THE WITTGENSTEINIAN REPRESENTATION HYPOTHESIS: IS LANGUAGE THE EFFECTIVE REFERENCE GEOMETRY OF MULTIMODAL ALIGNMENT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding why independently trained neural networks from different modalities exhibit shared representational structures, and what governs this cross-modal alignment, remains an open question in representation learning. Existing evidence relies on symmetric similarity measures, which can detect modal alignment but are structurally blind to its direction. We introduce directional modal alignment analysis using cycle-kNN, an asymmetric alignment measure, and apply it to dozens of independently trained unimodal models spanning point clouds, vision, and language. We uncover a consistent directional asymmetry: non-language modalities align with the neighborhood structure of language significantly more than language aligns with theirs. This pattern holds across all model families and scales, yet remains entirely invisible to symmetric measures. Mechanistic analysis traces this directionality to feature-density asymmetry, whereby language representations occupy the most compact regions of representational space. The Information Bottleneck framework provides a principled interpretation: optimization under compression promotes alignment with the discrete, compositional structures characteristic of language. We formalize this insight as the Wittgensteinian Representation Hypothesis: the semantic structure of language constitutes the effective reference geometry for cross-modal representation alignment.

1 INTRODUCTION

Understanding why representations from different modalities tend to converge, and identifying the endpoint of such convergence, is a fundamental question in multimodal representation learning. The Platonic Representation Hypothesis (PRH) (Huh et al., 2024) posits that independently trained models converge toward shared representational structures, an observation reinforced by theoretical work (Lu et al., 2025; Ziyin et al., 2025; Tasker et al., 2026) and empirical demonstrations of cross-modal alignment without paired data (Schnaus et al., 2025; Maniparambil et al., 2025).

However, **all empirical evidence relies on symmetric similarity measures**—CKA (Kornblith et al., 2019), mutual kNN (Huh et al., 2024), RSA (Kriegeskorte et al., 2008)—which by construction satisfy $s(\mathbf{X}, \mathbf{Y}) = s(\mathbf{Y}, \mathbf{X})$. No prior work employs an asymmetric measure to investigate the *direction* of this alignment (Klabunde et al., 2025). Like a ball rolling into a valley, if representation convergence possesses directional structure, symmetric measures are fundamentally blind to it.

We fill this gap with **directional convergence analysis** (Figure 1). Our key insight is that CYCLE-KNN—a known alignment measure whose inherent asymmetry has been consistently overlooked (Huh et al., 2024)—naturally encodes the direction of alignment. We systematically evaluate 50 independently trained unimodal models spanning three modalities along an **abstraction hierarchy**: 3D point clouds $\rightarrow$ 2D vision $\rightarrow$ language. Our core findings:

- **Alignment has direction.** CYCLE-KNN reveals consistent asymmetry: vision and point cloud representations approach language significantly more than the reverse—invisible to symmetric measures on the same pairs.
- **Language shows maximal intra-modality consensus**, consistent with proximity to the effective reference geometry.

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

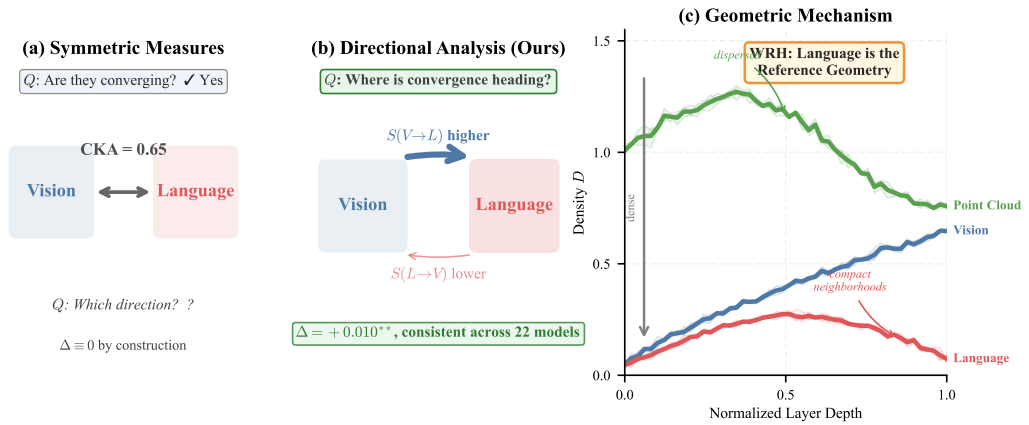


Figure 1: **Overview of directional convergence analysis.** (a) Symmetric measures (e.g., CKA) detect convergence but cannot reveal its direction. (b) CYCLE-KNN is inherently asymmetric; analyzing both directions reveals a consistent directional bias: vision representations approach language more than the reverse ($\Delta = +0.010$, $p < 0.05$, across all 16 model pairs). (c) Three modalities span an abstraction hierarchy; representational compactness (D) increases toward language. WRH posits that the most compact, semantically coherent modality—language, among those studied here—provides the effective reference geometry for alignment.

- **Directionality is scale-invariant.** $\Delta > 0$ toward language holds for every vision model (16/16), every point cloud model (5/5), and nearly every language model across ten families spanning four orders of magnitude.
- **Feature density asymmetry as geometric mechanism.** Language representations are among the most compact; the Information Bottleneck framework provides a principled interpretation.

These findings carry broad implications. If convergence is directional, then the effective reference geometry is not a generic shared structure but the most compact, semantically coherent modality—language, among those we study—consistent with the discrete, compositional nature of language acting as a natural compression target. This sharpens a question left open by PRH: if representations converge, *where*? Our analysis shows convergence is not mutual but asymmetric, following the abstraction hierarchy, and robust across ten model families spanning four orders of magnitude.

Methodologically, our work highlights a systematic blind spot: the exclusive reliance on symmetric measures in representation comparison has left the direction of alignment completely unexamined. We do not propose a new metric; rather, we recognize that CYCLE-KNN's asymmetry—catalogued by Huh et al. (2024) but never analyzed—encodes directional structure, and we build an analysis framework to extract it. This "directional convergence analysis" framework is useful independently of any specific hypothesis about the reference geometry.

Based on these findings, we propose the **Wittgensteinian Representation Hypothesis** (WRH): *cross-modal representational convergence is directional rather than mutual, and its direction follows an ordering over representational compactness and semantic coherence, with language occupying the extremal position among the modalities we examine.* The name invokes Wittgenstein's dictum "The limits of my language mean the limits of my world" (*Tractatus* 5.6), though we emphasize below that this naming reflects language's position in an empirically observed ordering, not a claim about language's unique ontological status. Unlike the Platonic Representation Hypothesis, which is agnostic about the convergence endpoint, WRH makes a falsifiable directional prediction—and our experiments confirm it across all 50 models.

The remainder of this paper is organized as follows. Section 2 defines CYCLE-KNN and its directional interpretation. Section 3 presents empirical findings on directional asymmetry, intra-modality consensus, and scale invariance. Section 4 provides mechanistic analysis. Section 5 formally states

WRH and positions it within the hypothesis landscape, and Sections 6–7 discuss related work and conclude.

2 PRELIMINARIES

2.1 CYCLE-KNN: AN ASYMMETRIC ALIGNMENT MEASURE

Let $\mathbf{X} \in \mathbb{R}^{N \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{N \times d_2}$ denote representations of N shared stimuli. Standard alignment measures—CKA (Kornblith et al., 2019), mutual kNN (Huh et al., 2024), RSA (Kriegeskorte et al., 2008)—are symmetric by construction: $s(\mathbf{X}, \mathbf{Y}) = s(\mathbf{Y}, \mathbf{X})$. They measure alignment *degree* but cannot reveal its *direction*.

CYCLE-KNN (Huh et al., 2024) computes alignment through a two-step nearest-neighbor cycle:

$$\text{CYCLE-KNN}(\mathbf{X} \to \mathbf{Y}; k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left[i \in \bigcup_{j \in \text{kNN}_{\mathbf{Y}}(i)} \text{kNN}_{\mathbf{X}}(j) \right] \tag{1}$$

where $\text{kNN}_{\mathbf{Y}}(i)$ denotes the *set* of indices of the k nearest neighbors of sample i in space $\mathbf{Y}$ (excluding i itself), $\text{kNN}_{\mathbf{X}}(j)$ the corresponding set in space $\mathbf{X}$, and the union aggregates the return sets of all k first-hop neighbors. The score is the fraction of samples whose two-hop cycle through $\mathbf{Y}$ returns to themselves. **Our key observation**: although Huh et al. (2024) note that CYCLE-KNN is the only asymmetric metric among eight they catalog, its asymmetry has never been analyzed. We show it carries structured directional information.

The asymmetry arises because when the first-hop space $\mathbf{Y}$ has more coherent neighborhoods, semantically similar samples cluster tightly, making the return hop more likely to succeed:

$$\text{CYCLE-KNN}(\mathbf{X} \to \mathbf{Y}) > \text{CYCLE-KNN}(\mathbf{Y} \to \mathbf{X}) \implies \mathbf{Y}\text{'s neighborhoods are more coherent.} \tag{2}$$

We define the **directional gap** $\Delta(\mathbf{X}, \mathbf{Y}) = \text{CYCLE-KNN}(\mathbf{X} \to \mathbf{Y}) - \text{CYCLE-KNN}(\mathbf{Y} \to \mathbf{X})$. A consistently positive Δ indicates that $\mathbf{Y}$'s structure is a more stable reference frame. Robustness is confirmed through synthetic experiments with matched dimensionality (Appendix B), k-sensitivity analysis ($k = 2 \ldots 50$; at k=1 the gap vanishes by construction, Appendix A), permutation tests, and controls for compactness, text granularity, and layer-selection bias (Appendices C–E).

2.2 EXPERIMENTAL SETUP

Models. We assemble 50 independently trained models spanning three modalities along an *abstraction hierarchy*: 5 **point cloud** models (PointMAE (Pang et al., 2022), PointBERT (Yu et al., 2022), PointGPT (Chen et al., 2023), ReCon++ (Qi et al., 2024); 16M–300M params), 16 **vision** models across supervised, self-supervised, and contrastive paradigms (ViT (Dosovitskiy et al., 2021), MAE (He et al., 2022), DINOv2 (Oquab et al., 2024), DINOv3 (Siméoni et al., 2025)), and 29 **language** models from five families (BLOOMZ (Muennighoff et al., 2023), LLaMA (Touvron et al., 2023), Qwen2.5 (Qwen Team, 2025a), Qwen3 (Qwen Team, 2025b), InternLM (Cai et al., 2024); 560M–72B). Full details are in Table 10 (Appendix G).

Abstraction hierarchy. We arrange the three modalities along a putative abstraction hierarchy as an organizing principle for the experimental design, motivated by each data source: **3D point clouds** encode raw physical geometry; **2D vision** projects 3D reality onto a plane, introducing viewpoint-dependent abstraction; **language**, as a symbolic system, abstracts low-level perceptual detail into discrete tokens. This arrangement is a design choice rather than an empirical claim about modalities; all directional findings are reported per modality pair and do not depend on the hierarchy.

Data and features. For vision–language alignment we use 1,024 image–text pairs from WiT (Srinivasan et al., 2021); for point cloud–language and point cloud–vision we use ShapeNet55 (Chang et al., 2015) with matched 3D objects, rendered images, and text descriptions. We extract per-layer feature matrices $\mathbf{X}_l \in \mathbb{R}^{N \times d}$ from every model, L2-normalized before cosine-similarity nearest-neighbor search.

Table 1: **Directional asymmetry across modality pairs.** Mean ± std of CYCLE-KNN in each direction ($k=10$), computed over all model pairs. The directional gap $\Delta > 0$ consistently indicates convergence toward the more abstract modality. Permutation test ($n=1000$) confirms significance. Symmetric measures (CKA, mutual kNN) show $\Delta \approx 0$ on the same pairs.

Direction Pair	CYCLE-KNN ($A \to B$)	CYCLE-KNN ($B \to A$)	Δ	p-value
Point Cloud → Language	0.628 ± 0.091	0.598 ± 0.098	+0.030	<0.01
Vision → Language	0.577 ± 0.075	0.566 ± 0.072	+0.010	<0.05
Point Cloud → Vision	0.720 ± 0.134	0.674 ± 0.133	+0.046	<0.01
Symmetric controls (same model pairs):				
CKA: Vis-Lang	0.410 ± 0.066		≈ 0	–
Mutual kNN: Vision–Language	0.385 ± 0.071		≈ 0	–

Metric computation. For each cross-modality model pair (M_x, M_y), we compute CYCLE-KNN in both directions at every layer pair (l_x, l_y), producing an $L_x \times L_y$ score matrix, and report the maximum (best-layer alignment) as the summary statistic. Symmetric measures (CKA, mutual kNN) are computed on the same pairs as controls. Default $k=10$. Best-layer selection is purely post-hoc; Appendix E re-computes Δ under seven layer-selection protocols and bounds the selection bias at ± 0.008, an order of magnitude below the reported effects.

3 Directional Convergence: Empirical Analysis

This section presents our core empirical findings. We demonstrate that representation convergence is not symmetric—it has a consistent direction pointing toward language.

3.1 Directional Asymmetry: Cycle-kNN Reveals the Direction of Alignment

We compute CYCLE-KNN in both directions for all cross-modality model pairs across the six directional combinations: language → vision, vision → language, language → point cloud, point cloud → language, vision → point cloud, and point cloud → vision (Figure 2).

Finding 1. *Across all model pairs, convergence exhibits a consistent directional asymmetry: non-language modalities align toward language more than language aligns toward them.*

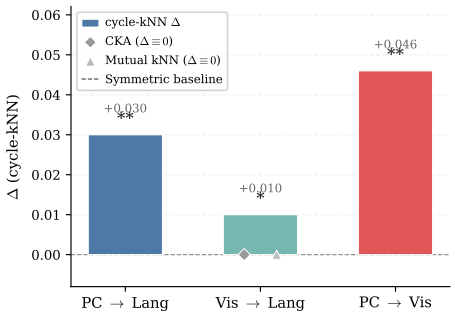


Figure 2: **Directional asymmetry across modality pairs.** Directional gap Δ = CYCLE-KNN$(A \to B)$ − CYCLE-KNN$(B \to A)$ per combination ($k = 10$); all $\Delta > 0$. Gray: symmetric measures (CKA, mutual kNN), $\Delta \equiv 0$. $^{**}p < 0.01$, $^{*}p < 0.05$ (permutation test, $n=1000$).

Robustness. We verify that the directional asymmetry is robust across $k = 2, 3, 5, 10, 20, 50$—the sign of Δ is consistent across all $k \geq 2$ (at k=1 the gap vanishes by construction; Appendix A). Permutation tests ($n=1000$) confirm statistical significance ($p < 0.01$ for all modality pairs; Table 1). On the same model pairs, CKA and mutual kNN yield $\Delta \approx 0$, confirming that directionality is invisible to symmetric analysis. The full pairwise score matrices (Figure 3) confirm this is systematic across all 638 model pairs. Three families of controls further rule out alternative explanations: compactness-only and semantics-destroyed controls (Appendix C), text-granularity controls (Appendix D), and layer-selection protocol analysis (Appendix E); an independent Information Imbalance comparison is provided in Appendix F.

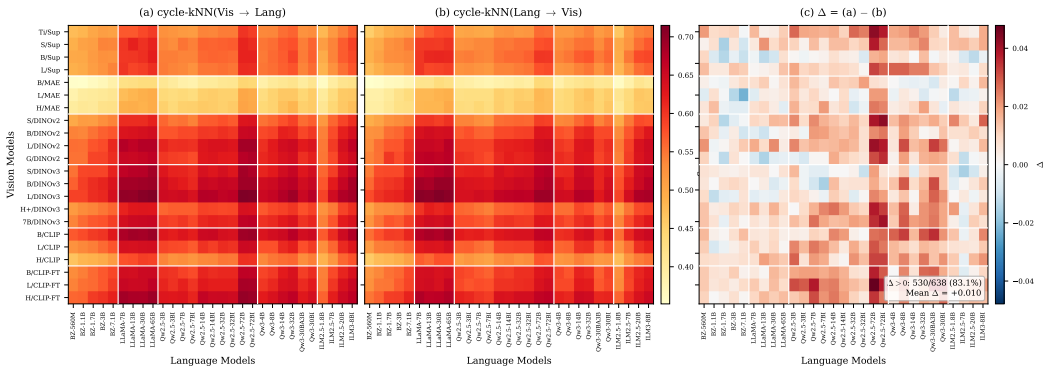


Figure 3: **Systematic directional asymmetry across all model pairs.** (a) CYCLE-KNN (Vision $\rightarrow$ Language) and (b) CYCLE-KNN (Language $\rightarrow$ Vision) score matrices for all 16 vision $\times$ 29 language model pairs ($k = 10$, WiT-1024 dataset). Both panels share the same color scale. Panel (a) is systematically brighter than (b), confirming that Language neighborhoods are more coherent when probed from Vision. (c) Element-wise difference Δ_{ij}: 530/638 pairs (83.1%) have $\Delta > 0$ (mean $\Delta = +0.010$), ruling out the asymmetry being driven by outlier models.

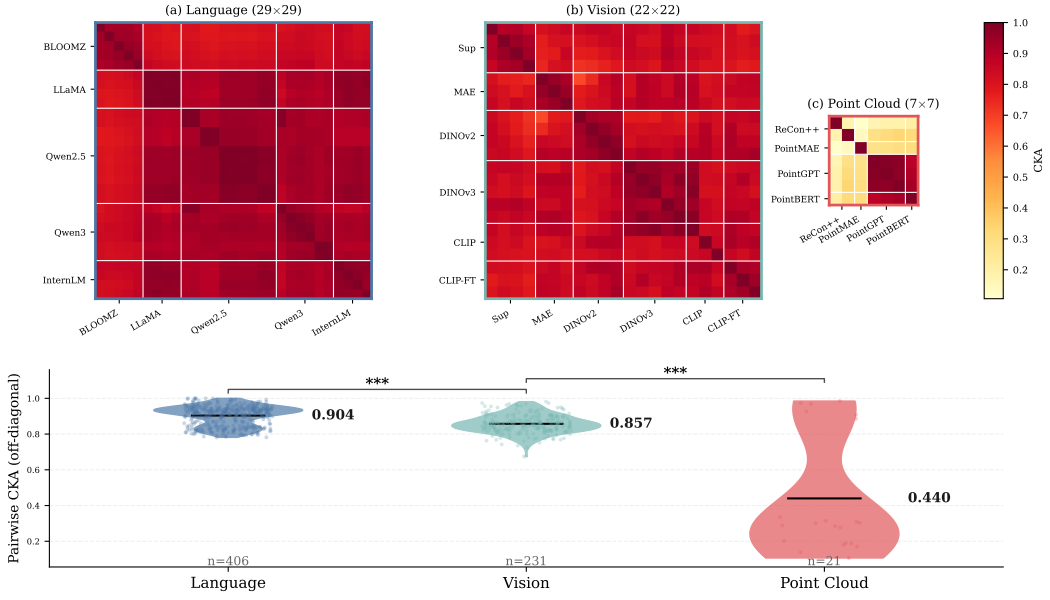


Figure 4: **Intra-modality representational consensus.** Top: pairwise CKA heatmaps (shared color scale). Bottom: violin plots confirm ordering Language $>$ Vision $>$ Point Cloud ($p < 0.001$, Mann–Whitney U). Consistent with language providing the effective reference geometry.

3.2 INTRA-MODALITY CONSENSUS: LANGUAGE REPRESENTATIONS AGREE MOST

Finding 2. *Language models exhibit the highest intra-modality representational consensus (mean CKA = 0.904), followed by vision (0.857), with point cloud showing the lowest (0.440).*

We compute pairwise alignment within each modality using both CKA and mutual kNN. Figure 4 and Table 2 show the results: language models exhibit uniformly high agreement, while point cloud models show high variance—consistent with language providing the effective reference geometry, around which representations show the least variance.

Table 2: **Intra-modality consensus.** Mean ± std of pairwise alignment within each modality.

Modality	#Pairs	CKA	Mutual kNN
Language	406	0.904 ± 0.055	0.562 ± 0.100
Vision	231	0.857 ± 0.051	0.502 ± 0.079
Point Cloud	21	0.440 ± 0.324	0.248 ± 0.236

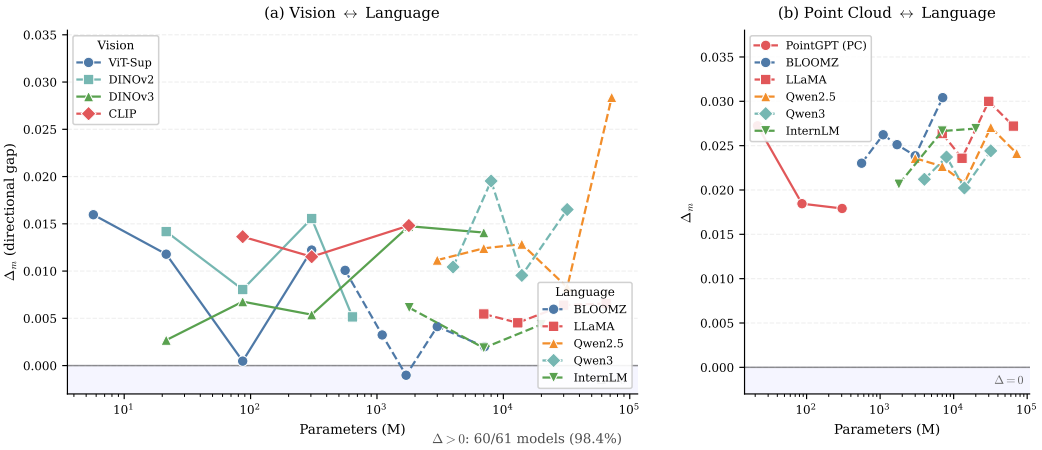


Figure 5: **Scale-invariant directionality.** Per-model Δ_m vs. parameter count for ten model families. (a) Vision↔Language. (b) PC↔Language. 60/61 combinations (98.4%) have $\Delta > 0$, confirming scale-invariance.

3.3 Directionality Is Robust Across Scales

Finding 3. *The directional asymmetry $\Delta > 0$ toward language holds for every vision model (16/16), every point cloud model (5/5), and nearly every language model (28/29 for Vision–Language, 29/29 for PC–Language), regardless of scale, architecture, or training paradigm.*

For each model m of modality A, we compute $\Delta_m = \overline{\text{cyc-kNN}}(A_m \to B) - \overline{\text{cyc-kNN}}(B \to A_m)$ averaged over all target-modality models. Figure 5 plots Δ_m vs. parameter count for ten families spanning four orders of magnitude (5.7M–72B). While neural scaling laws (Kaplan et al., 2020) predict smooth capability gains with scale, the directional asymmetry shows no such dependence—Δ remains positive regardless of parameter count.

4 Why Language? Mechanistic Analysis

Having established the empirical phenomenon, we investigate its underlying mechanism. We propose that the directionality is associated with **feature density asymmetry** and that this asymmetry has a principled interpretation under the Information Bottleneck framework.

4.1 Feature Density Differs Across Modalities

We measure representational compactness using pairwise mean distance:

$$D(\mathbf{X}) = \frac{2}{N(N-1)} \sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \tag{3}$$

where $\|\cdot\|_2$ denotes the ℓ_2 (Euclidean) norm; all nearest-neighbor computations likewise use standard Euclidean distance on ℓ_2-normalized features. Lower D indicates more compact representations (samples cluster more tightly).

Finding 4. *At alignment-optimal layers, language representations are among the most compact. Layer-wise density profiles reveal characteristic modality-specific patterns that correlate with the observed directional asymmetry.*

Specifically, vision models show D increasing monotonically across layers (early layers are dense, later layers sparse); language models follow an inverted-U trajectory where D first increases then decreases, reaching minimal density at intermediate-to-late layers—precisely the layers where cross-modal alignment peaks (Figure 6). These modality-specific patterns are consistent across model families, suggesting they reflect intrinsic properties of the input modality rather than architecture-specific artifacts.

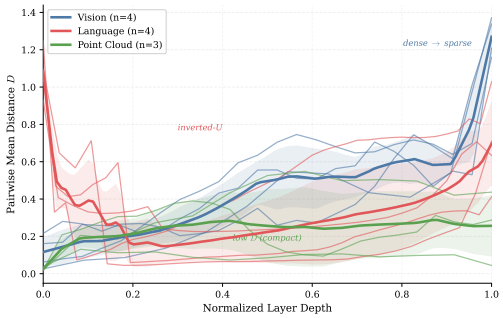


Figure 6: **Layer-wise representational density.** Pairwise mean distance D (ℓ_2-normalised features) vs. normalised depth. Vision (blue): monotonically increasing D; language (red): inverted-U, most compact at later layers; point cloud (green): variable. Bold: per-modality mean; shaded: ± 1 s.d.

The key observation: language models achieve maximal compactness at the same layers where cross-modal alignment is strongest, suggesting a geometric link—the coherent neighborhood structure of compact representations facilitates cycle-kNN return, producing the observed directional asymmetry. Crucially, this mechanism is *general*, not specific to language: in the point-cloud/vision pair, where neither side is language, we still observe a stable gap oriented toward the more compact modality (vision; $\Delta = +0.046$, positive for all 5 point-cloud models, Table 1). Controlled synthetic experiments (Appendix B) confirm that density asymmetry alone suffices to produce CYCLE-KNN directionality in matched-dimensionality settings, with Δ growing as the density gap widens; compactness and semantic-correspondence controls (Appendix C) show that compactness *without* semantic correspondence produces no directional gap, and that pathological collapse yields a qualitatively different signature from the modest, stable Δ we report.

4.2 Information Bottleneck: An Interpretive Account

The Information Bottleneck (IB) framework (Tishby et al., 2000; Tishby & Zaslavsky, 2015) offers an *interpretive lens*—not an independently tested empirical claim—for why language representations tend to be compact. Its objective trades off compression of the input against preservation of task-relevant information; as compression strengthens, optimal representations undergo phase transitions toward increasingly discrete, combinatorial structures (Strouse & Schwab, 2017), the structural form of natural language. On this reading, language is a highly optimized compression of perceptual experience: lower representational entropy manifests geometrically as tighter clustering (lower D) and predicts $\Delta > 0$ when language is the first-hop reference space. This is a *post-hoc* account consistent with our cross-sectional observations, not a dynamical claim about training trajectories; testing it directly would require trajectory-level evidence, which we leave to future work. The account also connects to independent theoretical lines: neural thermodynamics (Ziyin et al., 2025); universal statistical structure in natural datasets (Levi & Oz, 2025); Schrodi et al. (2025), who show information imbalance drives the modality gap and representation structure; IB regularization that improves multimodal alignment (Almudévar et al., 2025); and Gui et al. (2025), who show contrastive learning adapts to the intrinsic dimension of shared latent variables.

5 The Wittgensteinian Representation Hypothesis

Drawing together the empirical findings (Section 3) and mechanistic analysis (Section 4), we formally state:

> **The Wittgensteinian Representation Hypothesis (WRH).** *Among independently trained unimodal models, cross-modal representational convergence is directional rather than mutual: its direction is governed by an ordering over representational*

> *compactness and semantic coherence, with the more compressed, semantically coherent modality serving as the effective reference geometry for the others. Language occupies the extremal position among the three modalities examined here, which motivates the name.*

As noted in Section 1, the name borrows Wittgenstein's language-and-world metaphor (Wittgenstein, 1921) as a mnemonic for language's extremal position in this ordering; it implies no commitment about language's unique ontological status. Table 3 positions WRH within the hypothesis landscape.

Table 3: Comparison of representation convergence hypotheses. WRH characterizes convergence as directional, governed by an ordering over compactness and semantic coherence; its reference modality is the most compressed, semantically coherent one—language here, not a uniquely privileged endpoint.

Hypothesis	Endpoint	Direction	Scope	Method
PRH (Huh et al., 2024)	Stat. of reality	–	Inter-model	Symmetric
Sem. Hub (Wu et al., 2025)	Lang. internal hub	Intra-model	Intra-model	Causal int.
Indra (Lu et al., 2025)	Relational struct.	–	Inter-model	Symmetric
UNE (Tasker et al., 2026)	Gaussian geometry	–	Inter-model	Symmetric
Neural Thermo. (Ziyin et al., 2025)	Symmetry-constr.	–	Inter-model	Theoretical
WRH (Ours)	**Compact struct.**	**Toward compact**	**Inter-model**	**Asymmetric**

WRH differs from PRH in providing an *operationally testable* endpoint—the most compressed, semantically coherent modality—rather than an unobservable "statistical structure of reality"; from Semantic Hub (Wu et al., 2025) in addressing inter-model (not intra-model) convergence; and from Indra/UNE in employing asymmetric rather than symmetric tools.

Scope. We do not claim language subsumes perceptual information: perception carries information not captured by words, and non-linguistic agents can form useful world representations directly from perceptual input. Language's symbolic structure—which abstracts low-level detail while carrying high-level semantics for humans (Harnad, 1990; Lambon Ralph et al., 2017)—is thus *one* compressed view of the world (LeCun, 2022), not the only or final one.

Empirical support. Four converging lines of evidence support WRH: (1) directional asymmetry toward language across all three modality pairs (Section 3.1); (2) highest intra-modality consensus for language (Section 3.2); (3) scale-invariance across 50 models spanning four orders of magnitude (Section 3.3); and (4) a mechanistic link via feature-density asymmetry (Section 4). A supplementary directional-CKA analysis is consistent on the vision–language axis, with known caveats when feature dimensions differ (Appendix H.7).

6 Related Work

Representation convergence. The PRH (Huh et al., 2024) posits convergence toward a shared statistical model of reality, with formalizations including Indra representations (Lu et al., 2025), Universal Normal Embeddings (Tasker et al., 2026), Neural Thermodynamics (Ziyin et al., 2025), and learning-theoretic analyses (Insulla et al., 2025; Luthra et al., 2026). Empirically, frozen encoders can be bridged by lightweight projections (Maniparambil et al., 2025; Gröger et al., 2025), blind matching achieves non-trivial accuracy without paired data (Schnaus et al., 2025; Yacobi et al., 2025), and generative pretraining produces implicit alignment (Xiao et al., 2025). All rely exclusively on symmetric measures, leaving alignment direction unexamined.

Representation compatibility. Backward-compatible representation learning (Shen et al., 2020; Zhou et al., 2023; Biondi et al., 2026) formalizes whether *same-class samples stay close and different-class samples remain separated* across spaces—a class-level, inherently symmetric criterion. CYCLE-KNN differs in two ways. First, *granularity*: it operates at the instance level (can a specific sample's neighborhood be recovered through the other space?) rather than the class level—a strictly stronger condition that implies class-level compatibility but not vice versa. Second, *directionality*:

compatibility is symmetric by construction, whereas CYCLE-KNN is asymmetric for $k \geq 2$, enabling the directional analysis that compatibility-style criteria cannot provide.

Asymmetric and directional analysis. Acevedo et al. (2025) applied Information Imbalance (Glielmo et al., 2022) to independently trained models, observing text→vision asymmetry but without a directional hypothesis or mechanistic explanation. The Semantic Hub Hypothesis (Wu et al., 2025) describes language as a routing hub *within* multimodal models—complementary to our inter-model analysis. Recent surveys (Koutoupis et al., 2026; Tjandrasuwita et al., 2025) note the open question of which modality should anchor alignment; we provide the first systematic evidence.

Boundary conditions. Convergence is not unconditional: alignment depends on shared information structure (Tjandrasuwita et al., 2025), training objective is the primary determinant of representational consistency (Ciernik et al., 2025), and early-training perturbations can permanently alter representations (Kwok et al., 2025). Distributional closeness does not guarantee representational similarity (Nielsen et al., 2025; Smith et al., 2025), and Kabra et al. (2026) find near-random cross-modal similarity for certain pairs—underscoring the value of direct geometric measurement via CYCLE-KNN.

7 Discussion and Conclusion

Discussion. The Platonic Representation Hypothesis established that independently trained models converge but left *where* they converge unanswered—obscured by the exclusive use of symmetric measures. Directional convergence analysis shows that alignment is not symmetric: it follows an ordering over representational compactness and semantic coherence, with the most compressed, semantically coherent modality—language, among those we study—acting as the reference geometry. Methodologically, rather than proposing a new metric, we recognize that CYCLE-KNN's asymmetry encodes directional structure and build a framework to extract it.

Limitations and future work. Our hierarchy covers three modalities and 50 models; extending to audio, tactile, or code would strengthen generality, and the IB account remains interpretive, not a formal proof. CYCLE-KNN captures instance-level neighborhood recoverability, which is inherently local; concept-level geometry analyses (class-centroid distances, between-concept margins) remain a priority, though our collapse controls (Appendix C) already show concept-level structure alone yields a qualitatively different signature. Whether a modality more compressed than language could displace it as the reference geometry remains open; our exclusive use of unimodal models mitigates data-leakage concerns. Future work includes multi-metric validation (cf. Appendix F), tracking directionality during training, and formal links between IB optimality and linguistic structure.

AI Use Statement

AI-based tools were used only for auxiliary purposes: polishing prose, assisting with figure drafting, and debugging code. The CYCLE-KNN measure, the analysis pipeline, and all reported results were designed, implemented, and verified by the authors.

Reproducibility Statement

Anonymized analysis code and configuration files are provided as supplementary material for review; a public release will follow upon acceptance.

References

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images. *arXiv preprint arXiv:2505.17101*, 2025.

Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. Aligning multimodal representations through an information bottleneck. In *International Conference on Machine Learning (ICML)*, 2025.

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. A stationary (and therefore compatible) representation is all you need. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2411.05561.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gabor Csanyi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS Nexus*, 1(2), 2022.

Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With limited data for multimodal alignment, let the structure guide you. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.16895.

Yu Gui, Cong Ma, and Zongming Ma. Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.12473.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3): 335–346, 1990.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

Francesco Insulla, Shuo Huang, and Lorenzo Rosasco. Towards a learning theory of representation alignment. *arXiv preprint arXiv:2502.14047*, 2025.

Rishabh Kabra, Maks Ovsjanikov, Drew A. Hudson, Ye Xia, Skanda Koppula, et al. A mixed diet makes dino an omnivorous vision encoder. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2602.24181.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Stefanos Koutoupis, Michaela Areti Zervou, Konstantinos Kontras, Maarten De Vos, Panagiotis Tsakalides, and Grigorios Tsagkatakis. The more, the merrier: Contrastive fusion for higher-order multimodal alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2511.21331.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Devin Kwok, Gül Sena Altıntaş, Colin Raffel, and David Rolnick. The butterfly effect: Neural network training trajectories are highly sensitive to initial conditions. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.13234.

Matthew A. Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18:42–55, 2017.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022.

Noam Itzhak Levi and Yaron Oz. The underlying universal statistical structure of natural datasets. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2510.05197.

Jianglin Lu, Hailing Wang, Kuo Yang, Yitian Zhang, Simon Jenni, and Yun Fu. The indra representation hypothesis for multimodal alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Achleshwar Luthra, Priyadarsi Mishra, and Tomer Galanti. On the alignment between supervised and self-supervised contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.08852.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E. O'Connor. Harnessing frozen unimodal encoders for flexible multimodal alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2023.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. 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 (ICLR)*, 2021.

Beatrix Miranda Ginn Nielsen, Emanuele Marconato, et al. When does closeness in distribution imply representational similarity? an identifiability perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.03784.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, 2022.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision (ECCV)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025a.

Qwen Team. Qwen3 technical report. *arXiv preprint*, 2025b.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2503.24129.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: Examining model stitching. In *International Conference on Machine Learning (ICML)*, 2025.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

DJ Strouse and David J Schwab. The deterministic information bottleneck. *Neural Computation*, 29 (6):1611–1630, 2017.

Chen Tasker, Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. The universal normal embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *International Conference on Machine Learning (ICML)*, 2025.

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

Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Kegan Paul, 1921.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *arXiv preprint arXiv:2411.04986*, 2025.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.11693.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. Bt^2: Backward-compatible training with basis transformation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Liu Ziyin, Yizhou Xu, and Isaac Chuang. Neural thermodynamics: Entropic forces in deep and universal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Appendix

A CYCLE-KNN IS INHERENTLY ASYMMETRIC

We prove that CYCLE-KNN is asymmetric for $k \geq 2$, the regime used throughout this paper (k=10).

Definition. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N \times d}$ be two representation matrices of N shared stimuli. For $k \geq 1$:

$$\text{CYCLE-KNN}(\mathbf{X} \to \mathbf{Y}; k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[i \in \bigcup_{j \in \text{kNN}_{\mathbf{Y}}(i)} \text{kNN}_{\mathbf{X}}(j) \right] \tag{4}$$

where $\text{kNN}_{\mathbf{Y}}(i)$ denotes the k nearest neighbors of point i in space $\mathbf{Y}$ (excluding i itself), and similarly for $\text{kNN}_{\mathbf{X}}$.

Proposition (symmetry at k=1). For k=1 with unique nearest neighbors, $\text{CYCLE-KNN}(\mathbf{X}\to\mathbf{Y}; 1) = \text{CYCLE-KNN}(\mathbf{Y}\to\mathbf{X}; 1)$ always holds.

Proof. Let $A = \{i : \text{NN}_{\mathbf{X}}(\text{NN}_{\mathbf{Y}}(i)) = i\}$ and $B = \{i : \text{NN}_{\mathbf{Y}}(\text{NN}_{\mathbf{X}}(i)) = i\}$. Define $\varphi : A \to B$ by $\varphi(i) = \text{NN}_{\mathbf{Y}}(i)$. If $i \in A$, let $j = \text{NN}_{\mathbf{Y}}(i)$; then $\text{NN}_{\mathbf{X}}(j) = i$, so $\text{NN}_{\mathbf{Y}}(\text{NN}_{\mathbf{X}}(j)) = \text{NN}_{\mathbf{Y}}(i) = j$, hence $j \in B$. Injectivity: if $\varphi(i_1) = \varphi(i_2) = j$, then $\text{NN}_{\mathbf{X}}(j) = i_1 = i_2$. Surjectivity: for $j \in B$, set $i = \text{NN}_{\mathbf{X}}(j)$; then $\text{NN}_{\mathbf{Y}}(i) = j$ and $\text{NN}_{\mathbf{X}}(j) = i$, so $i \in A$ with $\varphi(i) = j$. Thus φ is a bijection, $|A| = |B|$, and the two scores coincide. □

Theorem (asymmetry at $k \geq 2$). For $k \geq 2$, there exist configurations where $\text{CYCLE-KNN}(\mathbf{X}\to\mathbf{Y}; k) \neq \text{CYCLE-KNN}(\mathbf{Y}\to\mathbf{X}; k)$.

Proof by construction. Let N=6, d=1, k=2 with $\mathbf{X} = (15, 26, 49, 60, 87, 90)^\top$ and $\mathbf{Y} = (34, 56, 58, 57, 63, 37)^\top$. All pairwise distances in both spaces are distinct at the k/k+1 boundary, eliminating tie-breaking ambiguity.

Direction $\mathbf{X} \to \mathbf{Y}$ (k=2): For each i, we compute $S_i = 2\text{NN}_{\mathbf{Y}}(i)$ and the return set $R_i = \bigcup_{j \in S_i} 2\text{NN}_{\mathbf{X}}(j)$:

$$\begin{aligned}
&i{=}1 : S_1 = \{6, 2\},\ R_1 = 2\text{NN}_{\mathbf{X}}(6) \cup 2\text{NN}_{\mathbf{X}}(2) = \{5, 4\} \cup \{1, 3\} = \{1, 3, 4, 5\} \ni 1\ \checkmark \\
&i{=}2 : S_2 = \{4, 3\},\ R_2 = \{3, 5\} \cup \{4, 2\} = \{2, 3, 4, 5\} \ni 2\ \checkmark \\
&i{=}3 : S_3 = \{4, 2\},\ R_3 = \{3, 5\} \cup \{1, 3\} = \{1, 3, 5\} \ni 3\ \checkmark \\
&i{=}4 : S_4 = \{2, 3\},\ R_4 = \{1, 3\} \cup \{4, 2\} = \{1, 2, 3, 4\} \ni 4\ \checkmark \\
&i{=}5 : S_5 = \{3, 4\},\ R_5 = \{4, 2\} \cup \{3, 5\} = \{2, 3, 4, 5\} \ni 5\ \checkmark \\
&i{=}6 : S_6 = \{1, 2\},\ R_6 = \{2, 3\} \cup \{1, 3\} = \{1, 2, 3\} \not\ni 6\ \times
\end{aligned}$$

Score: $\text{CYCLE-KNN}(\mathbf{X}\to\mathbf{Y}; 2) = 5/6$.

Direction $\mathbf{Y} \to \mathbf{X}$ (k=2):

$$\begin{aligned}
&i{=}1 : S_1 = \{2, 3\},\ R_1 = 2\text{NN}_{\mathbf{Y}}(2) \cup 2\text{NN}_{\mathbf{Y}}(3) = \{4, 3\} \cup \{4, 2\} = \{2, 3, 4\} \not\ni 1\ \times \\
&i{=}2 : S_2 = \{1, 3\},\ R_2 = \{6, 2\} \cup \{4, 2\} = \{2, 4, 6\} \ni 2\ \checkmark \\
&i{=}3 : S_3 = \{4, 2\},\ R_3 = \{2, 3\} \cup \{4, 3\} = \{2, 3, 4\} \ni 3\ \checkmark \\
&i{=}4 : S_4 = \{3, 5\},\ R_4 = \{4, 2\} \cup \{3, 4\} = \{2, 3, 4\} \ni 4\ \checkmark \\
&i{=}5 : S_5 = \{6, 4\},\ R_5 = \{1, 2\} \cup \{2, 3\} = \{1, 2, 3\} \not\ni 5\ \times \\
&i{=}6 : S_6 = \{5, 4\},\ R_6 = \{3, 4\} \cup \{2, 3\} = \{2, 3, 4\} \not\ni 6\ \times
\end{aligned}$$

Score: $\text{CYCLE-KNN}(\mathbf{Y}\to\mathbf{X}; 2) = 3/6 = 1/2$.

Since $5/6 \neq 1/2$, the metric is asymmetric. □

Why $k \geq 2$ breaks symmetry. The bijection proof for k=1 relies on the fact that NN returns a *single* point, creating a one-to-one pairing between success sets. For $k \geq 2$, the union $\bigcup_{j \in S_i} \text{kNN}_{\mathbf{X}}(j)$ aggregates return paths from multiple intermediate points. A space with compact clusters (small

pairwise distances, like $\mathbf{X}$ above where $\{87, 90\}$ and $\{15, 26\}$ form tight pairs) concentrates its k-neighborhoods on nearby points, creating redundant return paths that benefit queries *originating* in the other space. Conversely, a space with dispersed structure (like $\mathbf{Y}$ above where points 2–5 span only the range $[56, 63]$) offers fewer distinct return paths. This structural asymmetry—which space provides more coherent neighborhoods for the return hop—is the geometric basis of the directional convergence signal exploited throughout this paper (using k=10, as confirmed empirically in Appendix B).

B Synthetic Experiment Details

Synthetic generators. We use eight generators to produce paired feature spaces $(\mathbf{X}, \mathbf{Y})$ with controlled density asymmetry: (1) **Gaussian clusters**—C clusters with shared centers, variance ratio controls density difference; (2) **Concentric rings**—points on rings of varying radii, ring width controls density; (3) **Uniform grid**—points on a regular grid with Gaussian jitter, jitter magnitude controls density; (4) **Uniform disk**—points uniformly sampled from a disk, radius ratio controls density; (5) **Swiss roll**—points on a Swiss roll manifold, noise amplitude controls density; (6) **S-curve**—points on an S-curve, scaling factor controls density; (7) **Folded manifold**—a flat manifold with controlled folding, fold amplitude controls local density; (8) **High-dimensional Gaussian**—isotropic Gaussians in $d = 128$ dimensions, variance ratio controls density.

Protocol. For each generator, we produce 20 density ratios $\rho \in [1.0, 5.0]$ (where $\rho = D(\mathbf{Y})/D(\mathbf{X}) > 1$ means $\mathbf{X}$ is more compact). Both spaces share $N = 1000$ samples, $d = 128$ dimensions, and semantic labels (same point indices correspond to matched samples). We compute CYCLE-KNN in both directions with $k \in \{1, 5, 10, 20\}$ and report Δ as a function of ρ. Figure 7 shows the results.

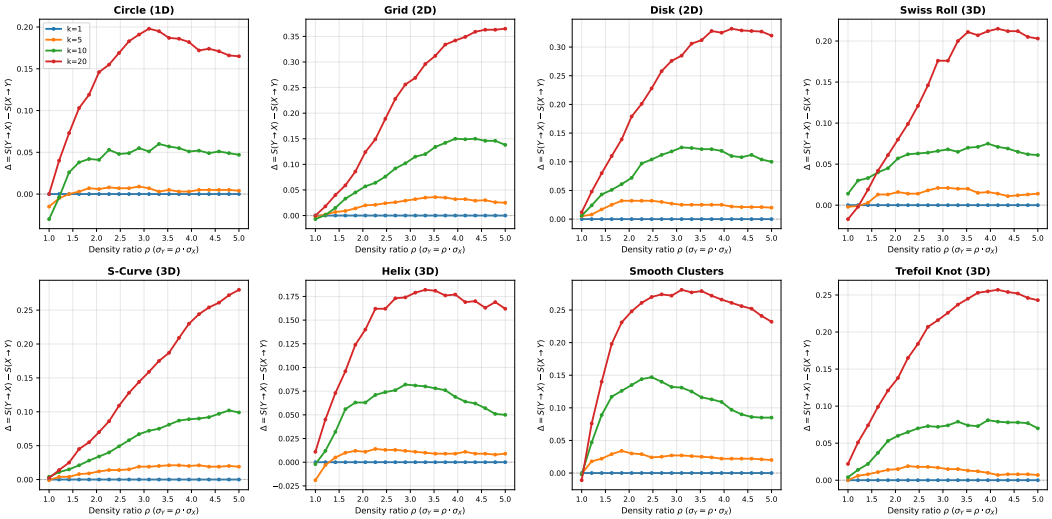


Figure 7: **Synthetic validation: Δ increases monotonically with density ratio ρ.** Each panel shows a different manifold generator (8 types spanning 1D–3D intrinsic dimensionality). X is a compact reference (σ_{base} noise) and Y is dispersed with noise scaled by $\rho \in [1, 5]$. For every $k \geq 5$, $\Delta = S(Y \to X) - S(X \to Y) > 0$ once $\rho > 1$ and increases monotonically, validating that CYCLE-KNN correctly detects asymmetric neighborhood density. Higher k values amplify the gap. At k=1 the two directions coincide by construction under unique nearest neighbors (Appendix A), so the k=1 curve carries no directional information and is shown only for reference.

C Compactness and Semantic-Correspondence Controls

A natural concern is that the directional gap Δ might be explained by compactness or density alone, rather than by semantically coherent neighborhood structure. We ran two families of controls, each on one representative model pair per modality pair.

(i) Pathological compactness does not mimic our effect. We collapse each representation to its class centroids (a concept-level operation that maximizes within-class compactness) and, separately, project to a random 1-D subspace. Table 4 reports the resulting forward/backward scores and gaps. Collapse inflates $|\Delta|$ to ≈ 0.2—an order of magnitude above the real effects—while *crashing both absolute scores* (to 0.42–0.66, far below the real cross-modal scores); random 1-D projection crashes both scores to ≤ 0.11. In contrast, the modal directionality reported in the main text arises under high representation similarity (mean absolute scores 0.57–0.80 across modality pairs; Tables 1 and 12), not under collapsed or degenerate representations.

Table 4: **Pathological controls** (cycle-kNN fwd / bwd $\rightarrow \Delta$) on representative model pairs. Collapse to class centroids inflates $|\Delta|$ (≈ 0.2, $10\times$ above real effects) while crashing both absolute scores; random 1-D projection crashes both scores to ≤ 0.11.

Condition	PC↔Lang	PC↔Vis	Lang↔Vis
Collapse to class centroids	0.62 / 0.42 → +0.192	0.61 / 0.44 → +0.175	0.66 / 0.45 → +0.210
Random 1-D projection	0.11 / 0.02 → +0.084	0.09 / 0.01 → +0.077	0.09 / 0.01 → +0.078

(ii) Removing semantics kills Δ. We destroy semantic correspondence while preserving compactness: (a) row-permutation, which shuffles the sample-to-feature correspondence and preserves the pairwise-distance distribution D exactly; and (b) D-matched random Gaussians, which replace each representation with a random Gaussian matched in dimension and density. Table 5 shows that all semantics-destroyed conditions fall within ± 0.004—negligible in magnitude and absent of any consistent direction. Compactness without semantic correspondence produces no directional gap. This is compatible with our interpretation: language provides a neighborhood structure that is simultaneously compact *and* semantically coherent (measured by p_{sem}, the fraction of k-NN neighbors sharing the ground-truth category: language $p_{\text{sem}} \approx 0.7$ vs. vision $p_{\text{sem}} \approx 0.2$–$0.5$ at the same D), which is precisely what an effective reference geometry requires.

Table 5: **Semantics-destroyed controls** (Δ, mean $\pm$ std) on the same representative model pairs. Row-permutation preserves D exactly; D-matched random Gaussians match density. All conditions fall within ± 0.004.

Condition	PC↔Lang	PC↔Vis	Lang↔Vis
Row-permutation (D preserved)	+0.002 $\pm$ 0.003	+0.001 $\pm$ 0.004	+0.004 $\pm$ 0.001
D-matched random Gaussians	+0.003 $\pm$ 0.002	+0.002 $\pm$ 0.003	+0.001 $\pm$ 0.003

Implications for the density mechanism. These controls also make explicit an assumption implicit in the density-based mechanism of Section 4: the link from compactness to CYCLE-KNN directionality presupposes that the representations involved come from well-trained models whose neighborhoods are semantically coherent. For unconstrained representations, compactness alone does not imply semantic coherence, and Tables 4–5 confirm that compactness without semantic correspondence yields no stable directional gap.

D Text Granularity Controls

A related concern is that the language-side compactness might be an artifact of coarse, category-level text (e.g., one-word class names), which would collapse many stimuli onto identical or near-identical captions. The ShapeNet55-34 text used in the paper consists of instance-level Cap3D captions (Luo et al., 2023): 1,024 unique captions with a mean length of 49.9 words. To verify that the directional

effect is not an artifact of text granularity, we constructed a granularity ladder from full descriptions down to bare class names. Table 6 shows an example for one ShapeNet chair instance; Table 7 reports the directional gap at each level across all 1,024 instances.

Table 6: **Granularity ladder** (example from one ShapeNet chair instance).

Level	Example text	Words
Full description (used in paper)	"A modern office chair featuring a plush green fabric upholstery, providing comfort and support. It has a high backrest and cushioned seat…"	∼50
First sentence	"A modern office chair featuring a plush green fabric upholstery, providing comfort and support."	∼17
Brief description	"A modern office chair"	∼4
Class name	"chair"	1

Table 7: **Directional gap across text granularity levels** (all 1,024 ShapeNet55-34 instances). Δ is stable from full descriptions to brief phrases (50→5 words: change -0.005, within the noise floor) and inflates only at bare 1-word class names ($+0.179$, $10\times$ above the paper's operating point), which are never used.

Level	Mean Δ	frac($\Delta > 0$)	Words
Full description (used in paper)	$\mathbf{+0.017}$	1.00	50
First sentence	$+0.019$	1.00	17
Brief description	$+0.012$	0.77	5
Class name	$+0.179$	0.05	1

From the full description to the brief description (50→5 words), Δ changes by only -0.005—within the noise floor. The effect inflates ($+0.179$) only at bare 1-word category names, which lie far outside the paper's operating point; notably, the fraction of pairs with $\Delta > 0$ drops to 0.05 there, showing that category-level text produces a qualitatively different (and pathological) signature rather than a strengthened version of the reported effect.

E Layer-Selection Protocol Analysis

The paper's summary statistic selects, for each direction independently, the best layer pair. To bound any bias introduced by this double selection, we re-computed Δ under seven layer-selection protocols (pure post-hoc re-analysis, no re-training). Let $S_{\text{fwd}}[i,j]$ and $S_{\text{bwd}}[j,i]$ be the CYCLE-KNN scores at layer pair (i,j) in each direction, and $D[i,j] = S_{\text{fwd}}[i,j] - S_{\text{bwd}}[j,i]$ the fixed-pair gap:

- **paper**: each direction independently picks its best layer pair (two selection steps, as in the main text);
- **fixed_fwd**: evaluate D at the forward direction's best pair (favors $\Delta > 0$);
- **fixed_bwd**: evaluate D at the backward direction's best pair (least favorable to $\Delta > 0$);
- **fixed_sym**: evaluate D at the shared best pair of the symmetric average $(S_{\text{fwd}} + S_{\text{bwd}}^{\top})/2$ (direction-neutral, one selection step);
- **fixed_last**: evaluate D at the true last layers (no selection at all);
- **all_mean / all_median**: mean and median of D over all real layer pairs (no selection).

A same-modality null control (Lang↔Lang, where no cross-modal gap exists) bounds the double-selection bias at ± 0.008—an order of magnitude below the reported effects (0.02–0.05). Table 8 reports the mean Δ and the fraction of pairs with $\Delta > 0$ under all seven protocols. All groups remain significant ($p < 10^{-6}$) under every protocol.

Table 8: **Mean Δ (and % of pairs with $\Delta > 0$) under seven layer-selection protocols.** All groups remain significant ($p < 10^{-6}$) under every protocol.

Group (n)	paper	fixed_fwd	fixed_bwd	fixed_sym	fixed_last	all_mean	all_median
PC→Lang (203)	+0.024, 100%	+0.035, 100%	+0.014, 86%	+0.024, 100%	+0.020, 97%	+0.030, 100%	+0.030, 100%
PC→Vis (154)	+0.041, 100%	+0.047, 100%	+0.033, 100%	+0.041, 100%	+0.043, 100%	+0.024, 100%	+0.021, 100%
Vis→Lang (638)	+0.010, 83%	+0.018, 93%	+0.002, 57%	+0.010, 80%	+0.017, 85%	+0.008, 98%	+0.007, 98%

F Information Imbalance Comparison

To check that the directional signal is not an artifact of CYCLE-KNN specifically, we compared Δ with Information Imbalance (II) (Glielmo et al., 2022) on identical inputs: the same 1,024 ShapeNet55-34 samples, under the direction-neutral `fixed_sym` layer protocol (Appendix E). The two metrics measure different quantities—CYCLE-KNN measures *neighborhood topology recoverability*, whereas II measures *rank averaging*—so their consistency profiles diverge by pair type (Table 9).

Table 9: **Consistency by pair type under the `fixed_sym` protocol**: fraction of pairs with a positive directional signal for CYCLE-KNN ($\Delta > 0$) vs. II ($\Delta_{\text{II}} > 0$).

Pair type	CYCLE-KNN: frac($\Delta > 0$)	II: frac($\Delta_{\text{II}} > 0$)
PC↔Language	**100%**	50%
Vision↔Language	83%	82%

Crucially, an independent team's II experiment on a different dataset (Flickr30k) gives the *same directional signal* as our CYCLE-KNN on the vision–language axis: Acevedo et al. (2025) report II(DeepSeek-V3 $\rightarrow$ DinoV2) ≈ 0.10 vs. the reverse ≈ 0.30, indicating that text representations are more predictive of visual representations than the opposite direction. Together with the metric-difference analysis above, this constitutes independent-dataset, independent-metric, independent-team validation of the language-ward directional alignment.

G Implementation Details

G.1 Feature Extraction

CYCLE-KNN computation. For each model pair, features are extracted at every layer. Features are L2-normalized before computing cosine similarity matrices. Nearest neighbors are found via exhaustive search (no approximate NN). $k = 10$ is used by default; full k-sensitivity results are in Appendix H.

Language models. For autoregressive language models, we extract hidden states at every layer and aggregate them via mean pooling over token positions. For each layer l, the feature matrix is $\mathbf{X}_l \in \mathbb{R}^{N \times d_{\text{hidden}}}$.

Vision models. For ViT-based models, we extract the [CLS] token representation at each transformer block. For models without a [CLS] token (e.g., MAE), we use spatial average pooling.

Point cloud models. For point cloud transformers, we extract the global representation at each block (typically obtained via average pooling over point tokens).

G.2 Evaluation Dataset

WiT-1024. We use N=1,024 image–text pairs from the Wikipedia-based Image Text (WiT) dataset (Srinivasan et al., 2021). Images are resized to 224×224 pixels (matching ViT input resolution). Text descriptions are tokenized using each language model's native tokenizer with a maximum sequence length of 128 tokens; representations are extracted via mean pooling over token positions (language) or [CLS] token (vision).

Table 10: **Complete model list.** All 50 models with individual parameter counts, grouped by modality along the abstraction hierarchy. Models marked with † use cross-modal supervision; all others are purely unimodal. *MoE architecture (30B total, 3B active).

Modality	Series	Model	#Params (M)	Objective
Point Cloud	PointMAE	PointMAE-S	16	Masked point modeling
	PointBERT	PointBERT-S	32	Masked point modeling
	PointGPT	PointGPT-S	21	Autoregressive
		PointGPT-B	42	
		PointGPT-L	150	
Vision	ViT-Sup	ViT-T/16	5.7	Supervised
		ViT-S/16	22	
		ViT-B/16	86	
		ViT-L/16	304	
	MAE	MAE-B/16	86	Masked imaging
		MAE-L/16	304	
		MAE-H/16	632	
	DINOv2	DINOv2-S/14	21	Self-distillation
		DINOv2-B/14	86	
		DINOv2-L/14	304	
		DINOv2-G/14	1,100	
	DINOv3	DINOv3-S/16	22	Self-distillation
		DINOv3-B/16	86	
		DINOv3-L/16	304	
		DINOv3-H/16	632	
		DINOv3-7B/16	7,000	
Language	BLOOMZ	BLOOMZ-560M	560	Causal LM
		BLOOMZ-1.1B	1,100	
		BLOOMZ-1.7B	1,700	
		BLOOMZ-3B	3,000	
		BLOOMZ-7.1B	7,100	
	LLaMA	LLaMA-7B	7,000	Causal LM
		LLaMA-13B	13,000	
		LLaMA-30B	30,000	
		LLaMA-65B	65,000	
	Qwen2.5	Qwen2.5-3B	3,000	Causal LM
		Qwen2.5-3B-Inst	3,000	
		Qwen2.5-7B	7,000	
		Qwen2.5-7B-Inst	7,000	
		Qwen2.5-14B	14,000	
		Qwen2.5-14B-Inst	14,000	
		Qwen2.5-32B	32,000	
		Qwen2.5-32B-Inst	32,000	
		Qwen2.5-72B	72,000	
		Qwen2.5-72B-Inst	72,000	
	Qwen3	Qwen3-4B	4,000	Causal LM
		Qwen3-8B	8,000	
		Qwen3-14B	14,000	
		Qwen3-32B	32,000	
		Qwen3-30B-A3B*	30,000	
		Qwen3-30B-A3B-Inst*	30,000	
	InternLM	InternLM2.5-1.8B	1,800	Causal LM
		InternLM2.5-7B	7,000	
		InternLM2.5-20B	20,000	
		InternLM3-8B-Inst	8,000	

ShapeNet. For point cloud experiments, we use $N{=}1{,}024$ samples drawn from ShapeNet (Chang et al., 2015). Each object is represented in three modalities: (i) a 3D point cloud with 8,192 points obtained via farthest-point sampling, (ii) a single rendered 2D image from a canonical viewpoint at $224{\times}224$ resolution, and (iii) an instance-level text caption generated by Cap3D (Luo et al., 2023) (mean 49.9 words; all 1,024 captions are unique). Appendix D reports granularity controls showing that the directional effect is stable from full descriptions to brief phrases and inflates only at bare category names, which are never used.

Stimulus alignment. For both datasets, the i-th sample across all modalities corresponds to the same semantic entity (image–caption pair for WiT; object instance for ShapeNet), enabling meaningful cross-modal nearest-neighbor computation with shared indices.

G.3 Hardware and Computation

All experiments were conducted on a single DGX Station equipped with 4× NVIDIA A800-SXM4 GPUs (80 GB each), 128 CPU cores, and 503 GB system memory. Feature extraction for all 50 models required approximately 120 GPU-hours, dominated by the largest language models: Qwen2.5-72B (∼18h per variant across all layers) and LLaMA-65B (∼14h). Vision models and point cloud models were substantially faster (<1h each). Alignment metric computation (cycle-kNN, CKA, mutual kNN, and dCKA across all model pairs and layer combinations) required approximately 40 GPU-hours total. The entire experimental pipeline, including feature extraction, metric computation, and analysis, completed within one week of wall-clock time on the single node.

H Additional Experimental Results

H.1 Full Cross-Modality Alignment Scores

Table 11 reports per-model directional alignment on the WiT dataset, and Table 12 reports the corresponding results on ShapeNet.

Table 11: **Per-model directional alignment summary (cycle-kNN, k=10).** For each model, we report the mean alignment score when that model serves as source (*fwd*) vs. target (*bwd*), averaged across all partners in the opposing modality. Δ = fwd − bwd.

(a) Language models (WiT, vs. 16 vision models)

Model	L→V	V→L	Δ
Qwen2.5-14B-Inst	0.499	0.509	−0.010
Qwen2.5-14B	0.516	0.519	−0.003
Qwen2.5-32B-Inst	0.532	0.531	+0.001
Qwen2.5-32B	0.537	0.541	−0.004
Qwen2.5-3B-Inst	0.553	0.555	−0.002
Qwen2.5-3B	0.609	0.615	−0.005
Qwen2.5-72B-Inst	0.612	0.616	−0.005
Qwen2.5-72B	0.613	0.620	−0.006
Qwen2.5-7B-Inst	0.619	0.626	−0.007
Qwen2.5-7B	0.553	0.568	−0.014
Qwen3-14B	0.553	0.561	−0.008
Qwen3-30B-A3B-Inst	0.549	0.563	−0.014
Qwen3-30B-A3B	0.544	0.555	−0.011
Qwen3-32B	0.568	0.582	−0.014
Qwen3-4B	0.570	0.581	−0.011
Qwen3-8B	0.573	0.583	−0.010
BLOOMZ-1b1	0.574	0.581	−0.007
BLOOMZ-1b7	0.603	0.633	−0.030
BLOOMZ-3b	0.604	0.631	−0.027
BLOOMZ-560m	0.555	0.565	−0.010
BLOOMZ-7b1	0.555	0.575	−0.020
LLaMA-13b	0.582	0.591	−0.010
LLaMA-30b	0.584	0.601	−0.017
LLaMA-65b	0.552	0.569	−0.017
LLaMA-7b	0.567	0.582	−0.015
InternLM2.5-1.8b	0.509	0.515	−0.006
InternLM2.5-20b	0.552	0.554	−0.002
InternLM2.5-7b	0.585	0.589	−0.004
InternLM3-8B-Inst	0.594	0.606	−0.012
Mean	0.566	0.576	−0.010

(b) Vision models (WiT, vs. 29 language models)

Model	V→L	L→V	Δ
DINOv3-7B/16	0.538	0.522	+0.016
DINOv2-B/14	0.560	0.548	+0.012
ViT-B/16-Sup	0.558	0.558	+0.000
MAE-B/16	0.552	0.539	+0.012
CLIP-B/16-ft	0.402	0.397	+0.005
CLIP-B/16	0.447	0.446	+0.001
DINOv3-B/16	0.455	0.448	+0.007
DINOv2-G/14	0.567	0.553	+0.014
MAE-H/14	0.597	0.589	+0.008
CLIP-H/14-ft	0.625	0.610	+0.016
CLIP-H/14	0.617	0.612	+0.005
DINOv3-H+/16	0.630	0.627	+0.003
CLIP-L/14-ft	0.647	0.640	+0.007
CLIP-L/14	0.660	0.655	+0.005
DINOv2-L/14	0.581	0.566	+0.015
ViT-L/16-Sup	0.609	0.595	+0.014
MAE-L/16	0.641	0.627	+0.014
DINOv3-L/16	0.594	0.582	+0.012
DINOv2-S/14	0.542	0.527	+0.015
ViT-S/16-Sup	0.606	0.592	+0.013
DINOv3-S/16	0.617	0.597	+0.020
ViT-T/16-Sup	0.639	0.625	+0.014
Mean	0.576	0.566	+0.010

CKA is mathematically symmetric—$\text{CKA}(\mathbf{X}, \mathbf{Y}) \equiv \text{CKA}(\mathbf{Y}, \mathbf{X})$—hence $\Delta_{\text{CKA}} = 0$ by construction. The mean cross-modal CKA for Language–Vision pairs on WiT is 0.410 ± 0.066, confirming that models share substantial representational structure even without any shared training signal. The directional asymmetry measured by cycle-kNN is therefore a genuine property of neighborhood geometry, not a measurement artifact.

H.2 Full Intra-Modality Consensus Scores

Table 13 reports the full intra-modality consensus scores for all three modalities.

Table 12: **Per-model directional alignment (ShapeNet, cycle-kNN k=10).** Point cloud models vs. language (29) and vision (16) model pools. $\Delta > 0$ in all cases confirms the WRH directionality prediction: alignment measured from the more abstract modality (PC) is consistently higher.

	PC ↔ Language		**PC ↔ Vision**	
PC Model	PC→L	L→PC	PC→V	V→PC
PointBERT-S	0.810	0.774	0.889	0.847
PointGPT-B	0.815	0.786	0.907	0.860
PointGPT-L	0.417	0.390	0.456	0.424
PointGPT-S	0.699	0.672	0.836	0.786
PointMAE-S	0.680	0.661	0.814	0.777
ReCon++-B	0.679	0.661	0.812	0.773
ReCon++-L	0.724	0.710	0.849	0.811
Mean	0.689	0.665	0.795	0.754
Δ (PC→ − →PC)	**+0.024**		**+0.041**	

Table 13: **Intra-modality representational consensus.** Average pairwise alignment (off-diagonal) within each modality, using both CKA and mutual kNN (k=10). Higher values indicate greater representational convergence among independently trained models. The ordering Language > Vision > Point Cloud is consistent across both metrics, supporting the WRH prediction that language representations provide the most effective reference geometry.

Modality	#Models	#Pairs	CKA	Mutual kNN (k=10)
Language	29	406	0.904 ± 0.055	0.562 ± 0.100
Vision	16	231	0.857 ± 0.051	0.502 ± 0.079
Point Cloud	5	21	0.440 ± 0.324	0.248 ± 0.236

The consensus ordering is robust: CKA ranks Language (0.904) > Vision (0.857) > Point Cloud (0.440), and mutual kNN confirms the same hierarchy (0.562 > 0.502 > 0.248). The notably lower consensus among point cloud models reflects greater architectural diversity in this relatively young field, where models range from masked autoencoders to autoregressive approaches with fundamentally different inductive biases.

H.3 Layer-Pair Alignment Heatmaps

Figure 8 shows the full layer-pair CYCLE-KNN heatmaps for representative model pairs from each cross-modality combination. For each model pair, we compute CYCLE-KNN between all possible layer combinations (source layer × target layer), yielding a 2D score matrix. Several consistent patterns emerge: (1) best alignment occurs at intermediate-to-late layers of both models, suggesting that deeper layers encode progressively more abstract, modality-agnostic structure; (2) the asymmetry $S(A \to B) > S(B \to A)$ is visible across the entire layer grid, not just at the optimal layer pair; (3) language models exhibit a characteristic "hot zone" in their upper layers that persists regardless of the partner modality, consistent with these layers achieving maximal representational compactness.

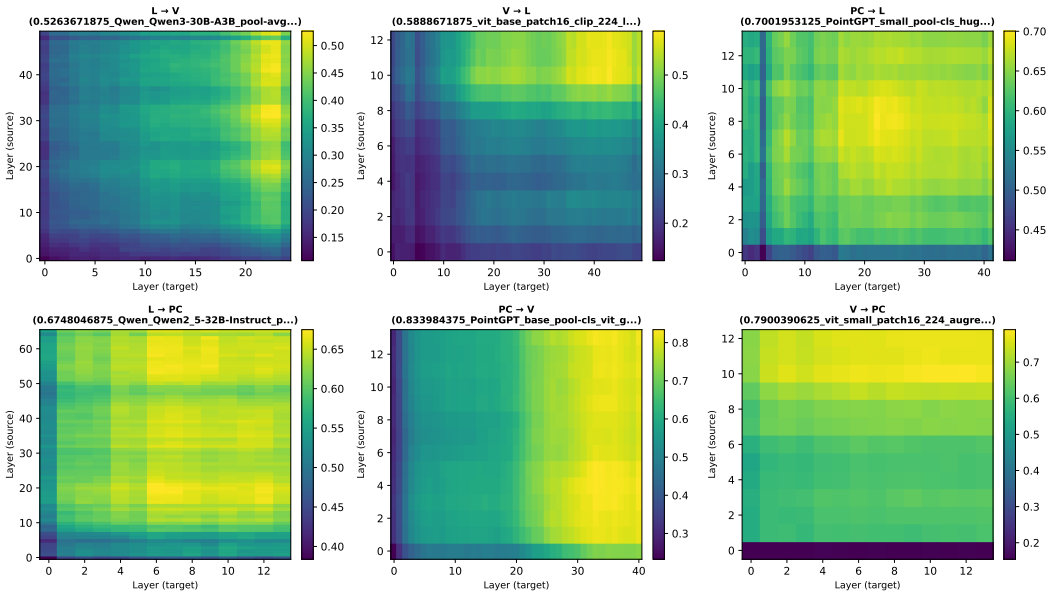


Figure 8: **Layer-pair CYCLE-KNN heatmaps** for representative model pairs from each cross-modality combination. Each panel shows the CYCLE-KNN score (color) for all layer combinations between a source model (y-axis) and a target model (x-axis). Top row: Language→Vision (Qwen2-0.5B → ViT-base), Vision→Language (ViT-base → Qwen2-0.5B), and 3D→Language (PointGPT → Qwen2). Bottom row: Language→3D, 3D→Vision, and Vision→3D. Best alignment typically occurs in the upper layers of both models, consistent with deeper layers encoding more abstract, modality-agnostic representations.

H.4 COMPLETE k-SENSITIVITY ANALYSIS

The choice of k in k-nearest-neighbor methods is often considered a sensitive hyperparameter. To rule out the possibility that our directional findings are an artifact of the default k=10, we systematically vary $k \in \{1, 3, 5, 10, 20, 50\}$ and recompute CYCLE-KNN in both directions for all three cross-modality pairs. Figure 9 shows the results. The sign of Δ remains positive across all $k \geq 2$ and all direction pairs without exception. At k=1, the gap vanishes by construction (the k=1 symmetry proposition in Appendix A: $\Delta \equiv 0$ under unique nearest neighbors), so k=1 is excluded from significance claims. The magnitude of Δ is smallest at small k and stabilizes for $k \geq 5$. Permutation tests confirm $p < 0.05$ at every $k \geq 2$, demonstrating that the observed asymmetry is a robust structural property of the representation spaces rather than a consequence of a particular neighborhood size.

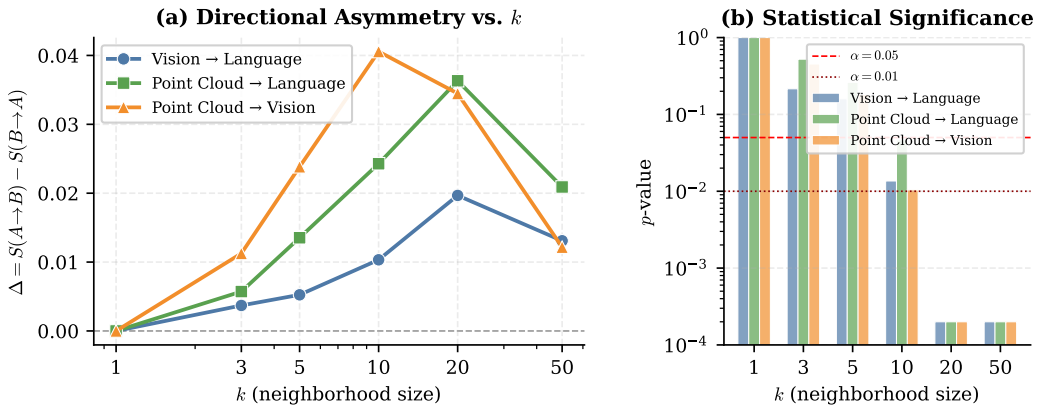


Figure 9: **k-Sensitivity analysis of directional asymmetry.** (a) The directional gap $\Delta = S(A \to B) - S(B \to A)$ remains positive and stable across $k \geq 2$ for all three direction pairs. The sign of Δ never flips, confirming that the observed directionality is not an artifact of the specific neighborhood size; at k=1, $\Delta \equiv 0$ under unique nearest neighbors (Appendix A), so k=1 is excluded from significance claims. (b) Permutation-test p-values remain below 0.05 for all $k \geq 2$, indicating statistical significance.

H.5 Complete Scaling Analysis

A critical question is whether the observed directionality is driven by a few large models or constitutes a universal property. Figure 10 extends the scaling analysis to all model families and directional combinations. For each model m of modality A, we compute the per-model directional gap $\Delta_m = \overline{S}(m \to B) - \overline{S}(B \to m)$ averaged over all partner models in modality B, and plot this against the model's parameter count. In the Vision→Language panel, 16/16 vision models (100%) have $\Delta > 0$; in the PC→Language panel, 5/5 point cloud models have $\Delta > 0$; in the PC→Vision panel, 5/5 models again have $\Delta > 0$. Importantly, no systematic correlation with scale is observed—the smallest models (ViT-T with 5.7M parameters) show comparable Δ to the largest (DINOv3-7B). This confirms that the WRH directionality is a fundamental structural property of the modality representations, not an emergent capability requiring scale.

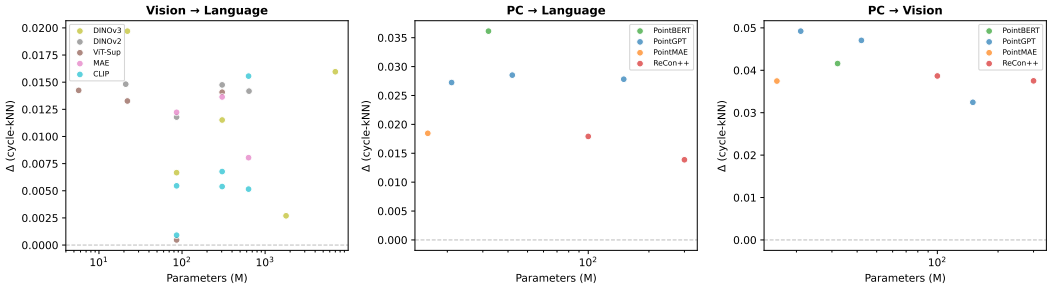


Figure 10: **Per-model directional gap Δ vs. parameter count** for all model families across three directional combinations: Vision→Language (left, 16 vision models), PC→Language (center, 5 point cloud models), and PC→Vision (right, 5 point cloud models). Each point represents one model; the y-axis is $\Delta_m = \overline{S}(m \to \cdot) - \overline{S}(\cdot \to m)$ averaged over all counterpart models. Nearly all points lie above the $\Delta = 0$ dashed line, confirming that the WRH directionality is scale-invariant and not driven by individual large models.

H.6 Density Profiles Across Layers

Figure 11 presents the complete layer-wise density profiles for all 50 models. We compute the pairwise mean distance D (on L2-normalized features) at each layer for every model, then group

curves by modality. Three distinct patterns emerge with high consistency within each modality: (1) **Language models** exhibit an inverted-U trajectory—D increases in early layers (token mixing disperses representations) then decreases in later layers as the model compresses toward task-relevant features, reaching maximum compactness near the final layers; (2) **Vision models** show monotonically increasing D across depth, reflecting progressive spatial abstraction that does not yield the same degree of compression; (3) **Point cloud models** maintain high D throughout with variable patterns, consistent with the geometric diversity of 3D feature learning. The cross-modality density ordering at alignment-optimal layers—$D(\text{language}) < D(\text{vision}) < D(\text{point cloud})$—directly predicts the observed CYCLE-KNN directionality via the density mechanism of Section 4.

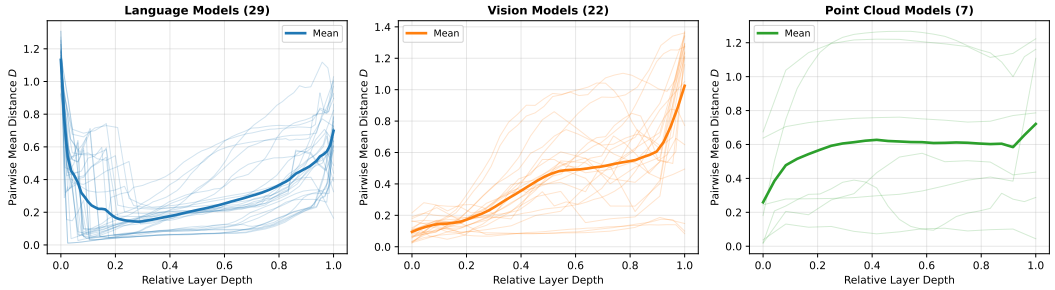


Figure 11: **Layer-wise pairwise mean distance D curves for all models**, grouped by modality: Language (29 models, left), Vision (16 models, center), and Point Cloud (5 models, right). Individual model curves are shown in light color; the bold curve indicates the modality mean. Language models exhibit an inverted-U profile (compression in final layers), vision models show monotonic increase, and point cloud models maintain a high baseline throughout. These distinct density signatures underpin the WRH: denser representations (language) serve as more effective reference geometries for representational alignment.

H.7 Supplementary Analysis: Directional CKA

As a supplementary exploration, we compute directional CKA (dCKA), adapted from Nguyen et al. (2021), which decomposes CKA asymmetrically: $\text{dCKA}(X \to Y) \neq \text{dCKA}(Y \to X)$ in general. We note that dCKA has a known limitation: when representations have mismatched feature dimensions, the asymmetry partly reflects dimensional differences rather than purely structural directionality. The results below should therefore be interpreted with this caveat in mind.

WiT (Language $\leftrightarrow$ Vision). dCKA aligns with the cycle-kNN direction: dCKA(V$\to$L) = 2.07 $>$ dCKA(L$\to$V) = 0.63 (fraction of pairs with positive Δ: 86.2%), consistent with vision representations aligning more readily *toward* language. However, the large dimension gap between vision encoders (typically $\geq$768-d) and language encoders (varying across families) means part of this asymmetry may reflect dimensional rather than structural differences.

ShapeNet (3D $\leftrightarrow$ Vision). dCKA shows the *opposite* direction to cycle-kNN: dCKA(V$\to$PC) = 17.4 $>$ dCKA(PC$\to$V) = 5.6, whereas cycle-kNN finds PC$\to$V $>$ V$\to$PC (Δ = +0.041). This divergence likely reflects both the dimensional sensitivity of dCKA and the fundamentally different properties measured by the two metrics: cycle-kNN captures neighborhood preservation under round-trip mapping (sensitive to local geometry), while dCKA captures variance-weighted similarity (sensitive to global structure). The WiT directional agreement is suggestive but not conclusive given dCKA's dimensional confound; the ShapeNet divergence further underscores that dCKA should be viewed as a supplementary exploration rather than independent confirmation.

I Broader Impact Statement

This work investigates the structural properties of learned representations across modalities. Our findings are primarily of scientific interest, contributing to the understanding of representation learning and multimodal alignment.

Potential positive impacts. Understanding directional convergence could enable more efficient cross-modal transfer learning, reducing computational cost by leveraging language as a natural alignment anchor rather than training expensive symmetric alignment objectives from scratch.

Potential risks. Our finding that language serves as the effective reference geometry should not be interpreted as language being the "correct" or "complete" representation of reality. Language inherently carries cultural biases, and if future systems over-rely on language-centric alignment, they may inherit and amplify these biases. We emphasize that WRH describes a *structural tendency* in current neural network training, not a normative claim about the superiority of any modality.