments above use i.i.d. Gaussian and heavy-tailed noise. To confirm that the width confounder is not tied to these specific generators, we also produce representations as $\mathbf{x}_i = f_x(\mathbf{a}_i)$ and $\mathbf{y}_i = f_y(\mathbf{a}_i)$, where $\mathbf{a}_i$ is a shared Gaussian input and f_x, f_y are *independent* random linear maps or MLPs, so the two representations share inputs but no systematic structure. Figure 17 shows that raw CKA still inflates with dimensionality d in every regime, while calibrated CKA stays at zero. Sharing inputs through unrelated networks thus produces width-driven inflation rather than genuine convergence, and calibration removes it regardless of how the representations are generated.

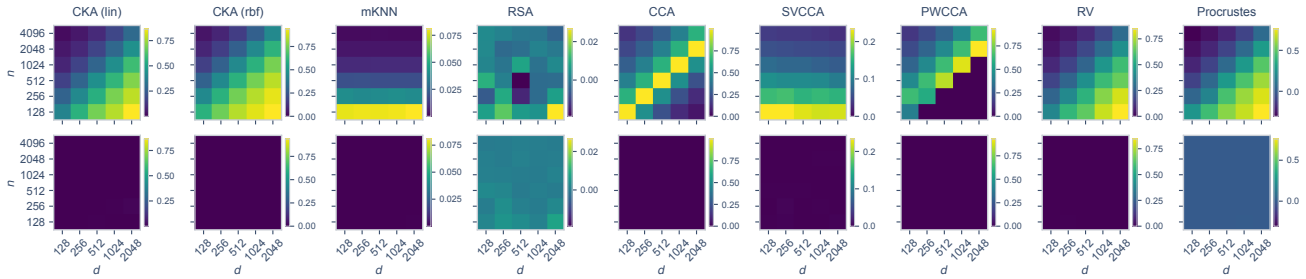


Figure 16. **Full null drift results (heavy-tailed).** Student-t ($\nu = 3$) noise. The pattern is consistent across all metrics: calibration eliminates spurious similarity regardless of noise distribution.

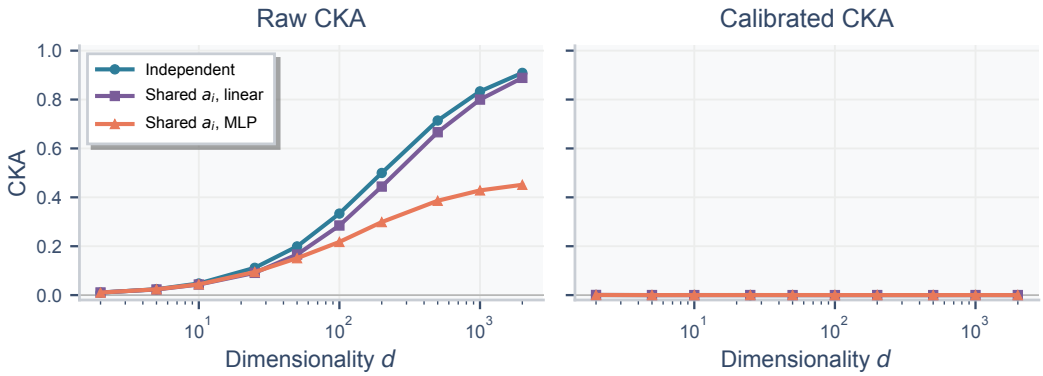


Figure 17. **The width confounder is robust to the generative process.** Even when $\mathbf{X}$ and $\mathbf{Y}$ are computed from the *same* inputs by *unrelated* random networks (linear or MLP), raw CKA (left) inflates with dimensionality d although no systematic alignment is present; calibrated CKA (right) stays at zero. The inflation is a finite-width artifact, not evidence of convergence.

E.8. Extended PRH alignment results (image–text)

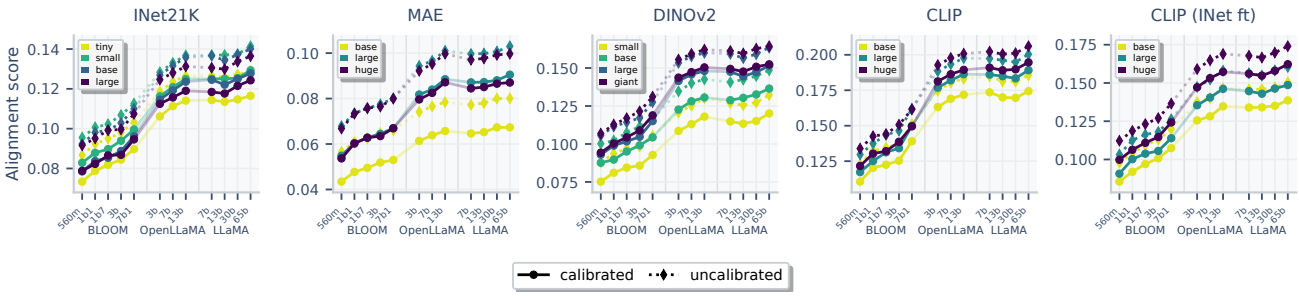
The main text establishes a divergence between local and global similarity metrics when applied to the Platonic Representation Hypothesis (PRH): neighborhood-based metrics retain significant cross-modal alignment after calibration, while spectral metrics lose their apparent convergence trend. A natural question is whether this finding is robust across model families and metric variants.

Here we present comprehensive results across all five vision model families in the PRH setting (DINOv2, CLIP, ImageNet-21K, MAE, and CLIP-finetuned) and a broad range of metrics spanning the local-to-global spectrum (Figures 18 to 20).

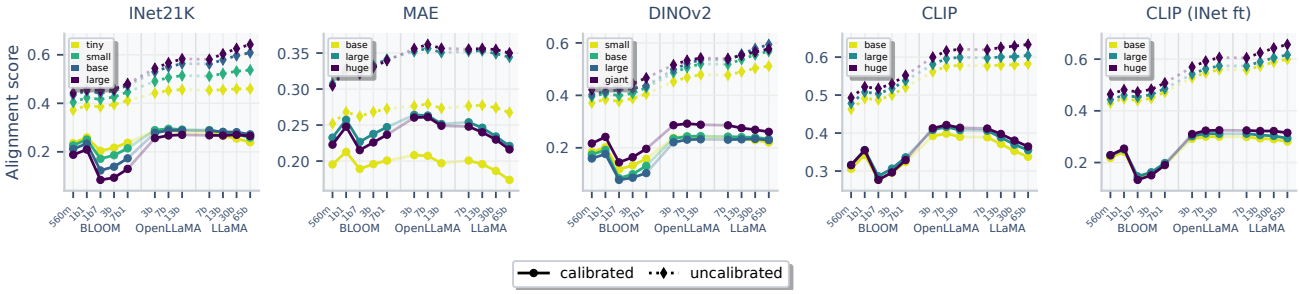
The results reinforce and extend the main text findings. Neighborhood metrics (mKNN, cycle-kNN, CKNNA) show a consistent alignment trend across all vision families with a neighborhood size of 10. This pattern holds for both self-supervised (DINOv2, MAE) and supervised (ImageNet-21K) pretraining objectives, as well as for both CLIP-aligned and CLIP-finetuned variants. Spectral metrics (CKA linear, CKA RBF, unbiased CKA, RV coefficient, SVCCA) show a different pattern: raw scores suggest increasing alignment with model scale, but calibrated scores show no such scaling trend.

Trend with model scale, before versus after calibration. For each metric separately, we measure the Pearson correlation between language-model capability (the model ranking of Huh et al. (2024)) and the similarity score across all model pairs, before and after calibration (Table 2). Global metrics lose most of their trend with scale after calibration, while local (neighborhood) metrics retain it across neighborhood sizes.

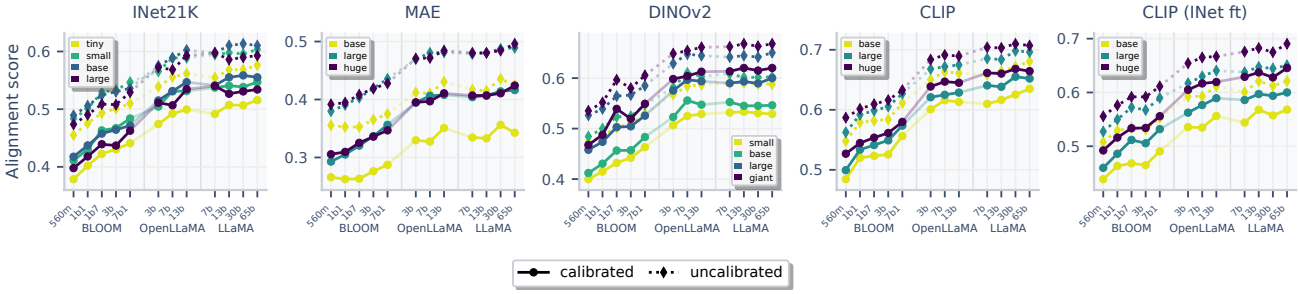
Statistical significance. Beyond calibrated scores, we report permutation p-values to quantify statistical evidence against the null hypothesis of no cross-modal alignment (Figure 21). All 204 vision–language model pairs are significant at $p < 0.05$, with most achieving $p < 0.005$ (the minimum achievable with $K = 200$ permutations) for both local and global metrics. This confirms that cross-modal similarity is statistically significant (*i.e.*, has some alignment) across all model pairs. The critical distinction between local and global metrics lies not in statistical significance but in the magnitude and trends of calibrated scores. Local metrics show substantial alignment above the null threshold that persists across scales, whereas



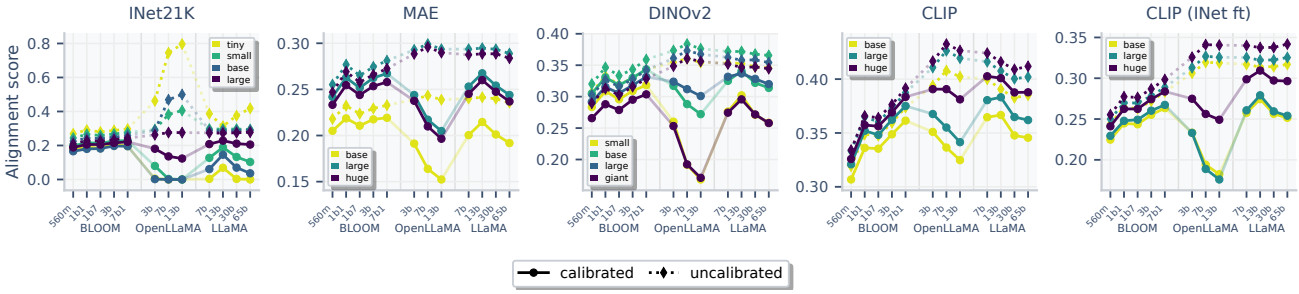
(a) mKNN: Neighborhood overlap.



(b) CKA RBF: Spectral alignment.

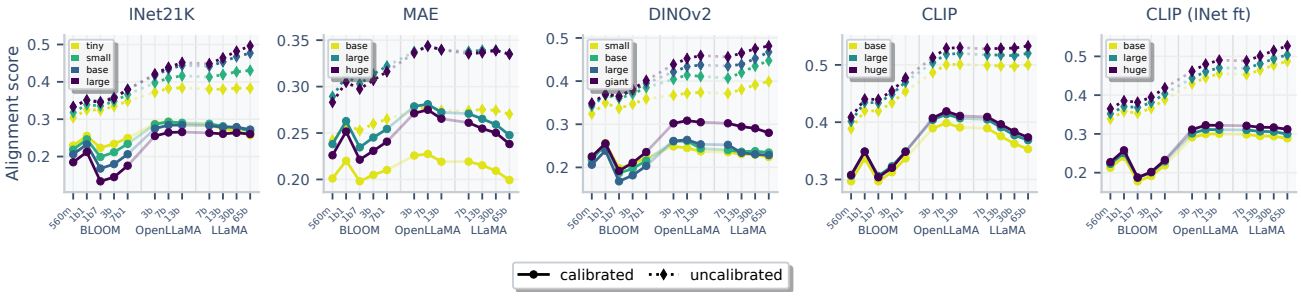


(c) cycle-kNN: Bidirectional consistency.

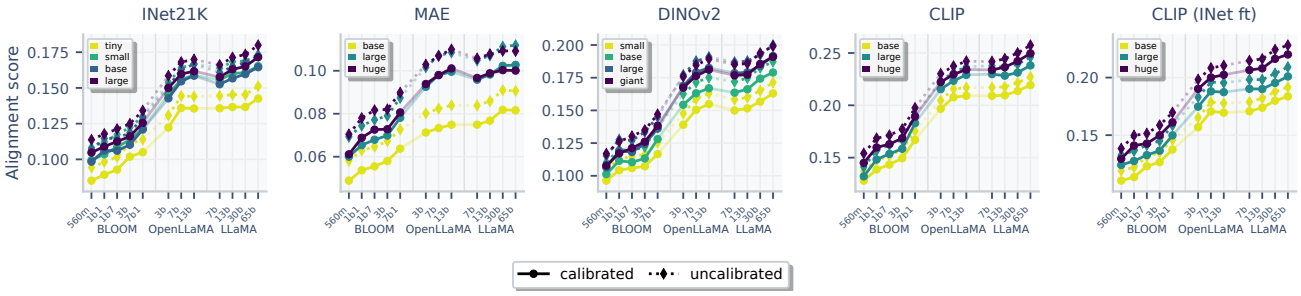


(d) Unbiased CKA.

Figure 18. **PRH alignment results (all vision families).** All five vision model families are shown (DINOv2, CLIP, ImageNet-21K, MAE, CLIP-finetuned). The divergence between local and global metrics is consistent across all families.

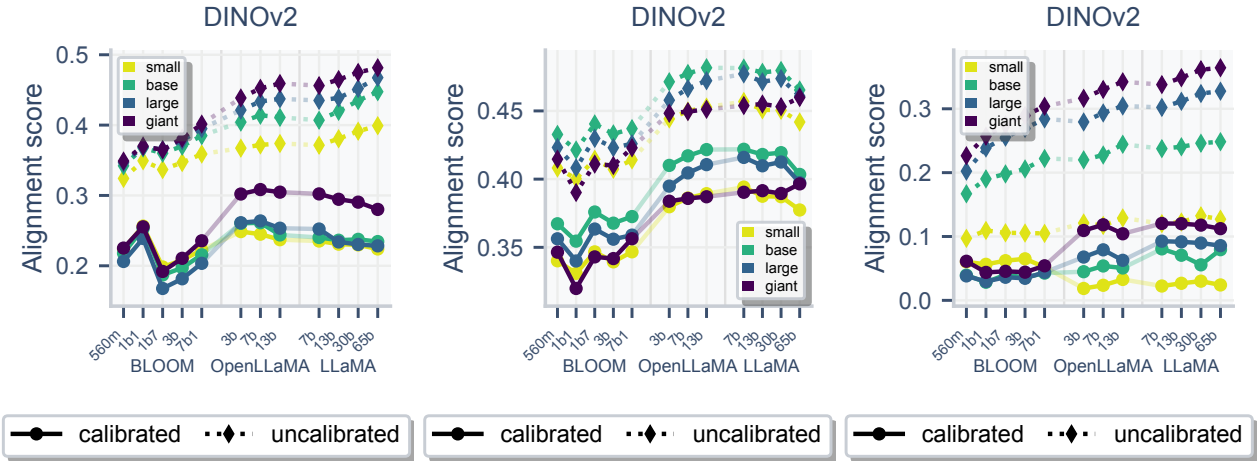


(a) CKA linear.



(b) CKNNA.

Figure 19. **Additional PRH metrics (all vision families).** CKA linear (a) shows the same loss of convergence trend as CKA RBF. CKNNA (b) shows consistent local alignment across all vision families.



(a) RV coefficient. *(b)* SVCCA. *(c)* Procrustes.

Figure 20. **Additional geometric and spectral metrics (DINOv2).** The RV coefficient (a), SVCCA (b), and Procrustes distance (c) confirm the same pattern as CKA: calibrated scores show no convergence trend with model scale, so the disappearance of global convergence is not specific to CKA but extends to shape- and geometry-based metrics. Due to the computational cost of permutation-based calibration for these metrics, we report results for DINOv2 only.

Table 2. **Trend of each similarity metric with model scale, before versus after calibration.** Pearson correlation between language-model capability and the similarity score across all vision–language model pairs in the PRH setting. Global metrics lose their trend with scale after calibration; local (neighborhood) metrics retain it.

Metric	Before	After
Global metrics (spectral and geometric)		
CKA (linear)	0.86	0.45
CKA (RBF)	0.83	0.35
Unbiased CKA	0.65	−0.01
Procrustes	0.89	0.39
RV coefficient	0.92	0.31
Local metrics (neighborhood)		
mKNN (k=10)	0.85	0.84
mKNN (k=20)	0.84	0.84
mKNN (k=50)	0.83	0.82
mKNN (k=100)	0.79	0.79
CKNNA (k=10)	0.87	0.87
CKNNA (k=20)	0.86	0.86
CKNNA (k=50)	0.86	0.86
CKNNA (k=100)	0.84	0.84
cycle-kNN (k=10)	0.85	0.85
cycle-kNN (k=20)	0.85	0.85
cycle-kNN (k=50)	0.84	0.85
cycle-kNN (k=100)	0.77	0.74

global metrics, although significant, show no convergence in calibrated effect sizes.

E.9. Extended video–language alignment results

The main text extends the PRH analysis to video–language alignment following Zhu et al. (2026). Here, we provide additional results to verify that the local-vs-global pattern observed for image–language alignment extends to the video modality.

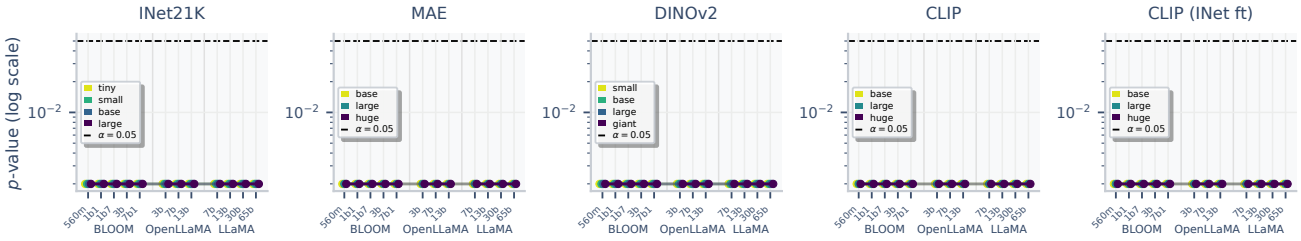
We use 1024 samples from the PVD (Bolya et al., 2025; Cho et al., 2025) test set. We evaluate video-native models (VideoMAE (Tong et al., 2022)) and, as a frame-level baseline, image models (DINOv2 and CLIP) applied to the middle frame of each video, comparing all against the same three language model families used in the image–language experiments (BLOOM, OpenLLaMA, LLaMA) at multiple scales. For VideoMAE we include both the fine-tuned scale series (small-/base/large/huge, fine-tuned on Kinetics) used in the main text and the corresponding non-fine-tuned checkpoints. Figure 22 shows results for spectral (CKA RBF) and neighborhood (mKNN, CKNNA) metrics.

The pattern mirrors the image–language findings. For spectral metrics, raw scores suggest alignment, whereas calibrated scores drop significantly, indicating that much of the apparent alignment is attributable to width and depth confounders. In contrast, neighborhood metrics retain significant alignment after calibration, confirming that video and language representations share local topological structure. This local alignment strengthens with the capability of the video encoder: both fine-tuning on Kinetics and increasing scale raise the calibrated neighborhood alignment, in line with the Aristotelian hypothesis that more capable representations converge more in local structure. Calibration removes the spectral inflation throughout.

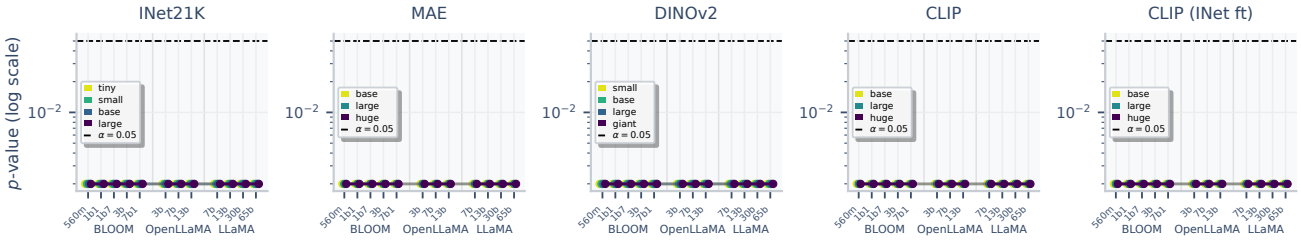
E.10. Characterizing the locality of cross-modal alignment

The main text establishes that local neighborhood metrics retain significant alignment after calibration, while global spectral metrics do not. We next characterize *how local* this alignment is. Both mKNN and CKA-RBF have hyperparameters that control their sensitivity to local versus global structure. By varying these parameters, we can characterize the scale at which cross-modal alignment emerges.

Experimental setup. We vary two locality parameters: the neighborhood size k in mKNN, testing $k \in \{10, 20, 50, 100\}$ where smaller values focus on immediate neighbors and larger values consider broader local structure, and the RBF kernel bandwidth σ in CKA-RBF, testing $\sigma \in \{0.1, 0.5, 2.0, 5.0\}$, which controls the length scale over which the kernel assigns



(a) mKNN ($k = 10$).



(b) CKA linear.

Figure 21. **Permutation p-values for PRH alignment.** All model pairs are significant at $p < 0.05$, with most achieving $p < 0.005$ for both local (a) and global (b) metrics. The difference between metric families lies in calibrated effect sizes, not significance.

significant weight.

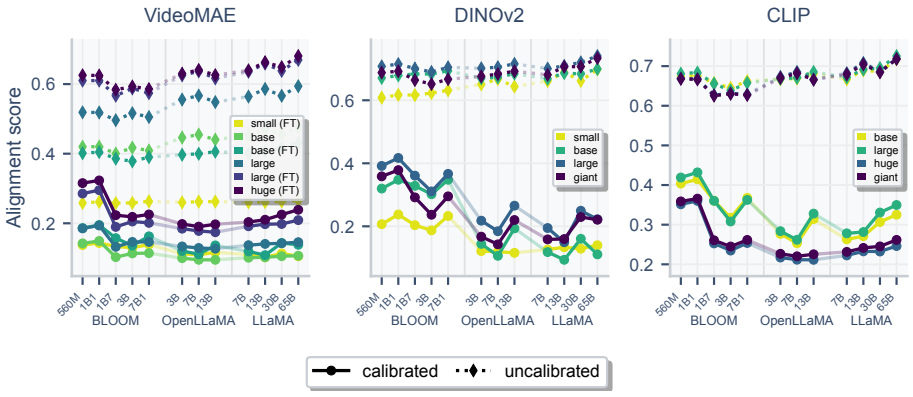
RBF bandwidth. The RBF (radial basis function) kernel is defined as $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2\right)$. The bandwidth σ determines the *length scale* of similarity. When σ is small (*e.g.*, 0.1), the kernel is sharply peaked: only very close points contribute significantly to the Gram matrix, making the similarity measure sensitive to *exact pairwise distances* in the immediate neighborhood. When σ is large (*e.g.*, 5.0), the kernel is broad: even moderately distant points contribute, and the similarity measure aggregates information over larger neighborhoods, becoming sensitive to coarser geometric structure.

Neighborhood size. For mKNN, the parameter k controls how many nearest neighbors are considered when measuring overlap. Small k (*e.g.*, 10) measures agreement on immediate neighbors, *i.e.*, the closest points to each sample, capturing fine-grained local topology. Large k (*e.g.*, 100) measures agreement on a broader neighborhood. With $n = 1000$ samples and $k = 100$, we ask whether the 10% closest points agree across representations. Crucially, mKNN is a *rank-based* metric: it asks *which* points are neighbors (ordinal information), not *how close* they are (cardinal information).

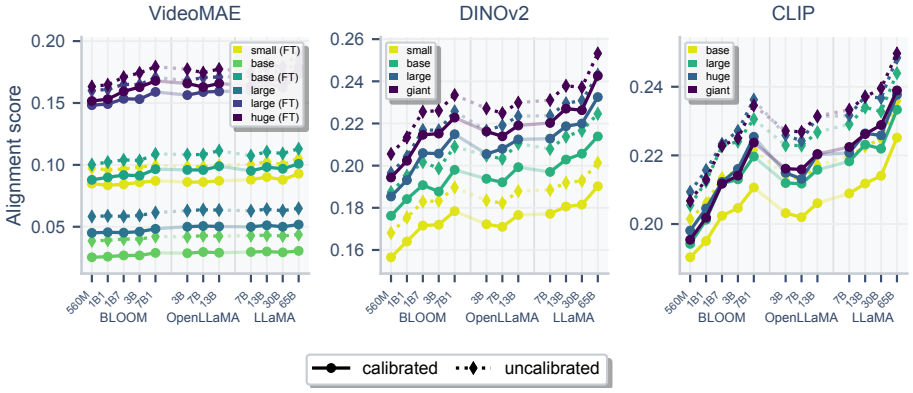
mKNN across k values. Figure 23 shows the PRH alignment results for mKNN with varying k. A consistent pattern emerges: all k values show significant alignment after calibration, with calibrated scores remaining well above zero even at $k = 100$. However, the scaling trend is most pronounced at small k. For $k = 10$, raw scores show a clear upward trend with model capacity that persists after calibration. At large k, this trend flattens even in raw scores. For $k = 100$, raw scores plateau for larger models, suggesting that broader neighborhood agreement is already saturated across model scales. This pattern indicates that scaling-driven improvement in alignment is concentrated at the finest topological level.

CKA-RBF across bandwidth values. Figure 24, and the accompanying p-values in Figure 25, show results for CKA-RBF with varying bandwidth σ, revealing a different pattern from mKNN. At $\sigma = 0.1$ (very local), there is no significant alignment after calibration: raw scores are near 1.0, reflecting the high similarity of any high-dimensional representations under a sharply peaked kernel. However, calibrated scores collapse to approximately zero with p-values exceeding 0.05 for most model pairs, indicating that the observed similarity is indistinguishable from chance. At $\sigma = 0.5$, alignment emerges, but with a flattening trend after calibration. Calibrated scores initially rise with model scale, then plateau and slightly decline for the largest models. At $\sigma = 2.0$ and $\sigma = 5.0$, significant alignment persists, but the calibrated trend also flattens, resembling the pattern observed for large-k mKNN: alignment exists, but scaling-driven improvement disappears after calibration.

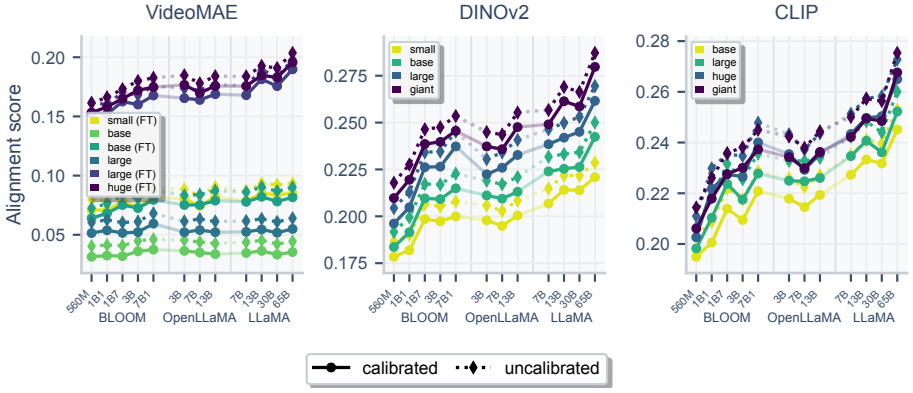
Topological versus metric alignment. The contrasting behavior of mKNN and small-σ CKA-RBF reveals a fundamental distinction in what "local alignment" means. On one hand, mKNN measures *topological* alignment: do the representations



(a) CKA RBF (spectral).

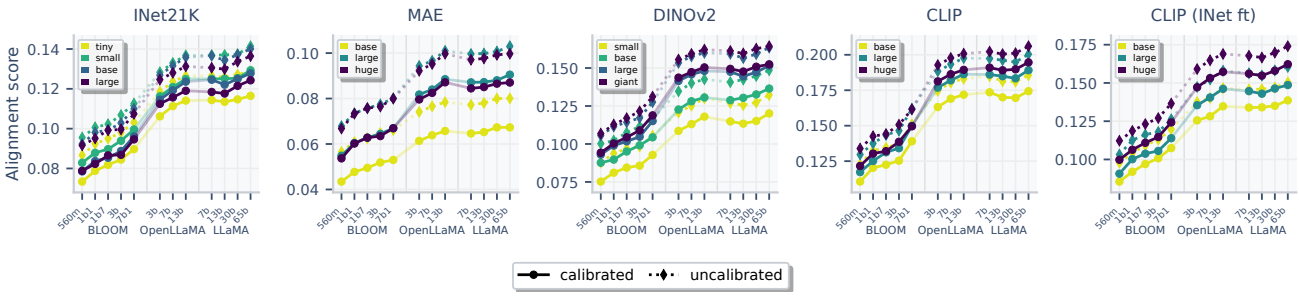


(b) mKNN ($k = 10$, neighborhood).

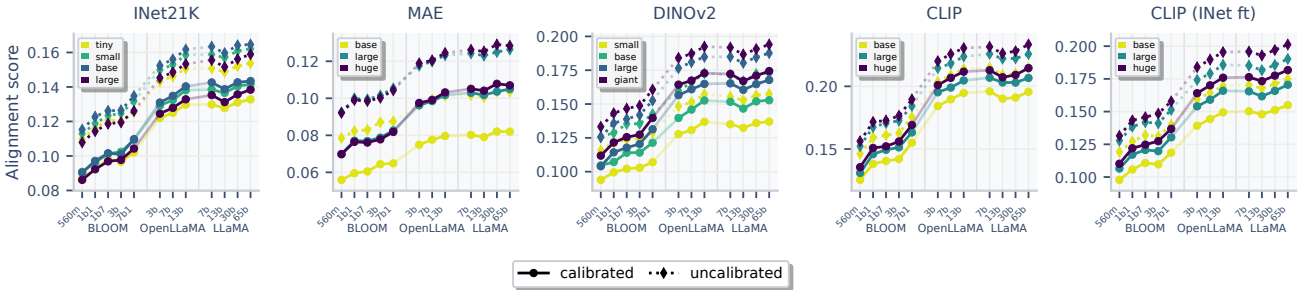


(c) CKNNA ($k = 10$, neighborhood).

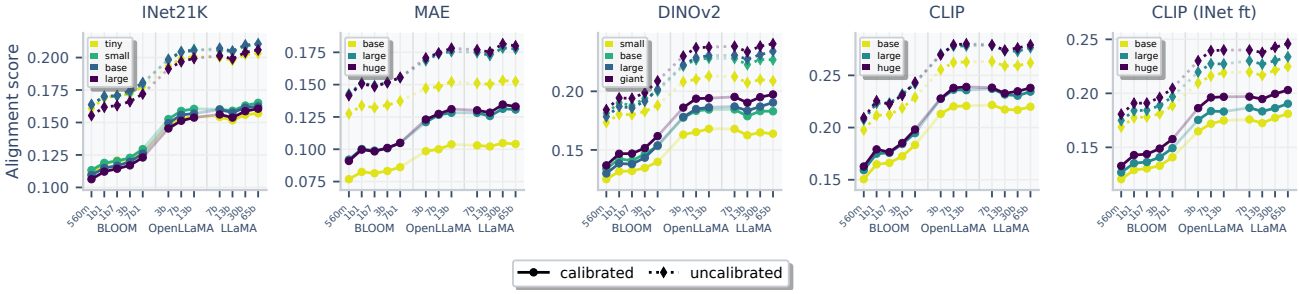
Figure 22. **Extended video–language alignment.** Each panel shows VideoMAE (fine-tuned and non-fine-tuned), DINOv2, and CLIP (frame-level) against the language model families; raw scores are dotted, calibrated solid. (a) Spectral alignment (CKA RBF) drops after calibration. (b,c) Neighborhood alignment is retained and strengthens with video-encoder capability.



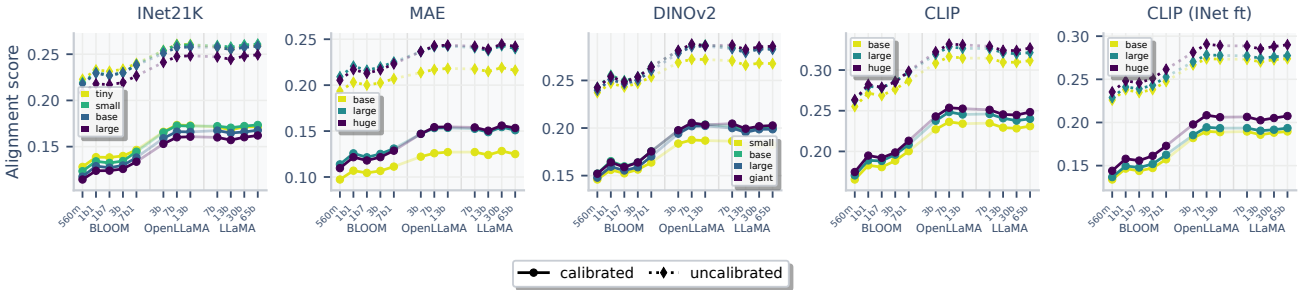
(a) mKNN ($k = 10$)



(b) mKNN ($k = 20$)

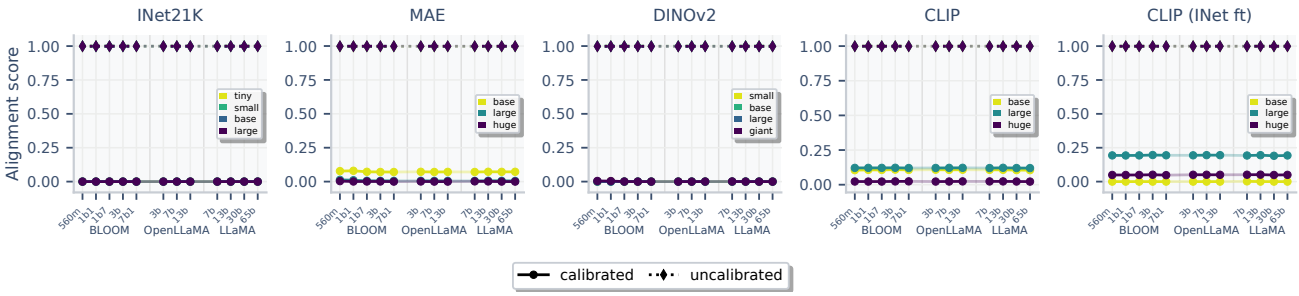


(c) mKNN ($k = 50$)

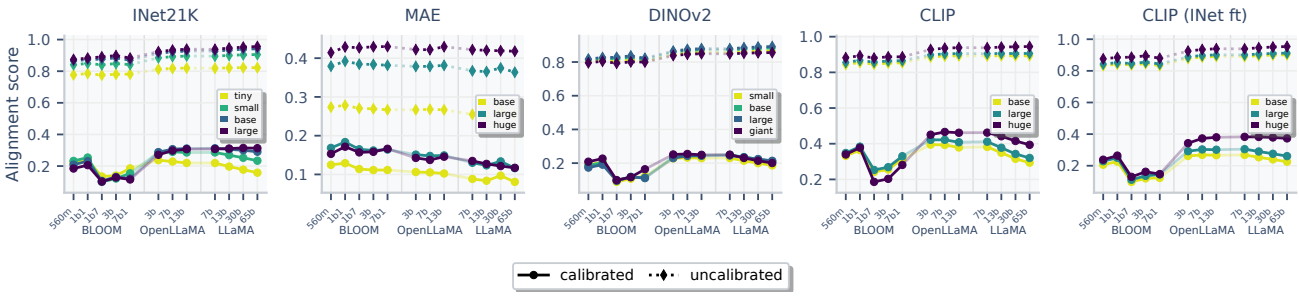


(d) mKNN ($k = 100$)

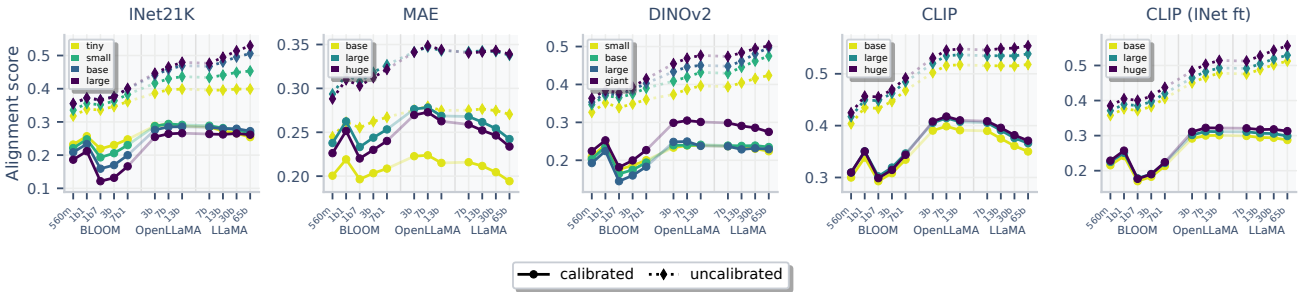
Figure 23. **PRH alignment with varying neighborhood size k for mKNN.** All k values show significant alignment after calibration. The scaling trend is clearest at small k and flattens at large k, suggesting scaling improvements are concentrated at the finest local scale.



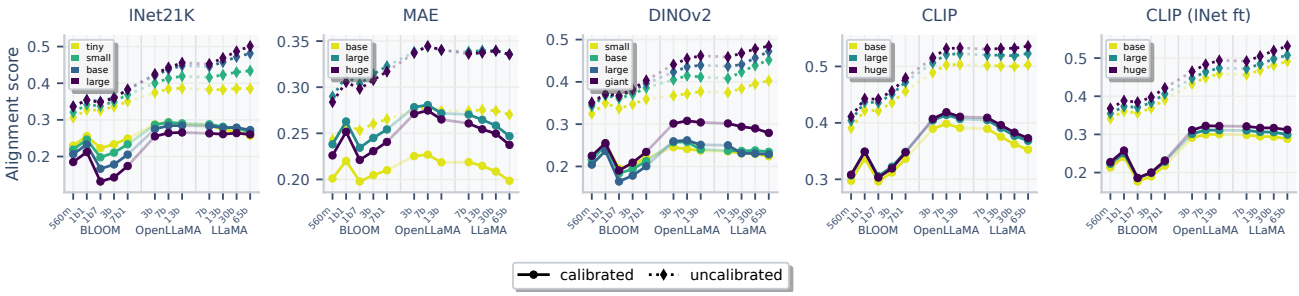
(a) CKA-RBF ($\sigma = 0.1$)



(b) CKA-RBF ($\sigma = 0.5$)

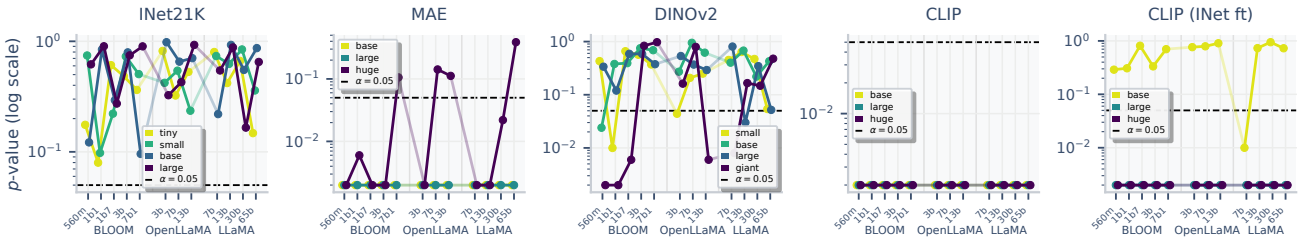


(c) CKA-RBF ($\sigma = 2.0$)

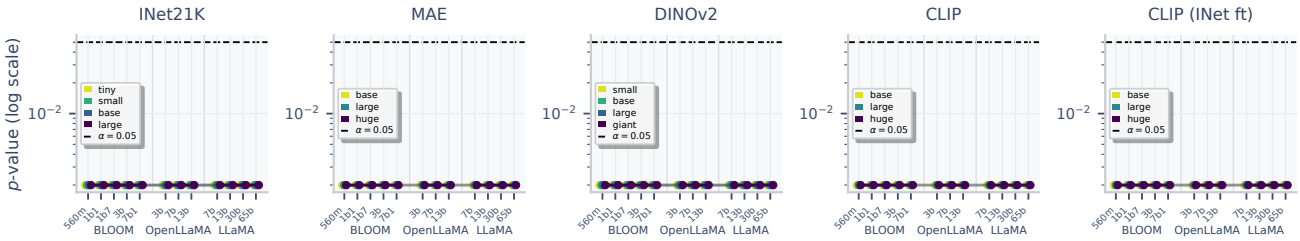


(d) CKA-RBF ($\sigma = 5.0$)

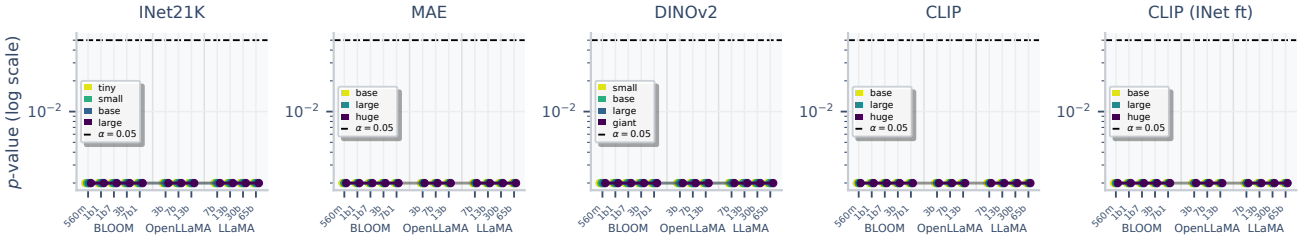
Figure 24. **PRH alignment with varying bandwidth σ for CKA-RBF.** At very small σ (a), no significant alignment remains after calibration. Larger σ values (b–d) show significant alignment, but the scaling trend flattens after calibration.



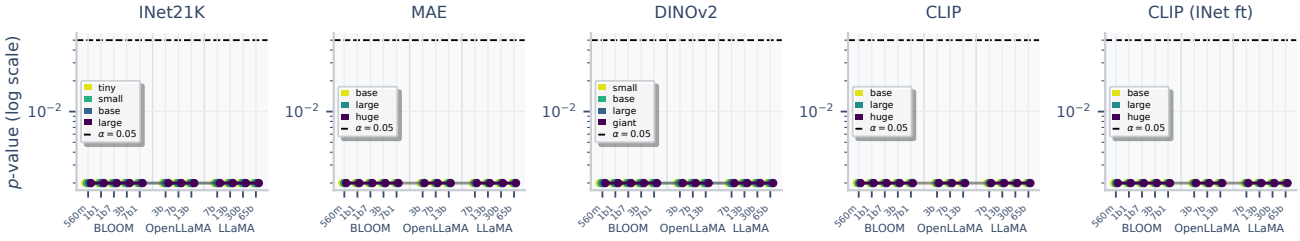
(a) CKA-RBF ($\sigma = 0.1$)



(b) CKA-RBF ($\sigma = 0.5$)



(c) CKA-RBF ($\sigma = 2.0$)



(d) CKA-RBF ($\sigma = 5.0$)

Figure 25. **Significance of PRH alignment with varying bandwidth σ for CKA-RBF.** Alignment with $\sigma = 0.1$ (a) is not significant for multiple models where larger bandwidths have significance (b–d).

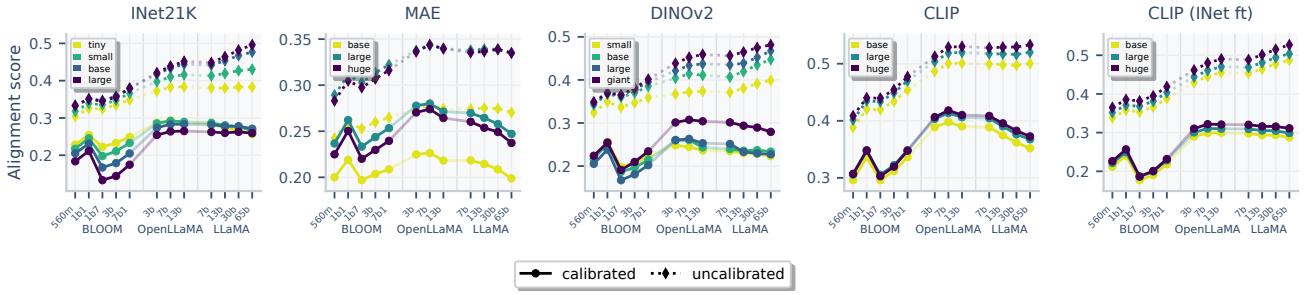
agree on *which* points are neighbors? This captures ordinal information where the ranking of distances matters but not their absolute values. On the other hand, small-σ CKA-RBF measures *metric* alignment: do the representations agree on *how close* neighbors are? This captures cardinal information where exact distance values matter.

The fact that mKNN shows alignment at all k values while small-σ CKA-RBF shows no alignment reveals that cross-modal representations agree on neighborhood identity (which points are close) but not on exact local distances (how close they are). This finding is consistent with the observation that different training objectives and architectures induce different distance scales in representation space while preserving the relative ordering of neighbors. The *Aristotelian* Representation Hypothesis should therefore be understood as convergence to shared *topological* structure rather than shared *metric* structure.

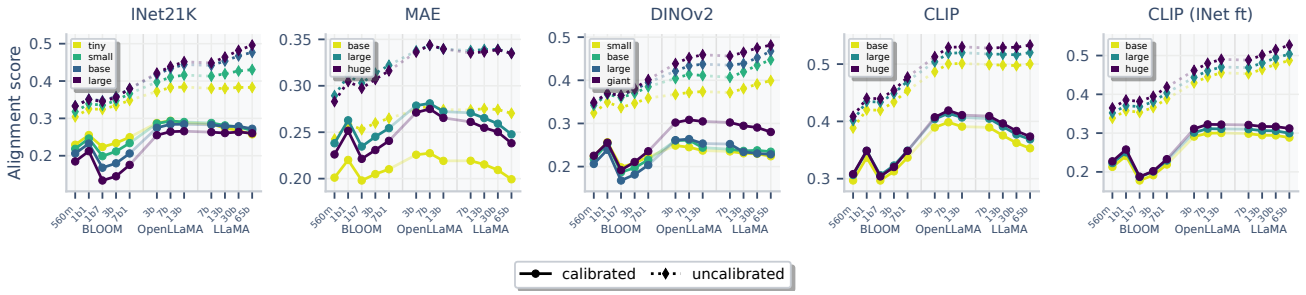
E.11. Sensitivity to significance level α

The main text uses a significance level of $\alpha = 0.05$ throughout. To confirm that the PRH conclusions are not sensitive to this choice, we repeat the PRH evaluation from Section 6.3 with $\alpha \in \{0.01, 0.05, 0.10\}$ for representative global (CKA linear, CKA RBF) and local (mKNN with $k = 10$) metrics.

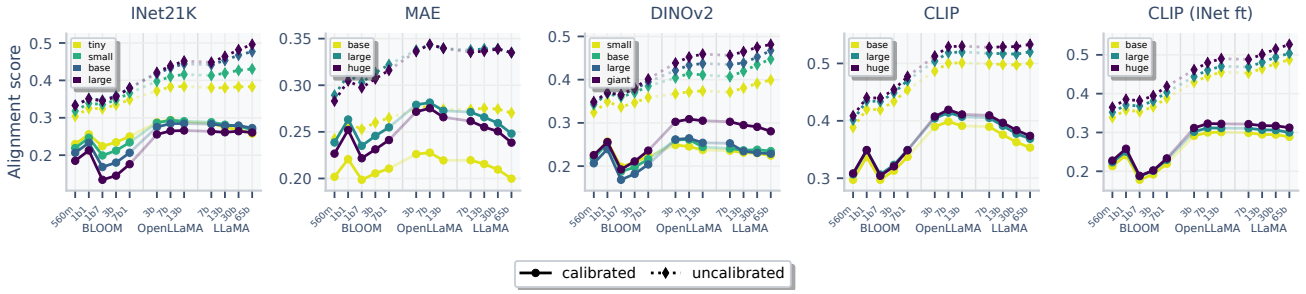
Figures 26 to 28 show that the conclusions are entirely invariant to the choice of α. For global metrics, calibrated scores show no convergence trend at any significance level. For local metrics, calibrated scores retain their alignment trend across all three α values. Stricter thresholds ($\alpha = 0.01$) produce slightly lower calibrated scores, while more permissive thresholds ($\alpha = 0.10$) produce slightly higher ones, but the qualitative pattern is unchanged. This confirms that our findings are not an artifact of a particular significance level.



(a) $\alpha = 0.01$

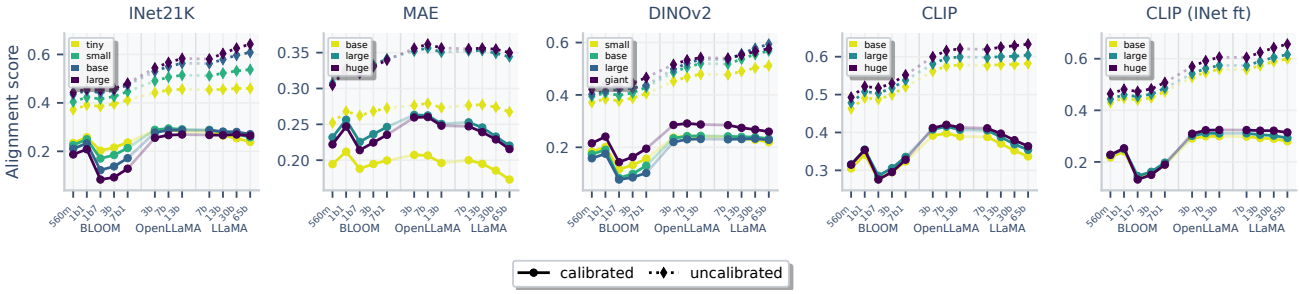


(b) $\alpha = 0.05$ (default)

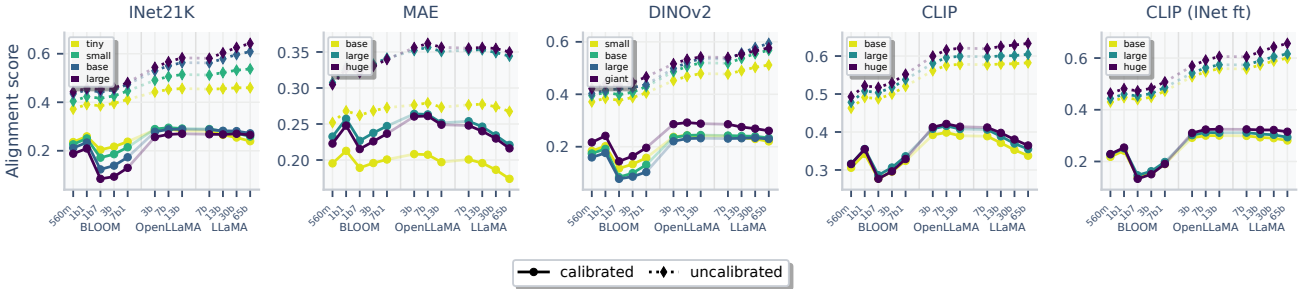


(c) $\alpha = 0.10$

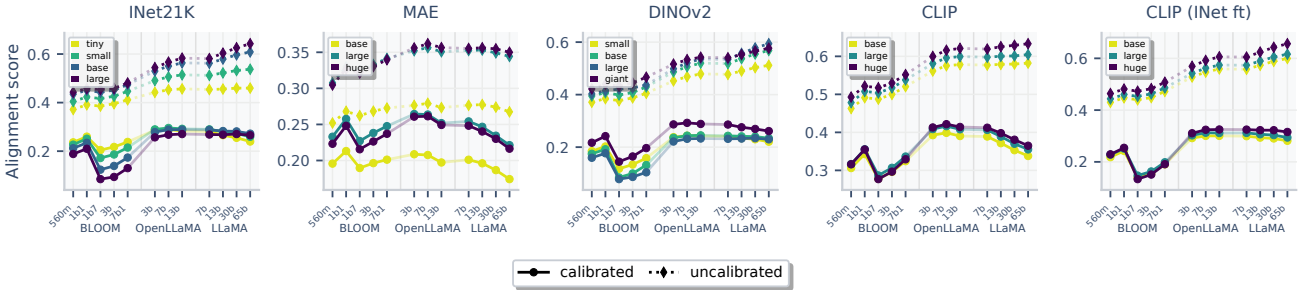
Figure 26. **Sensitivity to α for CKA linear.** Calibrated scores show no convergence trend regardless of significance level.



(a) $\alpha = 0.01$

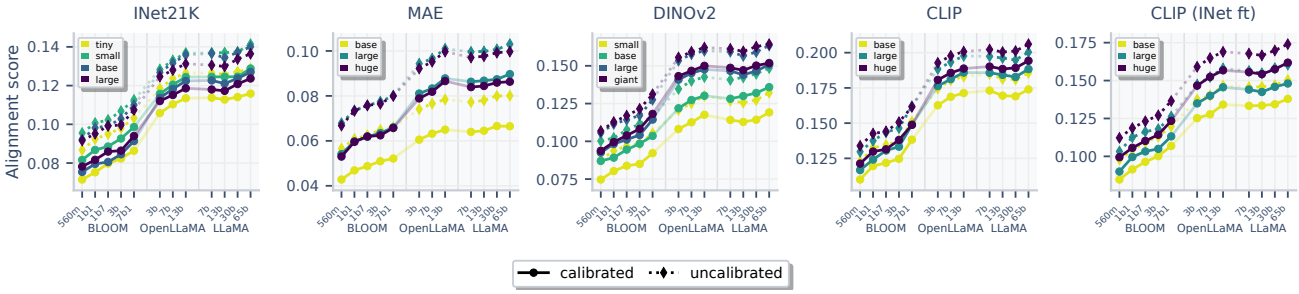


(b) $\alpha = 0.05$ (default)

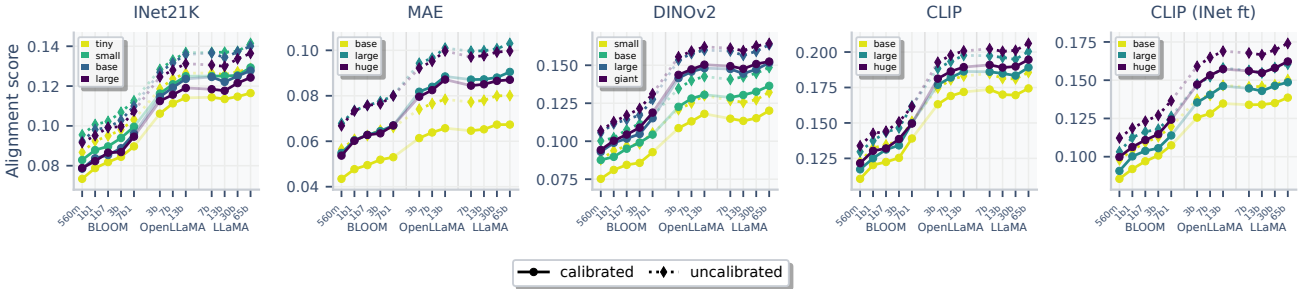


(c) $\alpha = 0.10$

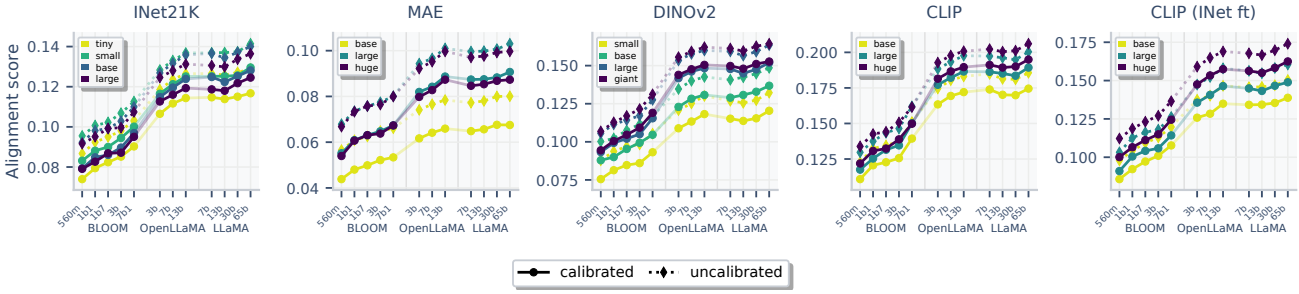
Figure 27. **Sensitivity to α for CKA RBF.** The same pattern holds: no convergence trend at any significance level.



(a) $\alpha = 0.01$



(b) $\alpha = 0.05$ (default)



(c) $\alpha = 0.10$

Figure 28. **Sensitivity to α for mKNN ($k = 10$).** Local alignment and its scaling trend persist across all significance levels.