Spectral Stratification of Semantic Abstraction in Vision-Language Models

Jeonghwan Cheon[1] **Marin Vogelsang**[2] **Lukas Vogelsang**[2] **Pawan Sinha**[2]
[1]Independent Researcher, Seoul, Republic of Korea,
[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
{cheon, ozaki, lvogelsa, psinha}@mit.edu

Abstract

Vision-language models now serve as general-purpose semantic embedding spaces, but how conceptual abstraction is organized within such spaces remains poorly understood. Here, we show that abstraction in these embeddings is spectrally stratified. Specifically, leading principal components encode broad conceptual structure, while later components carry progressively finer-grained distinctions. This pattern holds across multiple contrastive vision-language models and taxonomic datasets. Consistent with this stratification, rank-based interventions produce predictable shifts in retrieval: retaining only the leading components preserves coarse but degrades fine-grained retrieval, while removing them preferentially disrupts abstract category structure. Furthermore, rank-selective routing improves retrieval when the selected subspace matches the target level of abstraction. Finally, compact spectral subspaces that retain only leading components preserve the taxonomic and human-similarity structure of the full embedding and improve agreement with human odd-one-out judgments. Together, these results reveal that vision-language embeddings are not semantically homogeneous. Instead, hierarchical abstraction is ordered along the principal-component spectrum, and human-aligned structure is a recoverable, compact substructure of the full embedding.

1 Introduction

Vision-language models (VLMs) are now widely used for visual recognition and multimodal reasoning. By learning shared embedding spaces for images and language, they enable zero-shot classification and flexible image–text retrieval without task-specific supervision [1–6], building on earlier visual–semantic embedding models [7]. Their embeddings also increasingly serve as general-purpose semantic spaces, for instance to evaluate model–human representational alignment [8, 9]. However, it remains poorly understood how conceptual abstraction is organized within these spaces. Specifically, are broad and fine-grained distinctions distributed uniformly across the embedding, or are they organized along a latent geometric structure?

In human cognition, concepts are commonly described as organized in taxonomic hierarchies in which abstract categories subsume more specific ones [10–12]. The levels of these hierarchies are not equally salient. Basic-level categories such as *dog* occupy a privileged position, balancing within-category similarity against between-category distinctiveness, while superordinate (*animal*) and subordinate (*bulldog*) categories lie above and below them [13, 14]. In neural networks, learning-theoretic and geometric accounts connect semantic hierarchy to the statistical structure of data and concept representations [15, 16]. However, how hierarchical organization is reflected in representations that jointly encode perceptual and linguistic information remains an open question [17, 18].

Contrastive VLMs make this question directly testable, since they are trained on aligned image–text pairs at scale [1, 2, 6]. Different taxonomies, image benchmarks with hierarchical labels, and human

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

similarity judgments allow hierarchical structure in VLM embeddings to be tested across multiple notions of abstraction [19–26]. However, VLM embeddings are often evaluated as single, full-rank representations, leaving open how different components support different levels of abstraction. An exception is recent work that identifies coarse and fine structure in the principal components of CLIP text embeddings [27]. It remains unclear how reliably semantic depth is ordered along the principal-component spectrum, and how selecting spectral subspaces affects behavior.

Here, we address this question by combining analyses of representational geometry with a spectral decomposition of the embedding. Comparing distances among model embeddings with taxonomic or behavioral dissimilarities can show how their similarity structure is organized, beyond what accuracy alone reveals [28–30]. Principal component analysis orders directions of variation by explained variance, so rank-filtered subspaces let us ask which parts of the spectrum support what kinds of semantic structure [31–34]. If broad distinctions concentrate in leading components while fine-grained distinctions require higher-rank ones, the principal-component spectrum becomes an axis for probing and adjusting semantic granularity.

In this work, we show that conceptual abstraction in pretrained VLM embeddings is spectrally stratified. Specifically, leading principal components encode broad conceptual distinctions, while later components carry progressively finer-grained ones. This pattern generalizes across multiple contrastive VLMs and taxonomic datasets. Rank-based modulation changes retrieval predictably: retaining leading components preserves coarse but degrades fine-grained retrieval, and selecting a subspace matched to the target abstraction level improves coarse retrieval. Finally, we define *compact spectral subspaces* as low-rank subspaces that preserve taxonomic structure while truncating high-rank components. These subspaces preserve the taxonomic and human similarity structure of the full embedding and improve agreement with human odd-one-out judgments. Together, these results indicate that VLM embeddings order abstraction levels along the principal-component spectrum, and that human-aligned structure can be recovered from a compact subspace.

2 Preliminaries

2.1 Contrastive vision-language learning

We consider a collection of paired image–text samples $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^{N}$, where each image x_i is accompanied by a corresponding caption or textual description t_i. A contrastive vision-language model employs two separate encoders: an image encoder f_θ and a text encoder g_ϕ. These encoders project visual and linguistic inputs into a common d-dimensional representation space. The normalized embeddings for the image and text modalities are defined as

$$\mathbf{u}_i = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2}, \qquad \mathbf{v}_i = \frac{g_\phi(t_i)}{\|g_\phi(t_i)\|_2},$$

and their cross-modal similarity is measured by the cosine score $s_{ij} = \mathbf{u}_i^\top \mathbf{v}_j$. In line with CLIP-style learning frameworks [1, 2], the model is trained using a symmetric contrastive objective between images and texts. Given a minibatch of size B, the image-to-text and text-to-image losses are expressed as

$$\ell_i^{I\to T} = -\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)}, \qquad \ell_i^{T\to I} = -\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ji}/\tau)},$$

where τ denotes the temperature parameter. The overall training loss is then given by

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{2B} \sum_{i=1}^{B} \left(\ell_i^{I\to T} + \ell_i^{T\to I}\right).$$

This formulation can be viewed as an instance of InfoNCE-based representation learning, where aligned image–text pairs act as positive examples, while the remaining samples within the same minibatch are used as negative examples [35]. By leveraging naturally occurring language supervision at scale, the contrastive objective allows the model to learn shared multimodal representations. Such alignment further supports open-vocabulary recognition, since visual concepts can be associated with their corresponding textual descriptions [1, 2, 7].

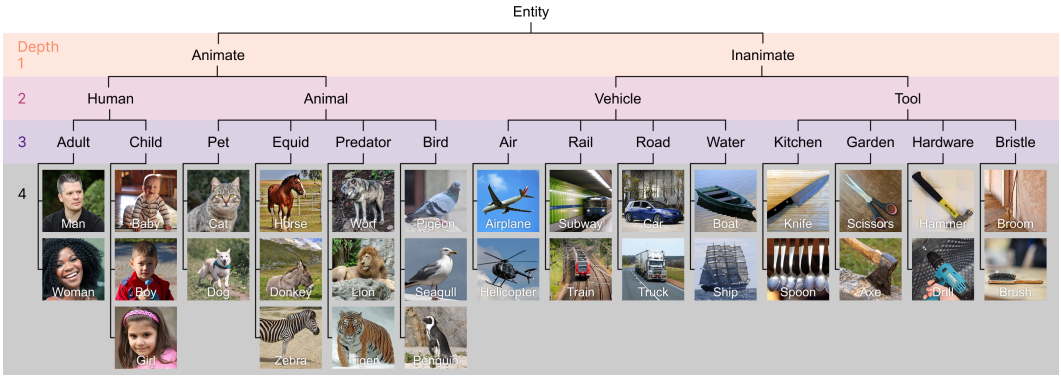


Figure 1: **A four-level probing taxonomy for measuring abstraction in VLM embeddings.** The taxonomy spans four hierarchical levels, from a broad animate–inanimate split at depth 1 to individual object concepts at depth 4. Intermediate levels separate entities into major categories and subcategories, providing a controlled hierarchy for evaluating how VLM embeddings organize concepts from coarse to fine structure.

In our experiments, we consider several commonly used vision-language backbones. CLIP-B/32 and CLIP-L/14 denote ViT-B/32 and ViT-L/14 variants of CLIP, respectively, which learn image–text alignment through large-scale contrastive pretraining [1, 36]. OpenCLIP follows the same training paradigm but is trained in an open-source setting, often using public datasets such as LAION [4, 5]. SigLIP-B/16 uses a ViT-B/16 backbone and replaces the softmax contrastive objective with a pairwise sigmoid loss for improved scalability [6, 36]. Together, these models span differences in scale, patch resolution, training data, and loss formulation.

2.2 Semantic abstraction and categorization

We use *semantic abstraction* to denote the level of generality at which an object is represented or categorized. In cognitive psychology, a concept is typically understood as a mental representation corresponding to a category of entities, allowing novel instances to be recognized, generalized over, and related to prior knowledge [11, 12]. Such categories are often organized taxonomically: a subordinate concept such as *bulldog* is nested under a basic-level concept such as *dog*, which is in turn nested under more abstract superordinate concepts such as *mammal* or *animal*. Abstraction thus varies along the hierarchy, from broad conceptual groupings to finer-grained distinctions among visually and semantically similar categories (Figure 1).

Psychological studies have established that not all levels of this hierarchy are equally salient. Basic-level categories occupy a privileged position because they tend to balance two competing demands: they are specific enough that their members share many perceptual and functional features but, at the same time, are general enough to group many similar instances [12–14]. Superordinate categories include more members, which share fewer diagnostic visual features, whereas subordinate categories are distinguished from one another only by finer details. Within each level, category membership is graded rather than purely definitional: category members differ in typicality, with typical exemplars sharing many properties with other members of the category and fewer properties with contrasting categories [12, 37–39]. Together, differences between levels and graded membership within them motivate evaluating representations by their similarity structure across levels of abstraction.

3 Results

3.1 Semantic abstraction is spectrally stratified in vision-language models

We measure abstraction with a four-level (D1–D4) probing taxonomy (Figure 1), ranging from the coarse animate–inanimate distinction to individual object concepts. Because each image carries a label at every depth, the same set of embeddings can be evaluated against four reference partitions of varying granularity (Figure 2a). This design allows us to test whether VLM embeddings encode different levels of abstraction at different locations along the principal-component spectrum, while controlling for image set and label vocabulary across depths.

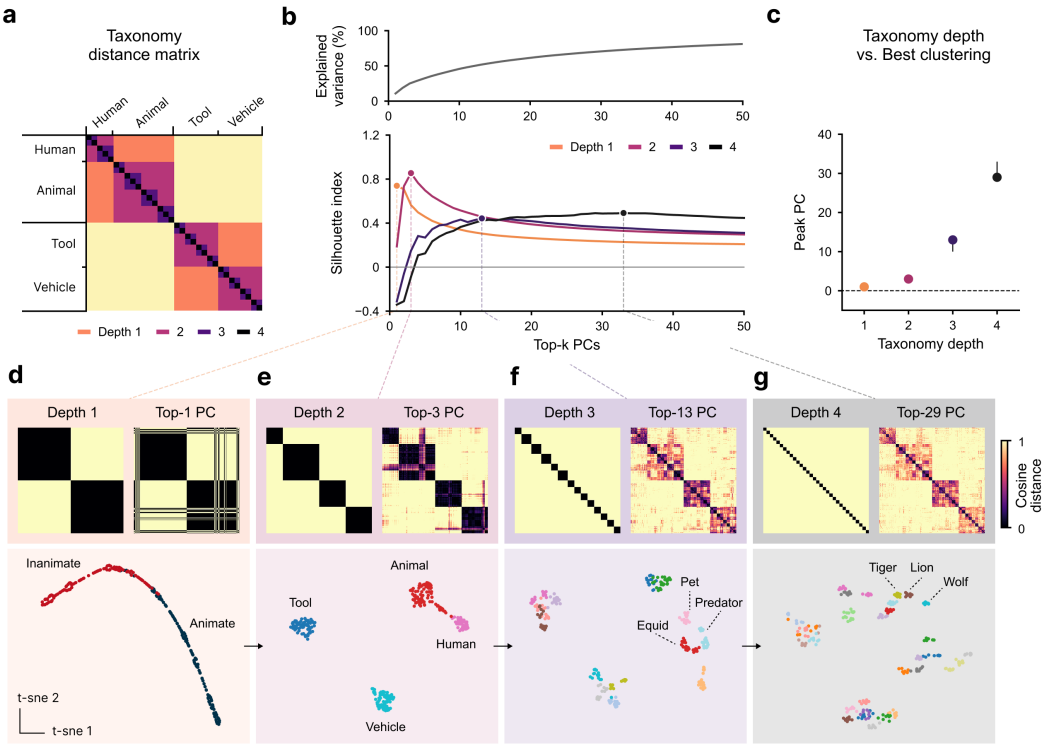


Figure 2: **Category abstraction shifts systematically across the principal-component spectrum.** **a,** Reference taxonomic distance matrix derived from our four-level probing hierarchy. **b,** Top: cumulative explained variance as a function of retained principal components in the CLIP B/32 image embedding. Bottom: silhouette index of taxonomic clusters at depths 1–4 across retained top-k PCs. **c,** Peak silhouette rank as a function of taxonomy depth. **d–g,** Top: representational distance matrices computed from the reference taxonomy (left) and from the rank-truncated embedding at the silhouette-optimal k (right) for depths 1 through 4. Bottom: t-SNE projections of the same rank-truncated embeddings, colored by category labels at each depth. Lower ranks reveal coarse partitions (Animate vs. Inanimate; Human, Animal, Tool, Vehicle), whereas higher ranks expose intermediate and finer-grained category structure.

We then evaluate category structure across rank-truncated subspaces of the CLIP B/32 image embedding. For each retained rank k, we project embeddings onto the top-k PCs and compute the silhouette index of taxonomic clusters at each of the four depths, yielding a curve that quantifies how well each depth is recovered as a function of k. Although cumulative explained variance grows smoothly with k and reaches about 80% only after several tens of components, silhouette curves for different depths peak at sharply distinct ranks (Figure 2b). The shallowest distinction (D1) can be recovered by the very first PC (Figure 2b, $k^* = 1$, 95% CI [1, 2]), while D2 categories peak within the top five PCs (Figure 2b, $k^* = 3$, 95% CI [3, 3]), D3 intermediate groupings near rank 13 (Figure 2b, $k^* = 13$, 95% CI [10, 13]), and D4 categories require about 30 components (Figure 2b, $k^* = 29$, 95% CI [29, 33]). The peak rank thus increases monotonically with taxonomy depth (Figure 2c), indicating that abstraction levels are ordered along the spectrum rather than uniformly distributed across the embedding.

This rank-dependent organization is also visible in the geometry of the embeddings. At the lowest retained ranks, representational distance matrices are dominated by the animate–inanimate partition, mirroring the depth 1 reference matrix. As k increases, finer block structure emerges within these halves. The four depth 2 categories separate at rank 3, intermediate groupings (depth 3), such as Pet, Predator, and Equid, separate at rank 13, and concept-level (depth 4) clusters at rank 29 (Figure 2d–g, top). Low-dimensional t-SNE projections of the same rank-truncated embeddings show a parallel transition (Figure 2d–g, bottom). As the number of retained PCs increases, the coarse Animate–Inanimate distinction progressively resolves into nested subclusters at depths 2 and 3, eventually revealing individual concepts such as Wolf, Lion, and Tiger at depth 4.

To examine whether this ordering emerges from the semantic hierarchy rather than arbitrary grouping, we test a shuffled taxonomy that preserves the tree structure while breaking the concept–label correspondence (Supplementary Results A.1). With the shuffled taxonomy, the previously observed strictly monotonic peak-k^* ordering collapses into near-flat or non-monotonic trajectories, suggesting that our results are not simply an artifact of arbitrary grouping. Moreover, changing branch and leaf composition preserves the coarse-to-fine progression (Supplementary Results A.2), as do removing backgrounds and controlling foreground size (Supplementary Results A.3). The progression also remains evident when every depth is reduced to binary comparisons, which hold the number of categories fixed, and within shared ranges of category separability (Supplementary Results A.4).

To examine the source of this ordering, we use synthetic hierarchies with different covariance structures. A hierarchy alone does not produce reliable ordering. Such ordering emerges when covariance energy decreases from coarse to fine distinctions, and higher-energy semantic directions in VLMs contribute more strongly to leading components (Supplementary Results A.5). Alternative decomposition, such as independent component analysis does not recover the same consistent ordering across depths (Supplementary Results A.6), further tying the progression to the variance-ranked PC spectrum. These results suggest that pretrained VLM embeddings are not semantically homogeneous. Coarse conceptual structure is concentrated in a small set of leading components, which together account for only a modest fraction of variance, whereas progressively finer category distinctions emerge from higher-rank components. The principal-component spectrum can thus be interpreted as a graded axis from broad to fine semantic structure.

3.2 Spectral abstraction generalizes across architectures, modalities, and taxonomies

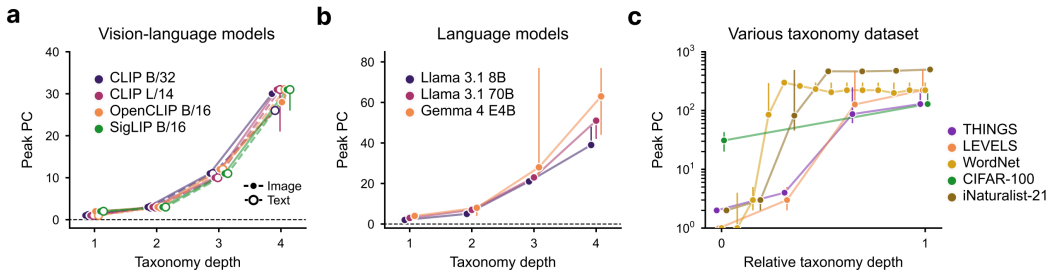


Figure 3: **Spectral abstraction generalizes across architectures, modalities, and taxonomies. a,** Peak silhouette rank as a function of taxonomy depth in image and text embeddings of various VLMs. Filled circles and solid lines denote images. Open circles and dashed lines denote text. Both modalities show increasing peak ranks from coarse to fine categories. **b,** The coarse-to-fine progression also appears in large language models on the same probing taxonomy. Points in **a** and **b** show median ranks across prompt rotations, with bars spanning the minimum and maximum. **c,** Peak silhouette rank across THINGS, LEVELS, WordNet, CIFAR-100, and iNaturalist-21 for CLIP B/32. Relative taxonomy depth places hierarchies of unequal length on a common scale. Points show full-data peaks, with bars indicating 95% bootstrap confidence intervals.

We next ask whether the observed spectral organization is specific to the visual modality of VLMs or generalizes across modalities and models. Across CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16, we find that both image and text embeddings show the same progression of peak rank increasing with taxonomy depth (Figure 3a, median $k^* = 1$–2 at depth 1 and 26–31 at depth 4). Differences in architecture and modality change the absolute peak ranks but preserve their ordering across depths (Supplementary Results A.7). The progression is also recovered using text-derived and joint image–text PCA bases (Supplementary Results A.8). Thus, a coarse-to-fine organization is shared by the visual and linguistic representations within these models. Our spectral analysis also captures hierarchical organization in hierarchy-aware models, including the hyperbolic VLMs (Supplementary Results A.9).

This pattern also extends to language-model representations of the same concepts. In Llama 3.1 8B, Llama 3.1 70B [40], and Gemma 4 E4B [41], broad conceptual distinctions are recovered within the first few components, while finer distinctions require substantially higher ranks (Figure 3b, median $k^* = 2$–4 at depth 1 and 39–63 at depth 4). This progression recurs across variations in wording

and context (Supplementary Results A.10), suggesting that spectral stratification extends beyond the shared image–text embedding space of VLMs. By comparison, vision-only models retain fine-grained category structure but show weaker separation of broad semantic categories (Supplementary Results A.7). This contrast suggests a potential role for language supervision in shaping spectrally organized semantic hierarchies in visual representations.

Spectral organization also generalizes across taxonomic datasets. We repeat the analysis on THINGS, LEVELS, WordNet, CIFAR-100, and iNaturalist-21 [19, 22–26], which differ in category coverage, hierarchy depth, and granularity. Across all five datasets, coarse partitions are recovered at lower ranks than fine-grained categories (Figure 3c). The separation is especially pronounced in WordNet and iNaturalist-21, where the coarsest levels peak within the first few components and finer distinctions require hundreds. The relationship between spectral rank and semantic granularity therefore recurs across hierarchies with markedly different category counts and branching structures. The progression also transfers with fixed, common PCA bases estimated from external ImageNet or THINGS images (Supplementary Results A.11). Together, these results support spectral rank as a common axis for indexing semantic granularity across model architectures, modalities, and taxonomic structures.

3.3 Spectral interventions control abstraction-dependent behavior

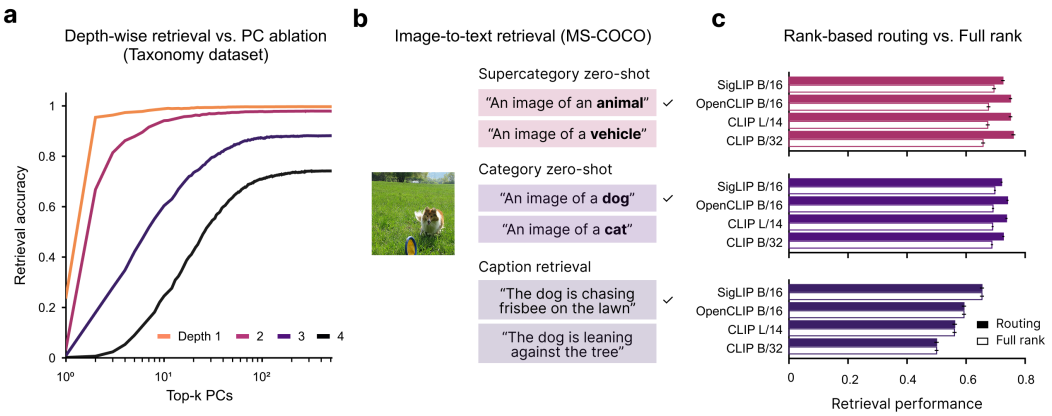


Figure 4: **Rank-based spectral interventions selectively control abstraction-dependent retrieval. a,** Image-to-text retrieval accuracy on the probing taxonomy as a function of retained top-k PCs, evaluated separately at depths 1–4. Coarse retrieval saturates with very few leading components, whereas finer-grained retrieval improves gradually and requires substantially higher ranks. **b,** Schematic of the MS-COCO retrieval task. For each image, candidate captions are constructed at three levels of abstraction: supercategory prompts (e.g., *animal*, *vehicle*), category prompts (e.g., *dog*, *cat*), and full natural-language captions describing the scene. **c,** Retrieval performance on MS-COCO at the supercategory (top), category (middle), and caption (bottom) levels, comparing the full-rank embedding with rank-routed embeddings whose retained subspace is matched to the target abstraction level. Rank-based routing improves retrieval over the full-rank embedding when the selected subspace is matched to the target level of abstraction.

The preceding analyses establish that abstraction is organized across the PC spectrum. To examine how this organization affects model behavior, we ask whether rank-based interventions on the embedding produce predictable shifts in retrieval, evaluating image-to-text retrieval on the probing taxonomy separately at each depth while varying the retained top-k PCs. Retrieval curves are strongly depth-dependent: D1 retrieval saturates near ceiling with only two or three components (Figure 4a, top-1 $= 0.955$ at $k = 2$ vs. max 0.997), D2 retrieval plateaus by approximately the top ten PCs, and D3 and D4 retrieval improve gradually well beyond rank 30 (Figure 4a, depth-4 max 0.742 at $k = 406$). In addition, to localize where in the spectrum each abstraction level resides rather than how many leading components are required to span it, we evaluate each PC band $[k_{\text{lo}}, k_{\text{hi}})$ in isolation (Supplementary Results A.12). The silhouette index at each depth peaks in a strictly later band than its predecessor, indicating that distinct rank ranges support different abstraction levels. Removing leading PCs yields a complementary effect, preferentially weakening coarse category structure in visual embeddings. In language models, leading-PC removal improves fine distinctions among related concepts while weakening broader partitions (Supplementary Results A.10).

To test whether rank selection generalizes beyond our taxonomy, we also evaluate retrieval on MS-COCO with prompts at three abstraction levels: supercategory prompts such as *animal* or *vehicle*, category prompts such as *dog* or *cat*, and full natural-language captions describing the scene (Figure 4b). These targets require progressively more detailed image–text matching, from broad semantic grouping to finer-grained caption semantics. For each abstraction level, we compare the full-rank embedding with a rank-routed embedding whose retained subspace is selected to match the target level, using a held-out validation split to choose k. Across all four backbones, rank routing improves supercategory and category retrieval, whereas caption-level retrieval shows negligible changes relative to the full-rank baseline (Figure 4c, CLIP B/32 i2t supercategory top-1 $0.658 \rightarrow 0.760$, $\Delta = +0.103$), with the largest absolute gains at the supercategory level. These gains in coarse retrieval persist when the PCA basis is fitted to external ImageNet or THINGS images (Supplementary Results A.11).

Thus, the spectral stratification reported in the previous sections is not only descriptive but can also be used to control behavior. Without retraining or modifying model parameters, selecting a subspace matched to the target abstraction level recovers the structure that is relevant for that level while suppressing higher-rank residuals that appear to add noise to coarse retrieval.

3.4 Compact subspaces preserve semantic structure and improve behavioral alignment

The taxonomy and retrieval results suggest that much of the abstract semantic structure in VLM embeddings is concentrated in a compact portion of the spectrum. We therefore ask whether taxonomic and human-similarity structure are preserved after removing high-rank residual components. Indeed, Spearman correlation between embedding distances and the reference taxonomic distance matrix peaks at a low retained rank and declines slightly thereafter (Figure 5a, $k^* = 5$, $\rho = 0.540$ vs. full-rank $\rho = 0.537$). Alignment with human behavioral similarity on THINGS shows a similar profile, with peak correlation at a comparable rank and only marginal change at higher ranks (Figure 5b,

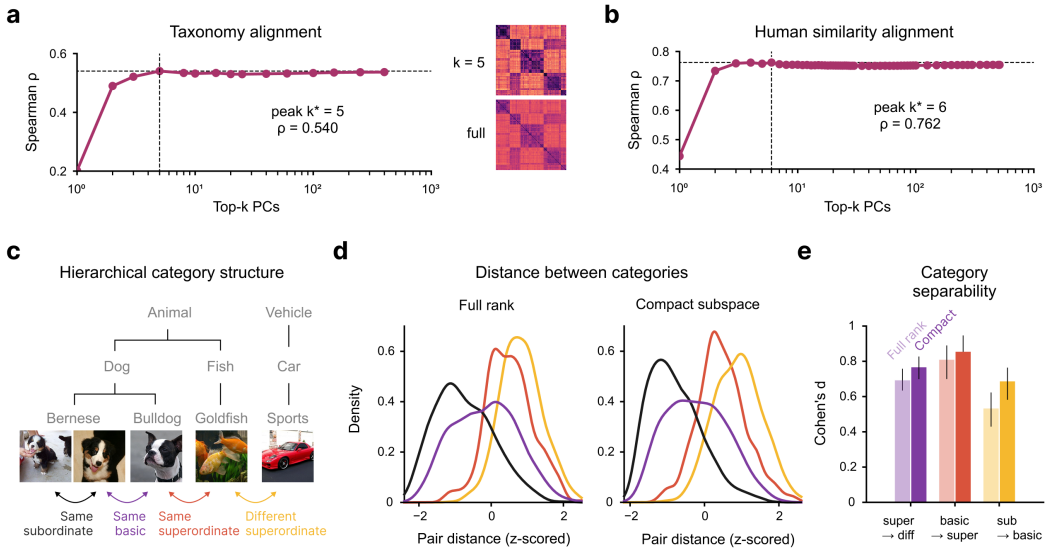


Figure 5: **Compact spectral subspaces preserve hierarchical category structure and human-aligned similarity. a,** Spearman correlation between embedding-derived pairwise distances and the reference taxonomic distance matrix as a function of retained top-k PCs. Inset matrices compare the rank-5 and full-rank distance structures. **b,** Spearman correlation between embedding distances and human similarity judgments on THINGS as a function of retained top-k PCs. **c,** Schematic of the hierarchical pair structure used to evaluate category separability. Pairs are classified as same-subordinate, same-basic, same-superordinate, or different-superordinate based on their lowest common ancestor. **d,** Distributions of z-scored pairwise distances by hierarchical relation, computed from the full-rank embedding (left) and the compact spectral subspace (right). **e,** Category separability quantified as Cohen's d between adjacent hierarchical relations: superordinate vs. different-superordinate, basic vs. superordinate, and subordinate vs. basic.

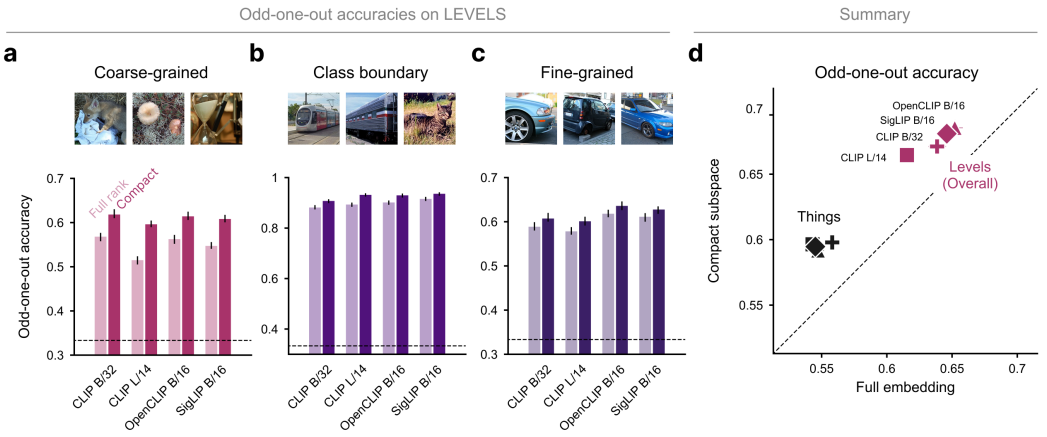


Figure 6: **Compact spectral subspaces improve alignment with human similarity judgments. a–c,** Odd-one-out accuracy on the LEVELS benchmark across coarse-grained (a), class-boundary (b), and fine-grained (c) trials, evaluated for CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16. Compact spectral subspaces (dark bars) outperform the full-rank embedding (light bars) at all three difficulty levels and across all four backbones. Dashed lines indicate chance. **d,** Summary comparison of full-rank and compact-subspace odd-one-out accuracy on LEVELS (overall) and THINGS. Each marker denotes a single VLM backbone. All points fall above the diagonal, indicating consistently higher behavioral alignment after spectral compression.

$k^* = 6$, $\rho = 0.762$ vs. full-rank $\rho = 0.755$). These modest benefits hold across all four backbones and nearby retained ranks (Supplementary Results A.13). These results indicate that later components are not necessary for preserving the taxonomic and pairwise human-similarity structure captured by the full embedding. To verify that this preservation is specific to the variance-ordered spectral structure rather than a generic effect of dimensionality reduction, we construct a dimensionality-matched random-subspace control by sampling k PCs uniformly at random. Random subspaces of the same dimensionality fall below the full-rank value at every rank on both taxonomy and human-similarity readouts, whereas the leading compact spectral subspace preserves the alignment structure.

To understand what this compact subspace preserves, we examine pairwise category distances stratified by hierarchical relations. Using a reference hierarchy (Animal vs. Vehicle at the superordinate level; Dog, Bird, and Truck at the basic level; Retriever, Poodle, Owl, and Semi at the subordinate level), we classify pairs as same-subordinate, same-basic, same-superordinate, or different-superordinate based on their lowest common ancestor (Figure 5c). In the full-rank embedding, distance distributions for these four pair types are broad and substantially overlapping. In the compact spectral subspace, the distributions shift apart and narrow, producing sharper separation organized by hierarchical distance (Figure 5d). Quantifying separability with Cohen's d between adjacent hierarchical relations confirms this pattern: compact subspaces increase separability across superordinate, basic-level, and subordinate comparisons (Figure 5e, superordinate→different-superordinate Cohen's d $0.53 \rightarrow 0.68$, the largest absolute increase among adjacent boundaries).

We then evaluate compact subspaces directly against human odd-one-out judgments. On LEVELS, compact subspaces improve odd-one-out accuracy on coarse-grained trials, where the task depends on broad superordinate distinctions (Figure 6a, CLIP B/32 $0.568 \rightarrow 0.618$, $\Delta = +0.050$, 95% CI [0.037, 0.065]). Improvements are also observed on class-boundary trials, which require identifying category boundaries (Figure 6b), and on fine-grained trials, which require discerning subtle distinctions within the same category (Figure 6c). Although the gains are most pronounced for coarse-grained distinctions, they extend across all three LEVELS trial types. Aggregating across both LEVELS and THINGS, every backbone falls above the diagonal in the summary scatter, indicating that spectral compression consistently increases odd-one-out accuracy relative to the full-rank embedding (Figure 6d). To confirm that the behavioral gain is not simply due to dimensionality reduction, we again compare the compact spectral subspace against a dimensionality-matched random-PC baseline (Supplementary Results A.14). Random subspaces with the same number of components perform essentially at the full-rank level across all four backbones, so the odd-one-out gain is specific to

the variance-ranked components retained by the spectral compact subspace. The improvement also transfers across datasets using fixed external bases, from THINGS to LEVELS and from ImageNet to THINGS (Supplementary Results A.11).

These results suggest that human-aligned semantic structure is a recoverable, compact substructure of the full embedding. A compact spectral subspace, obtained without modifying model parameters, preserves hierarchical category structure of the full-rank representation and modestly but consistently increases agreement with human odd-one-out judgments across multiple datasets and backbones. The remaining higher-rank components may carry information useful for fine-grained perceptual matching but appear to dilute the coarse, taxonomically organized structure that dominates human similarity judgments.

4 Discussion

We show that semantic abstraction in pretrained vision-language embeddings is spectrally stratified. Low-rank principal-component subspaces preserve broad taxonomic structure, whereas progressively higher-rank components support finer-grained distinctions. This organization generalizes across architectures, modalities, and taxonomies, and also has functional consequences: retaining leading components preserves coarse but degrades fine-grained retrieval, removing them disrupts abstract category structure selectively, and rank-selective routing improves retrieval when the selected subspace matches the target abstraction level. Compact subspaces of leading components preserve the taxonomic and human-similarity structure of the full embedding and increase agreement with human odd-one-out judgments. Together, these results suggest that VLM embeddings are not semantically homogeneous, but contain a recoverable substructure in which hierarchical abstraction is organized along the PC spectrum.

Our findings complement prior work on interpreting vision-language representations. Large-scale contrastive models are studied primarily as transferable zero-shot representations [1, 2, 5, 6], while interpretability work often focuses on identifying human-interpretable units or concepts inside these models. For example, prior work identifies multimodal neurons in CLIP [42], and language supervision is used to automatically describe hidden units in vision networks [43]. Other work evaluates full-representation alignment between models and human judgments or neural responses [8, 44, 45]. Our contribution is different: rather than assigning semantic labels to individual neurons or evaluating the full embedding as a single object, we decompose the embedding spectrum and show that different rank ranges support different abstraction levels. This also distinguishes our result from embedding post-processing methods, where top components are often treated as dominant directions to remove [34]; in VLMs, leading components can carry meaningful abstract category structure, and removing them selectively damages coarse semantic organization.

Related work links semantic hierarchy to both learning dynamics and representational geometry. Saxe et al. show how singular modes of structured data govern hierarchical differentiation in deep linear networks [15]. Park et al. characterize categorical concepts and hierarchical relations through polytopes and orthogonal subspaces in a transformed language-model representation space [16]. LiteEmbed reports a coarse-to-fine separation in CLIP's text principal components and uses this structure to adapt embeddings for rare classes [27]. These accounts motivate examining how semantic depth is ordered along the variance spectrum of pretrained representations and how that ordering supports abstraction-dependent behavior.

The spectral hierarchy we observe also connects to cognitive and neuroscientific accounts of hierarchical representation. Cognitive theories hold that concepts are organized across levels of abstraction, from superordinate to basic-level and subordinate categories [10, 12–14]. Similarity-based theories further emphasize that categorization depends on graded relations among objects rather than discrete labels alone [37, 39]. Studies of object vision show that human and primate IT representations contain both categorical and continuous structure [46–48]. This perspective also connects to broader accounts of hierarchical organization in the brain and biologically inspired models [49–52]. Here, retained spectral rank provides an axis along which broad conceptual partitions and fine distinctions become accessible within the same embedding.

Finally, our results contribute to the growing literature on human-aligned machine learning. Recent work shows that model representations can be made more human-aligned by using human behavioral data as an explicit alignment target. For example, a recent study trains a teacher model to imitate

human judgments and then fine-tunes foundation models through a human-aligned distillation objective, obtaining large improvements across multi-level similarity tasks [26]. Related work identifies latent dimensions underlying human–model representational alignment [53] or evaluates models using neural and behavioral benchmarks [45]. In contrast, our method does not introduce a human-alignment objective, alter model parameters, or finetune on human judgments; it simply truncates the spectrum of an off-the-shelf VLM embedding to isolate a compact subspace. The resulting gains are therefore expected to be smaller than those obtained by direct alignment or fine-tuning. However, this difference is also the main scientific implication: to some extent, human-aligned structure is already present in pretrained VLMs as a recoverable spectral substructure, even when it may be partially obscured in the full-rank embedding.

5 Broader impacts

This work is primarily a representational analysis of pretrained vision-language models and does not introduce a new deployed model, training dataset, or decision-making system. Its potential positive impacts include improving the interpretability and controllability of multimodal embedding spaces, especially by providing simple spectral tools for diagnosing which components support coarse versus fine semantic distinctions. Such tools may help practitioners audit retrieval, classification, and human-alignment behavior without retraining model parameters. At the same time, the same ability to modulate semantic granularity could be misused or could have unintended effects if applied in deployed retrieval or classification systems without appropriate validation. For example, rank-filtered embeddings may amplify biases already present in pretrained VLMs, alter performance unevenly across demographic or culturally specific categories, or encourage overly rigid taxonomic interpretations of concepts that are socially or contextually fluid. We therefore view this method as an analysis and auditing tool rather than a fairness or safety guarantee, and recommend that any downstream use be accompanied by dataset-specific bias evaluation, subgroup performance analysis, and transparency about which spectral subspaces are used.

6 Limitations

Several limitations qualify the scope of our conclusions. First, our experiments cover a limited range of contrastive VLMs, vision-only controls, datasets, large language models, and retrieval settings. Second, improvements in pairwise human-similarity correlations are modest in absolute magnitude, even when gains in odd-one-out accuracy are more pronounced. These findings should therefore be interpreted as evidence that compact spectral subspaces can improve certain alignment metrics, rather than as a complete account of human conceptual similarity. Finally, we do not claim that VLMs implement human conceptual mechanisms or that principal components correspond directly to neural dimensions. Establishing broader connections between our findings and neuroscience remains an important direction for future research.

7 Code availability

Python 3.12 (Python software foundation) with PyTorch 2.12 is used to perform the simulation. The code used in this work is available at `https://github.com/jeonghwan-cheon/spectral-abstraction`.

Acknowledgments and Disclosure of Funding

The authors acknowledge the MIT Office of Research Computing and Data for providing high performance computing resources that have contributed to the research results reported within this paper. Marin Vogelsang was supported by the Japan Society for the Promotion of Science (JSPS) Overseas Research Fellowship and the Yamada Science Foundation, and Lukas Vogelsang was supported by a grant from the Simons Foundation International to the Simons Center for the Social Brain at MIT.

References

[1] Radford, A. *et al.* Learning transferable visual models from natural language supervision. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 8748–8763 (PMLR, 2021).

[2] Jia, C. *et al.* Scaling up visual and vision-language representation learning with noisy text supervision. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 4904–4916 (PMLR, 2021).

[3] Li, J., Li, D., Xiong, C. & Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K. *et al.* (eds.) *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, 12888–12900 (PMLR, 2022).

[4] Schuhmann, C. *et al.* LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **35**, 25278–25294 (2022).

[5] Cherti, M. *et al.* Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829 (2023).

[6] Zhai, X., Mustafa, B., Kolesnikov, A. & Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986 (2023).

[7] Frome, A. *et al.* Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems* **26** (2013).

[8] Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A. & Kornblith, S. Human alignment of neural network representations. In *International Conference on Learning Representations* (2023).

[9] Muttenthaler, L. *et al.* Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems* **36**, 50978–51007 (2023).

[10] Collins, A. M. & Quillian, M. R. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior* **8**, 240–247 (1969).

[11] Smith, E. E. & Medin, D. L. *Categories and concepts* (Harvard University Press, 1981).

[12] Murphy, G. *The big book of concepts* (MIT press, 2002).

[13] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology* **8**, 382–439 (1976).

[14] Rosch, E. Principles of categorization. In *Cognition and categorization*, 27–48 (Lawrence Erlbaum, 1978).

[15] Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* **116**, 11537–11546 (2019).

[16] Park, K., Choe, Y. J., Jiang, Y. & Veitch, V. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations* (2025).

[17] Rogers, T. T. & McClelland, J. L. *Semantic cognition: A parallel distributed processing approach* (MIT press, 2004).

[18] Lupyan, G. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology* **3**, 54 (2012).

[19] Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM* **38**, 39–41 (1995).

[20] Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

[21] Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755 (Springer, 2014).

[22] Van Horn, G. *et al.* The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8769–8778 (2018).

[23] Krizhevsky, A. Learning multiple layers of features from tiny images. Tech. Rep., University of Toronto (2009).

[24] Hebart, M. N. *et al.* Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS one* **14**, e0223792 (2019).

[25] Hebart, M. N. *et al.* Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* **12**, e82580 (2023).

[26] Muttenthaler, L. *et al.* Aligning machine and human visual representations across abstraction levels. *Nature* **647**, 349–355 (2025).

[27] Agarwal, A., Karanam, S. & Gandhi, V. Liteembed: Adapting clip to rare classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 6133–6142 (2026).

[28] Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**, 249 (2008).

[29] Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* **17**, 401–412 (2013).

[30] Nili, H. *et al.* A toolbox for representational similarity analysis. *PLOS Computational Biology* **10**, e1003553 (2014).

[31] Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).

[32] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417 (1933).

[33] Jolliffe, I. T. Principal components in regression analysis. In *Principal Component Analysis*, 129–155 (Springer, 1986).

[34] Mu, J. & Viswanath, P. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations* (2018).

[35] Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[36] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

[37] Tversky, A. Features of similarity. *Psychological Review* **84**, 327 (1977).

[38] Medin, D. L. & Schaffer, M. M. Context theory of classification learning. *Psychological Review* **85**, 207 (1978).

[39] Nosofsky, R. M. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**, 39 (1986).

[40] Grattafiori, A. *et al.* The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[41] Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770* (2026).

[42] Goh, G. *et al.* Multimodal neurons in artificial neural networks. *Distill* **6**, e30 (2021).

[43] Oikarinen, T. & Weng, T.-W. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations* (2023).

[44] Peterson, J. C., Abbott, J. T. & Griffiths, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science* **42**, 2648–2669 (2018).

[45] Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv* 407007 (2018).

[46] Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).

[47] Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nature Neuroscience* **17**, 455–462 (2014).

[48] Konkle, T. & Caramazza, A. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience* **33**, 10235–10242 (2013).

[49] Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature reviews neuroscience* **15**, 536–548 (2014).

[50] Mai, G. *et al.* Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations* (2020).

[51] Dordek, Y., Soudry, D., Meir, R. & Derdikman, D. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife* **5**, e10094 (2016).

[52] Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *The Journal of neuroscience* **28**, 2539–2550 (2008).

[53] Mahner, F. P., Muttenthaler, L., Güçlü, U. & Hebart, M. N. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence* **7**, 848–859 (2025).

[54] Desai, K., Nickel, M., Rajpurohit, T., Johnson, J. & Vedantam, S. R. Hyperbolic image-text representations. In Krause, A. *et al.* (eds.) *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, 7694–7731 (PMLR, 2023).

[55] Pal, A. *et al.* Compositional entailment learning for hyperbolic vision-language models. In *The 13th International Conference on Learning Representations* (2025).

[56] Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024).

[57] He, K. *et al.* Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988 (2022).

[58] Lin, T.-Y., Maire, M., Belongie, S. *et al.* Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, 740–755 (2014).

[59] Russakovsky, O., Deng, J., Su, H. *et al.* ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).

A Supplementary results

A.1 Randomly permuted taxonomy control

We test whether spectral stratification follows the semantic correspondence between concepts and their taxonomic labels by permuting concept-to-path assignments. This preserves the tree structure and the number of categories at each depth while breaking their correspondence with the image embeddings. The original progression of peak ranks gives way to near-flat or non-monotonic trajectories across 20 permutations (Figure S1). Mean shuffled peak ranks are 8.6 ± 8.6, 26.1 ± 3.9, 28.7 ± 2.9, and 24.0 ± 0.0 at D1–D4, respectively, and increasing depth ordering appears in only 2/20 permutations. Leaf-level clustering is unchanged by relabeling, while the coarse partitions lose their distinct low-rank peaks. Under a representative permutation, model distance matrices no longer reproduce the broad blocks in the shuffled reference taxonomy. Thus, the spectral progression tracks the semantic assignment of concepts within the hierarchy, rather than the tree's marginal category counts alone.

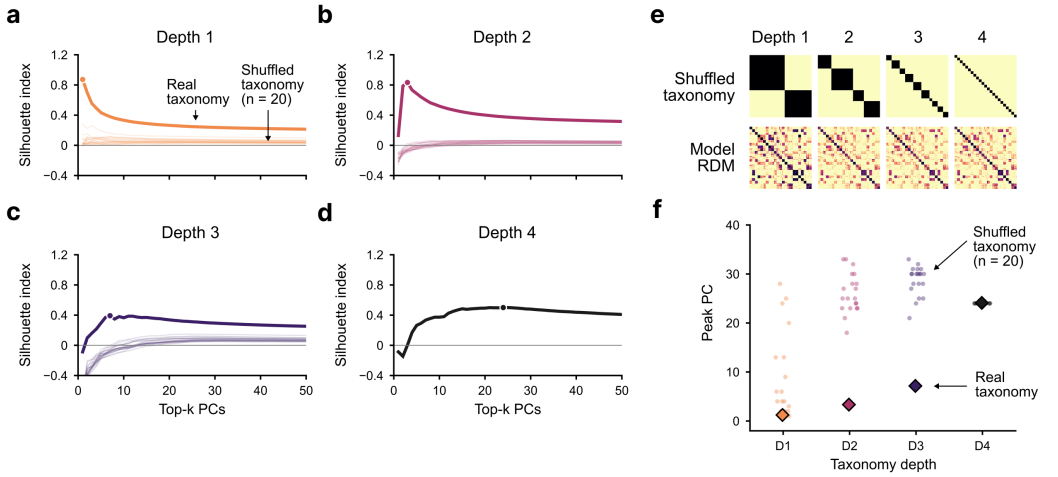


Figure S1: **Spectral organization under randomly permuted taxonomic labels. a–d,** Silhouette versus retained top-k PCs for D1–D4, respectively, comparing the original CLIP B/32 taxonomy (thick lines) with 20 shuffled assignments (thin lines). **e,** Reference taxonomic matrices (top) and rank-truncated model distance matrices (bottom), ordered by a representative shuffled assignment. **f,** Peak ranks for each permutation, with original ranks marked by diamonds.

A.2 Sensitivity to alternative taxonomy constructions

We examine whether the rank–depth progression depends on a particular branch or leaf composition of the probing taxonomy. First, we remove each of the four D2 branches in turn and repeat the analysis on the remaining hierarchy. Second, we construct 50 balanced resamples of the taxonomy. Strictly increasing peak ranks remain in three of four branch-exclusion variants for both CLIP B/32 and SigLIP B/16 (Table S1). Across balanced resamples, strict ordering occurs in 50/50 CLIP variants and 46/50 SigLIP variants. Absolute peak ranks vary with the retained concepts, particularly at the leaf level, but broad distinctions continue to occupy the leading components and finer distinctions peak later. The progression is therefore stable across these changes to branch coverage and leaf composition.

Table S1: **Peak-rank progression across alternative taxonomy constructions.** Entries give median selected ranks; brackets give minimum–maximum ranges for repeated constructions. Strict counts report $k^*_{D1} < k^*_{D2} < k^*_{D3} < k^*_{D4}$ across 1 original, 4 branch-exclusion, or 50 balanced variants. Each variant uses reference-only concept-centroid PCA, validation rank selection, and disjoint test images.

Backbone	Construction	D1	D2	D3	D4	Strict
CLIP B/32	Original	1	3	11	30	1/1
CLIP B/32	Leave one branch out	1 $[1, 2]$	2 $[2, 2]$	7.5 $[7, 9]$	22 $[18, 24]$	3/4
CLIP B/32	Balanced resamples	1 $[1, 2]$	3 $[3, 3]$	7 $[5, 10]$	15 $[13, 15]$	50/50
SigLIP B/16	Original	2	3	11	31	1/1
SigLIP B/16	Leave one branch out	1 $[1, 2]$	2 $[2, 2]$	8.5 $[8, 10]$	23 $[19, 26]$	3/4
SigLIP B/16	Balanced resamples	2 $[1, 2]$	3 $[2, 3]$	7 $[5, 10]$	15 $[13, 15]$	46/50

A.3 Controls for background and foreground size

We repeat the spectral analysis after removing image backgrounds and controlling for foreground object size. The foreground-only condition places segmented objects on a common background and uses 320 images from 32 concepts. The size-controlled condition additionally matches foreground object area and retains 226 images from 29 concepts, excluding hammer, knife, and scissors, whose spiky shapes make it infeasible to match their pixel areas through resizing. Both conditions preserve strictly increasing peak ranks in CLIP B/32 and SigLIP B/16 (Table S2). Background removal yields D1–D4 ranks of $(1, 3, 13, 27)$ and $(1, 3, 11, 27)$, respectively; the size-controlled subset yields $(1, 2, 6, 19)$ and $(1, 4, 6, 24)$. Coarse distinctions remain concentrated in the first few components, while finer categories require additional ranks. The accompanying Silhouette scores show clear category separation under both controls. These results indicate that spectral ordering is not driven by background context or foreground object size.

Table S2: **Spectral ordering and category separability after visual controls.** Silhouette values are evaluated at the selected rank for each depth. Within each condition, PCA fitting, rank selection, and scoring use the same images; the size-controlled subset has foreground-area error ≤ 0.02.

Backbone	Condition	k^* (D1–D4)	Silhouette (D1–D4)
CLIP B/32	Foreground only	$(1, 3, 13, 27)$	$(0.8778, 0.7946, 0.4439, 0.4504)$
CLIP B/32	Size controlled	$(1, 2, 6, 19)$	$(0.8767, 0.7465, 0.5282, 0.5996)$
SigLIP B/16	Foreground only	$(1, 3, 11, 27)$	$(0.8040, 0.8292, 0.5807, 0.6086)$
SigLIP B/16	Size controlled	$(1, 4, 6, 24)$	$(0.6394, 0.7229, 0.6117, 0.6163)$

A.4 Cluster-count and separability controls

Because deeper taxonomy levels contain more categories, we construct binary comparisons between sibling groups at each depth, holding cluster count fixed at two. Semantic tasks receive equal weight when their peak ranks are summarized. The median peak ranks increase from D1 to D4 across all four backbones, with strict progression in three backbones and a D1–D2 tie in OpenCLIP (Table S3). Thus, later peaks for finer distinctions also appear when every comparison has the same number of clusters.

We next restrict comparisons to overlapping ranges of full-rank silhouette across depths. Central-90% support and full empirical support produce similar rank trajectories: D1 peaks at ranks 1–2 and D4 at ranks 24–28 (Table S4). All eight backbone–support combinations are non-decreasing, and six are strictly increasing. This assignment-level support restriction provides a descriptive comparison within a shared range of full-space separability. Together, the two controls show that the observed progression extends beyond the increase in cluster count and remains evident after restricting the range of separability represented at each depth.

Table S3: **Peak ranks in cluster-count-matched sibling comparisons.** Entries give median ranks, with 95% intervals from 1,000 semantic-task cluster bootstrap resamples in brackets. Every task contains two balanced sibling groups; medians weight semantic tasks equally. There are 1, 2, 19, and 22 semantic tasks at D1–D4, respectively. The same 12 images per concept are used for PCA, rank selection, and scoring.

Backbone	D1	D2	D3	D4
CLIP B/32	1 [1, 1]	2.5 [2, 3]	9 [7, 11]	21.5 [14, 26]
CLIP L/14	1 [1, 1]	2.5 [2, 3]	9 [5, 10]	22.5 [19, 27]
OpenCLIP B/16	2 [2, 2]	2 [1, 3]	7 [4, 9]	21.5 [17.5, 25.5]
SigLIP B/16	1 [1, 1]	2 [1, 3]	10 [6, 11]	23.5 [19.5, 28]

Table S4: **Peak ranks within overlapping full-rank separability support.** Common support is defined across depths from assignment-level full-rank silhouette, using central-90% or full empirical intervals. Eligible assignments are aggregated within semantic tasks before depth medians are computed.

Backbone	Support	Interval	k^* (D1–D4)	Tasks D1/D2/D3/D4
CLIP B/32	Central 90%	[0.2456, 0.3346]	$(1, 2, 10, 24)$	1/2/10/6
CLIP L/14	Central 90%	[0.1994, 0.2820]	$(1, 2.5, 9, 28)$	1/2/7/6
OpenCLIP B/16	Central 90%	[0.2238, 0.3118]	$(2, 2, 7.5, 27)$	1/2/10/4
SigLIP B/16	Central 90%	[0.2022, 0.2951]	$(1, 2, 11, 27)$	1/2/12/5
CLIP B/32	Full empirical	[0.2343, 0.3509]	$(1, 2.5, 10, 24)$	1/2/14/6
CLIP L/14	Full empirical	[0.1871, 0.3042]	$(1, 2.5, 9, 28)$	1/2/11/6
OpenCLIP B/16	Full empirical	[0.2167, 0.3296]	$(2, 2, 7, 26)$	1/2/12/5
SigLIP B/16	Full empirical	[0.1826, 0.3151]	$(1, 2, 11, 26)$	1/2/14/7

A.5 Synthetic hierarchies and semantic covariance organization

We use synthetic embeddings to examine which covariance organizations produce a reliable four-level spectral progression. The hierarchical conditions share the probing taxonomy's 2/4/14/32 tree, while level-specific covariance energy and nuisance variance vary across conditions. Level energy denotes the latent top eigenvalue: it is equal across depths in S1, decreases as $1{:}0.5{:}0.25{:}0.125$ in S2, and follows the reverse schedule in S3; S4 adds dominant nuisance variance to S2. Descending coarse-to-fine energy produces strict ordering in 100/100 repetitions, compared with 7/100 for flat leaf geometry, 2/100 for equal level energy, and 0/100 for ascending energy or dominant nuisance variance (Table S5). Equal-energy hierarchies can already favor a partial coarse-to-fine progression; descending energy makes the complete four-level ordering strict and consistent. The same descending-energy condition remains strictly ordered across all tested combinations of embedding dimension (512, 768) and noise level ($\sigma = 0.05, 0.10, 0.20$). The synthetic comparison therefore identifies covariance organization as a determinant of the reliability and compactness of spectral ordering.

We relate this account to actual VLM embeddings using raw parent-conditioned centroid increments, $\boldsymbol{\beta}_c = \boldsymbol{\mu}_c - \boldsymbol{\mu}_{\mathrm{parent}(c)}$. For a node with S_c descendant leaves, its support-weighted energy is $\widehat{C}_c = S_c\|\boldsymbol{\beta}_c\|^2$, and its contribution to the leading subspace is $A_c(k) = q_c\|P_k\boldsymbol{\beta}_c\|^2$, where q_c is its descendant-leaf fraction. Controlling for depth, energy correlates positively with top-5 and top-10 contribution across all four backbones (Table S6; top-5 $\rho = 0.566$–0.767, top-10 $\rho = 0.786$–0.903). These directions also respond more selectively to their own descendants than to matched shuffled assignments, with selectivity contrasts of 0.245–0.311. The directions are not orthogonalized: the analysis connects spectral organization to the support and energy of overlapping semantic distinctions.

Table S5: **Ordering in controlled synthetic hierarchies** ($D = 512$, noise $\sigma = 0.10$). Rank entries give medians, with interquartile ranges (25th–75th percentiles) across 100 seeds in brackets. The hierarchical conditions share the 2/4/14/32 tree and use 64 independent samples per leaf in each reference, validation, and test split. Ranks are selected on validation data; strict-ordering counts are over the same seeds. S4 D3/D4 optima reach the full-rank search boundary.

Covariance organization	D1	D2	D3	D4	Strict
S0: Flat leaf geometry	7 [3, 17]	30 [27.75, 31]	31 [31, 31]	30 [30, 31]	7/100
S1: Equal level energy	2 [1, 3]	5.5 [4, 7]	31 [31, 31]	31 [30, 31]	2/100
S2: Descending level energy	1 [1, 1]	3 [3, 3]	13 [12, 13]	31 [30, 31]	100/100
S3: Ascending level energy	31.5 [1, 46]	39.5 [35, 45]	31 [31, 31]	18 [18, 18]	0/100
S4: Descending + nuisance	17 [17, 18]	511 [507, 512]	512 [512, 512]	512 [512, 512]	0/100

Table S6: **Semantic covariance energy, leading-subspace contribution, and descendant selectivity.** Top-5 and top-10 columns report partial Spearman correlations between $\widehat{C}_c$ and $A_c(k)$ controlling for depth. Selectivity is the true-minus-shuffled descendant contrast on held-out images; brackets give 95% bootstrap intervals.

Backbone	Top-5 contribution	Top-10 contribution	Selectivity over shuffle
CLIP B/32	0.767 [0.542, 0.886]	0.903 [0.794, 0.950]	0.262 [0.244, 0.281]
CLIP L/14	0.566 [0.336, 0.725]	0.886 [0.759, 0.938]	0.245 [0.230, 0.260]
OpenCLIP B/16	0.612 [0.377, 0.762]	0.819 [0.640, 0.896]	0.311 [0.291, 0.332]
SigLIP B/16	0.617 [0.382, 0.780]	0.786 [0.572, 0.890]	0.245 [0.227, 0.264]

A.6 Comparison with independent component analysis

We compare PCA with independent component analysis to examine whether cumulative coarse-to-fine ordering extends to a decomposition organized by statistical independence. FastICA is fit after external ImageNet PCA whitening, and independent components are ordered by their contribution energy in the original embedding space. Under the primary rank-256 configuration, PCA shows increasing normalized peak ranks in every backbone, whereas the median ICA trajectory is strictly increasing only for CLIP L/14 (Table S7). Strict ordering occurs in 3/10, 3/10, 2/10, and 0/10 ICA seeds for CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16, respectively.

Changing whitening rank or the FastICA nonlinearity also changes the recovered trajectories (Table S8). Some configurations expose an increasing sequence, but the sequence is not consistent across seeds, ranks, and backbones. This comparison locates spectral abstraction specifically in the variance-ranked PCA spectrum: a decomposition can preserve category information without recovering the same stable cumulative ordering of abstraction levels.

Table S7: **Normalized cumulative peak ranks under PCA and ICA.** Both decompositions are fit to external ImageNet images. ICA uses rank-256 whitening and logcosh FastICA, with components ordered by scale-invariant original-space contribution energy. PCA ranks are normalized by embedding dimension and ICA ranks by whitening rank. ICA tuples are medians over 10 seeds; rank selection and scoring use the same curated images.

Backbone	PCA (D1–D4)	ICA (D1–D4)	ICA strict seeds
CLIP B/32	$(0.004, 0.008, 0.092, 0.273)$	$(0.607, 0.869, 0.863, 0.863)$	3/10
CLIP L/14	$(0.001, 0.003, 0.168, 0.253)$	$(0.768, 0.852, 0.887, 0.900)$	3/10
OpenCLIP B/16	$(0.004, 0.006, 0.090, 0.266)$	$(0.451, 0.531, 0.688, 0.338)$	2/10
SigLIP B/16	$(0.003, 0.004, 0.171, 0.182)$	$(0.734, 0.588, 0.621, 0.691)$	0/10

Table S8: **ICA ordering across whitening ranks, seeds, and nonlinearities.** Rank tuples are medians across converged fits. Strict counts report the number of converged seeds with strictly increasing D1–D4 peak ranks. Ten seeds are attempted per configuration; one OpenCLIP rank-256 cube fit fails before convergence evaluation.

Backbone	Rank	Nonlinearity	Median k^* (D1–D4)	Converged	Strict
CLIP B/32	64	logcosh	$(27.5, 53, 50, 25)$	10/10	0/10
CLIP B/32	128	logcosh	$(98.5, 100, 93.5, 92)$	10/10	1/10
CLIP B/32	256	cube	$(155.5, 208, 209, 211.5)$	10/10	1/10
CLIP B/32	256	logcosh	$(155.5, 222.5, 221, 221)$	10/10	3/10
CLIP B/32	512	logcosh	$(239, 347.5, 327, 306.5)$	10/10	0/10
CLIP L/14	64	logcosh	$(45.5, 55.5, 57, 58)$	10/10	3/10
CLIP L/14	128	logcosh	$(112, 107, 103.5, 114.5)$	10/10	0/10
CLIP L/14	256	cube	$(239.5, 241.5, 245.5, 238.5)$	10/10	0/10
CLIP L/14	256	logcosh	$(196.5, 218, 227, 230.5)$	10/10	3/10
CLIP L/14	768	logcosh	$(475, 543, 541, 513)$	3/10	0/3
OpenCLIP B/16	64	logcosh	$(23.5, 50, 49, 38.5)$	10/10	0/10
OpenCLIP B/16	128	logcosh	$(49.5, 110, 113.5, 112.5)$	10/10	2/10
OpenCLIP B/16	256	cube	$(175, 161, 153, 132)$	9/9	0/9
OpenCLIP B/16	256	logcosh	$(115.5, 136, 176, 86.5)$	10/10	2/10
OpenCLIP B/16	512	logcosh	$(90, 234, 248, 172)$	8/10	1/8
SigLIP B/16	64	logcosh	$(19, 48.5, 48.5, 45.5)$	10/10	1/10
SigLIP B/16	128	logcosh	$(97.5, 99, 98.5, 100.5)$	10/10	2/10
SigLIP B/16	256	cube	$(245.5, 233, 224.5, 191.5)$	10/10	0/10
SigLIP B/16	256	logcosh	$(188, 150.5, 159, 177)$	10/10	0/10
SigLIP B/16	768	logcosh	$(481, 481, 441.5, 452.5)$	4/10	0/4

A.7 Spectral organization across VLM architectures and modalities

We extend the main-text vision-encoder analysis to CLIP L/14, OpenCLIP B/16, and SigLIP B/16. All three reproduce the increasing peak-rank sequence across the four probing depths (Figure S2a–c; numerical values for the original image analysis are in Table S21). Their rank-truncated distance matrices progressively reveal the hierarchy, from the animate–inanimate distinction to intermediate and leaf-level clusters (Figure S3b–d).

The same organization appears in the text encoders. Applying each backbone's image-derived basis to both modalities yields increasing peak ranks for image and text embeddings in all 10 prompt-template rotations (Table S9). Text D1 peaks at ranks 1–2 and text D4 at ranks 26–31, closely tracking the image-side progression. Thus, the shared image–text space supports a similar spectral organization in both modalities.

Vision-only DINOv2, MAE, and supervised ViT controls also retain fine category information, but their coarse-level silhouettes are lower and their distance matrices show less distinct broad partitioning (Figure S2d–f; Figure S3e–g). This comparison highlights the clarity of coarse semantic organization in the evaluated VLMs; the language-model results in Section A.10 further establish that spectral ordering extends beyond paired vision–language encoders.

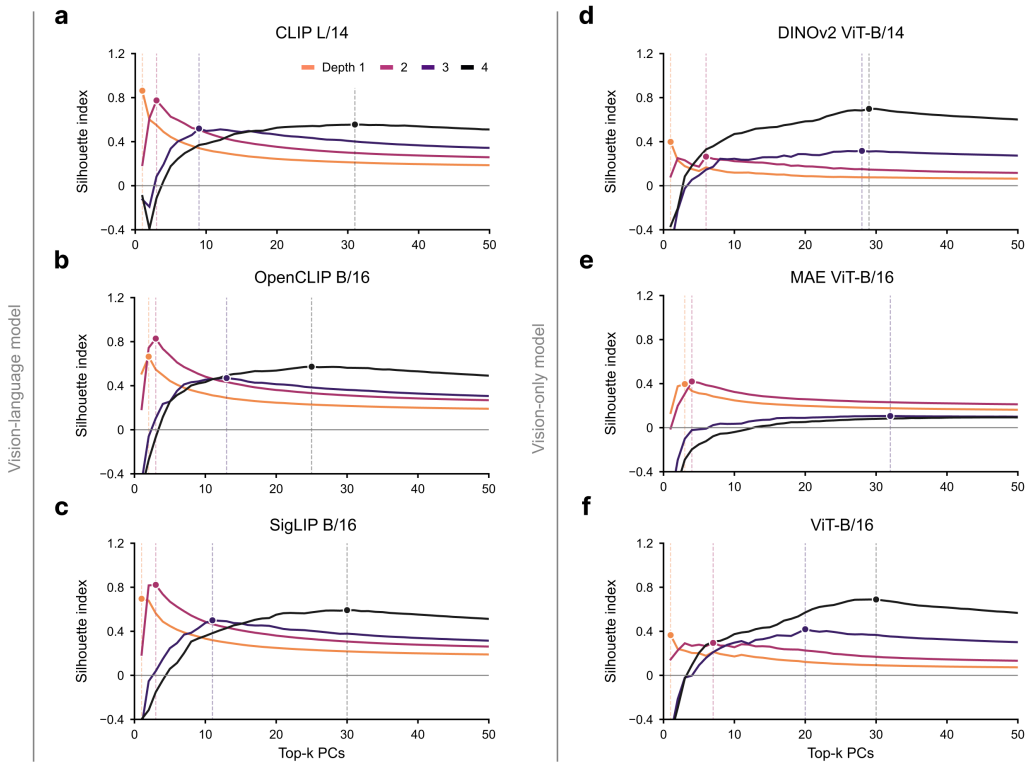


Figure S2: **PC-wise category clustering across vision–language and vision-only models.** Silhouette at D1–D4 as a function of retained top-k PCs. **a–c,** CLIP L/14, OpenCLIP B/16, and SigLIP B/16. **d–f,** DINOv2 ViT-B/14, MAE ViT-B/16, and supervised ViT-B/16. Colored dots and dashed lines mark each depth's peak rank.

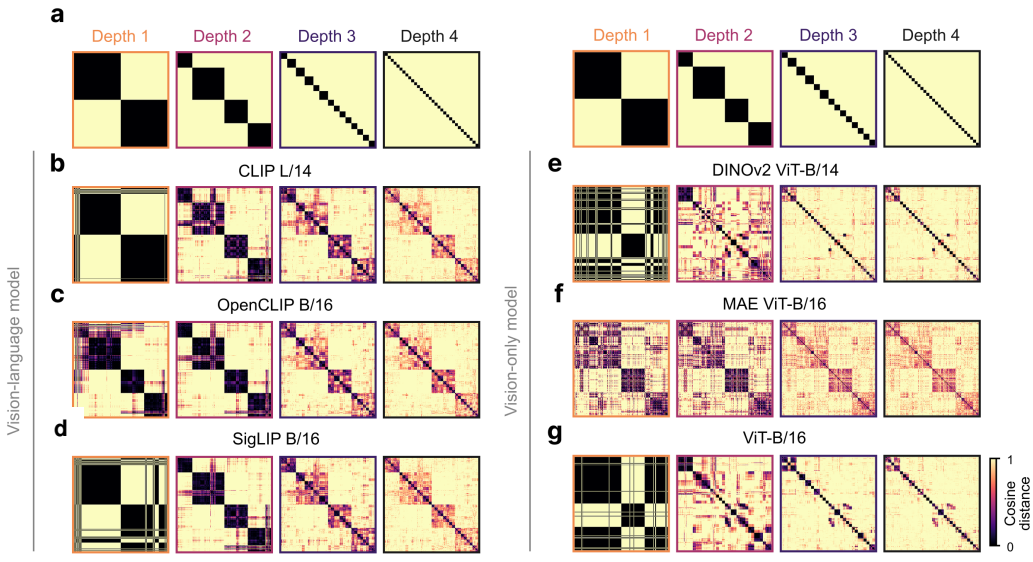


Figure S3: **Rank-truncated representational distance structure across model families. a,** Reference taxonomic distance matrices at D1–D4. **b–d,** Rank-truncated cosine distance matrices for CLIP L/14, OpenCLIP B/16, and SigLIP B/16, at each depth's silhouette-optimal rank. **e–g,** The corresponding analysis for DINOv2 ViT-B/14, MAE ViT-B/16, and supervised ViT-B/16.

Table S9: **Peak-rank ordering in image and text embeddings under the same image-derived PCA basis.** Entries give median ranks, with minimum–maximum ranges across 10 three-template/two-template rotations in brackets. Both modalities show strict ordering in every rotation for every backbone. Image PCA uses reference concept centroids; image validation/test sets and text validation/test templates are disjoint.

Backbone	Modality	D1	D2	D3	D4
CLIP B/32	Image	1 [1, 1]	3 [3, 3]	11 [11, 11]	30 [30, 30]
CLIP B/32	Text	1 [1, 1]	3 [3, 3]	11 [11, 11]	26 [26, 27]
CLIP L/14	Image	1 [1, 1]	3 [3, 3]	10 [10, 10]	31 [31, 31]
CLIP L/14	Text	1 [1, 1]	3 [3, 3]	10 [10, 12]	31 [21, 31]
OpenCLIP B/16	Image	2 [2, 2]	3 [3, 3]	12 [12, 12]	28 [28, 28]
OpenCLIP B/16	Text	1 [1, 1]	3 [3, 3]	12 [12, 12]	31 [28, 31]
SigLIP B/16	Image	2 [2, 2]	3 [3, 3]	11 [11, 11]	31 [31, 31]
SigLIP B/16	Text	2 [2, 2]	3 [3, 3]	11 [11, 11]	31 [26, 31]

A.8 Spectral ordering across image, text, and joint PCA bases

We vary the source of the PCA basis to test whether the image-side progression depends on deriving principal directions from visual examples. In addition to image-reference centroids, we fit bases to individual text prompts and to balanced image–text reference rows. Image-derived and joint bases yield strict D1–D4 ordering in every prompt rotation for all four backbones (Table S10). Text-derived bases also yield strict ordering in all rotations for CLIP L/14, OpenCLIP B/16, and SigLIP B/16. CLIP B/32 exhibits an early D1–D2 tie in the median text-basis trajectory, followed by a clear increase at D3 and D4.

Basis source changes the absolute ranks at which distinctions peak, while preserving the broad separation between leading coarse structure and later fine structure. For example, CLIP L/14 progresses from $(1, 3, 10, 31)$ under the image basis to $(2, 3, 16, 30)$ under the text basis and $(2, 5, 16, 42)$ under the joint basis. Together with the cross-modal comparison in Section A.7, these results show that spectral ordering is recoverable from multiple sources of covariance structure in the shared embedding space.

Table S10: **Image-embedding peak ranks across image-derived, text-derived, and joint PCA bases.** Entries give medians, with minimum–maximum ranges across 10 prompt-template rotations in brackets; strict counts refer to ordering within each rotation. Three text templates participate in basis fitting and rank selection; the other two form the text test set. Joint bases weight image and text rows equally.

Backbone	PCA source	D1	D2	D3	D4	Strict
CLIP B/32	Image	1 [1, 1]	3 [3, 3]	11 [11, 11]	30 [30, 30]	10/10
CLIP B/32	Text	1 [1, 1]	1 [1, 3]	17 [16, 18]	33 [32, 33]	4/10
CLIP B/32	Joint image–text	2 [2, 2]	4 [4, 4]	12.5 [11, 19]	40 [40, 41]	10/10
CLIP L/14	Image	1 [1, 1]	3 [3, 3]	10 [10, 10]	31 [31, 31]	10/10
CLIP L/14	Text	2 [1, 2]	3 [2, 5]	16 [12, 18]	30 [29, 31]	10/10
CLIP L/14	Joint image–text	2 [2, 2]	5 [4, 5]	16 [14, 17]	42 [41, 46]	10/10
OpenCLIP B/16	Image	2 [2, 2]	3 [3, 3]	12 [12, 12]	28 [28, 28]	10/10
OpenCLIP B/16	Text	1 [1, 1]	4 [3, 4]	15 [9, 16]	26 [26, 33]	10/10
OpenCLIP B/16	Joint image–text	3 [3, 3]	4 [4, 4]	15 [14, 16]	33 [31, 36]	10/10
SigLIP B/16	Image	2 [2, 2]	3 [3, 3]	11 [11, 11]	31 [31, 31]	10/10
SigLIP B/16	Text	1 [1, 2]	4 [3, 4]	10 [10, 11]	33.5 [32, 34]	10/10
SigLIP B/16	Joint image–text	2 [2, 2]	4 [4, 4]	15 [14, 15]	36 [32, 37]	10/10

A.9 Spectral abstraction in hyperbolic vision–language models

We compare Euclidean VLM (CLIP) with hyperbolic VLMs (MERU [54] and HyCoCLIP [55]) at matched ViT-S/16 and ViT-B/16 scales. For hyperbolic embeddings, we perform PCA in the tangent space at the hyperboloid origin. We evaluate both a common cosine-distance track in the linear representations and a native-geometry track that reconstructs hyperbolic points before computing geodesic distances. Both tracks recover strictly increasing D1–D4 peak ranks for all six checkpoints (Table S11). Under the common metric, for example, ViT-B/16 CLIP, MERU, and HyCoCLIP yield $(1, 3, 12, 15)$, $(2, 3, 12, 17)$, and $(2, 3, 13, 28)$.

Held-out silhouettes show that the models differ in how strongly particular depths separate, without a uniform hyperbolic advantage across depths and architectures (Table S12). HyCoCLIP B/16, for example, has higher common-metric leaf separation than matched CLIP, while its intermediate-depth separation is lower. The shared rank progression is the consistent result: spectral abstraction remains observable both in Euclidean embeddings and in representations trained with explicit hyperbolic inductive biases. Replacing the origin anchor with the reference-set Fréchet mean preserves every selected hyperbolic rank, with a maximum test-silhouette shift below 0.001.

Table S11: **Peak-rank progression in matched Euclidean and hyperbolic VLMs.** All six 512-dimensional checkpoints are trained on the same GRIT setup. PCA uses 128 reference images; ranks are selected on validation images and held fixed on the test split. The common metric is cosine distance; the native metric is cosine for CLIP and Lorentz geodesic distance for MERU/HyCoCLIP.

Architecture	Model	Common k^* (D1–D4)	Native k^* (D1–D4)
ViT-B/16	CLIP	$(1, 3, 12, 15)$	$(1, 3, 12, 15)$
ViT-B/16	MERU	$(2, 3, 12, 17)$	$(2, 3, 8, 17)$
ViT-B/16	HyCoCLIP	$(2, 3, 13, 28)$	$(2, 3, 9, 22)$
ViT-S/16	CLIP	$(1, 3, 9, 18)$	$(1, 3, 9, 18)$
ViT-S/16	MERU	$(1, 3, 9, 21)$	$(1, 3, 10, 21)$
ViT-S/16	HyCoCLIP	$(2, 3, 9, 25)$	$(2, 3, 9, 25)$

Table S12: **Held-out silhouette in matched Euclidean and hyperbolic VLMs.** Common- and native-metric columns use their corresponding selected ranks from Table S11.

Architecture	Model	Common metric (D1–D4)	Native metric (D1–D4)
ViT-B/16	CLIP	$(0.5748, 0.7310, 0.3984, 0.4020)$	$(0.5748, 0.7310, 0.3984, 0.4020)$
ViT-B/16	MERU	$(0.5279, 0.6794, 0.4025, 0.4138)$	$(0.4041, 0.5084, 0.2791, 0.2600)$
ViT-B/16	HyCoCLIP	$(0.5893, 0.6915, 0.3742, 0.4599)$	$(0.4713, 0.5020, 0.2448, 0.3104)$
ViT-S/16	CLIP	$(0.6297, 0.7294, 0.3815, 0.3279)$	$(0.6297, 0.7294, 0.3815, 0.3279)$
ViT-S/16	MERU	$(0.6094, 0.7100, 0.3848, 0.3021)$	$(0.4512, 0.5548, 0.2644, 0.1885)$
ViT-S/16	HyCoCLIP	$(0.4984, 0.6234, 0.3604, 0.3918)$	$(0.4011, 0.4528, 0.2449, 0.2433)$

A.10 Spectral abstraction in modern LLMs

We examine the same 32 concepts in Llama 3.1 8B, Llama 3.1 70B, and Gemma 4 E4B, using mean-pooled hidden states over the concept's label tokens [40, 41]. Prompt families provide separate PCA-reference, rank-validation, and test instances. At the penultimate layer, median peak ranks progress from $(2, 5, 21, 39)$ in Llama 8B to $(3, 7, 23, 51)$ in Llama 70B and $(4, 8, 28, 63)$ in Gemma (Table S13). Both Llama models are strictly ordered in all five prompt rotations; Gemma is strictly ordered in four and non-decreasing in all five. Median peak-rank sequences are also strictly increasing at the middle and final layers in every model. Held-out silhouette profiles and rank-truncated distance matrices illustrate this depth-dependent organization within a fixed prompt-family rotation (Figures S4 and S5). The ordering therefore generalizes across model family, parameter scale, and layer, with coarse distinctions represented by substantially fewer components than sibling and leaf distinctions.

We next remove leading PCs, choosing the removal count on validation sibling-leaf discrimination. The resulting test-family changes reveal a granularity-dependent trade-off (Table S14). Parent-conditioned leaf silhouette increases by 0.0079, 0.0074, and 0.0246, respectively, while D1 decreases

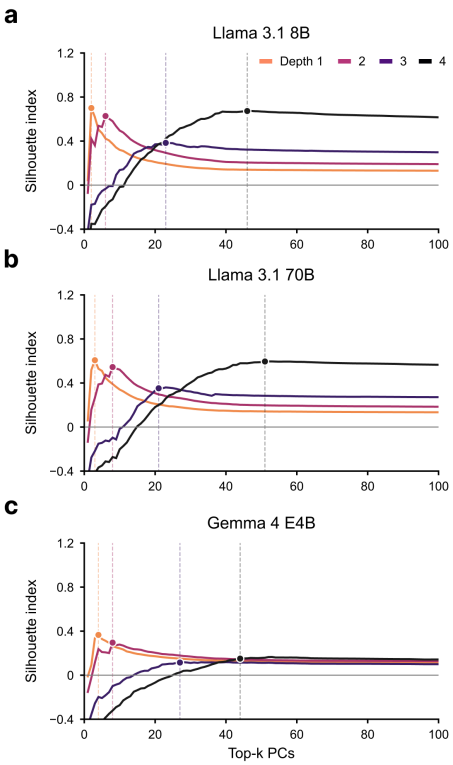


Figure S4: **PC-wise category clustering in language-model representations. a–c,** Held-out silhouette at D1–D4 as a function of retained top-k PCs for Llama 3.1 8B, Llama 3.1 70B, and Gemma 4 E4B, respectively. Colored dots and dashed lines mark validation-selected ranks, evaluated on the held-out test family using mean-restored reconstructions. All panels use the penultimate layer and the first prompt-family rotation; summaries across all five rotations are reported in Table S13.

by approximately 0.08–0.09 and D2 by 0.08–0.13. Matched random-PC removal changes coarse silhouettes by less than 0.002 in magnitude in the corresponding primary control. Leading directions thus carry broad semantic structure even where their removal slightly improves distinctions among sibling leaves. This extends the all-but-the-top comparison to contemporary language representations and identifies the abstraction level at which its benefit and cost occur [34].

Table S13: **Spectral ordering in modern LLMs at the penultimate layer.** Ranks are medians across five prompt-family rotations; brackets give their ranges. Strict and non-decreasing columns count rotations satisfying the respective ordering criterion. Each rotation uses 384 reference prompts (maximum rank 383), with separate validation and test families.

Model	D1	D2	D3	D4	Strict	Non-decr.
Llama 3.1 8B	2 [2, 3]	5 [5, 6]	21 [19, 23]	39 [38, 48]	5/5	5/5
Llama 3.1 70B	3 [2, 3]	7 [6, 8]	23 [21, 23]	51 [42, 53]	5/5	5/5
Gemma 4 E4B	4 [3, 4]	8 [4, 9]	28 [24, 77]	63 [44, 77]	4/5	5/5

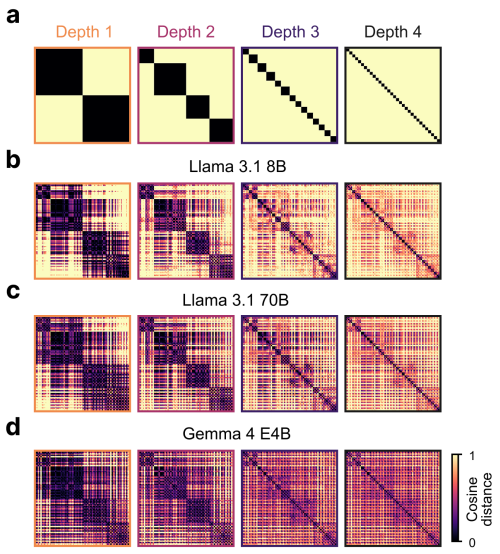


Figure S5: **Rank-truncated representational distance structure in language models. a,** Reference taxonomic distance matrices at D1–D4. **b–d,** Cosine distance matrices for Llama 3.1 8B, Llama 3.1 70B, and Gemma 4 E4B at their depth-specific validation-selected ranks. Matrices use the same 128 held-out prompts as Figure S4, ordered by the hierarchy, with four prompt variants per concept retained separately. Centered top-k projections are L2-renormalized without mean reintroduction, following Figure S3. All matrices share a 0–1 display range; larger cosine distances saturate at the upper color.

Table S14: **Granularity-dependent effects of leading-PC removal in modern LLMs.** Entries give medians, with minimum–maximum ranges across five prompt-family rotations in brackets. Removal count m^* is selected separately for each rotation by validation parent-conditioned D4 improvement. Changes are evaluated on test families relative to the mean-centered $m = 0$ baseline. $D4 \mid D3$ measures discrimination among sibling leaves within their D3 parent.

Quantity	Llama 3.1 8B	Llama 3.1 70B	Gemma 4 E4B
m^*	6 [4, 6]	8 [4, 8]	12 [8, 12]
$\Delta D1$	-0.0891 [$-0.1144, -0.0691$]	-0.0801 [$-0.0974, -0.0539$]	-0.0816 [$-0.1308, -0.0641$]
$\Delta D2$	-0.1261 [$-0.1406, -0.1099$]	-0.0813 [$-0.1015, -0.0553$]	-0.0949 [$-0.1843, -0.0761$]
$\Delta D3$	-0.0229 [$-0.0449, -0.0160$]	-0.0079 [$-0.0263, 0.0096$]	-0.0243 [$-0.0766, 0.0051$]
$\Delta(D4 \mid D3)$	0.0079 [0.0049, 0.0198]	0.0074 [0.0035, 0.0162]	0.0246 [$-0.0086, 0.0347$]

A.11 Transfer with fixed external PCA bases

We test whether the spectral organization transfers when PCA is estimated outside the evaluation taxonomy. For each backbone, we fit one basis to individual ImageNet images and another to individual THINGS images, then freeze each basis for the downstream analyses. Both external bases preserve strictly increasing D1–D4 selected ranks across all four backbones (Table S15). The numerical ranks shift with the reference distribution: CLIP B/32, for example, yields $(2, 4, 39, 175)$ with ImageNet and $(1, 5, 66, 182)$ with THINGS. Thus, a transferable coarse-to-fine progression remains even when the reference set and fitting unit differ from the curated concept-centroid analysis.

The same bases support abstraction-dependent retrieval gains on MS-COCO (Table S16). For CLIP B/32, ImageNet-based routing improves supercategory top-1 accuracy by 13.2 percentage points and category accuracy by 3.1 points; THINGS-based routing yields gains of 15.5 and 3.4 points. Every backbone improves supercategory retrieval under both bases, while category gains are smaller. Caption R@5 remains close to full-rank performance, consistent with detailed captions requiring a larger portion of the spectrum. Each abstraction level selects its rank on validation images before evaluation on disjoint test queries.

External compact subspaces also improve cross-dataset odd-one-out accuracy (Table S17). Transferring the THINGS basis to LEVELS yields gains of 1.23–2.15 percentage points across backbones; transferring the ImageNet basis to THINGS yields gains of 0.24–3.72 points. SigLIP B/16 improves by 1.49 and 3.72 points in the two transfers, respectively. With both the basis and validation-selected rank fixed during testing, these transfers connect spectral organization to retrieval and behavioral agreement beyond the dataset used to estimate the basis.

Table S15: **Taxonomic rank–depth progression under fixed external PCA bases.** Entries give selected ranks or depth slopes, with 95% intervals from 1,000 leaf-stratified image bootstrap resamples in brackets. Bases are fit to individual ImageNet or THINGS images and transferred to the curated taxonomy without refitting. Ranks are selected on validation images.

Backbone	Basis	D1	D2	D3	D4	Depth slope
CLIP B/32	ImageNet	2 [1, 2]	4 [3, 4]	39 [39, 61]	175 [103, 178]	55.4 [34.0, 58.1]
CLIP B/32	THINGS	1 [1, 1]	5 [5, 6]	66 [66, 106]	182 [178, 193]	60.4 [59.2, 67.7]
CLIP L/14	ImageNet	1 [1, 2]	2 [2, 7]	130 [63, 145.4]	169 [169, 169]	63.2 [56.3, 65.8]
CLIP L/14	THINGS	1 [1, 1]	5 [5, 6]	122 [114, 127]	167 [164, 222]	61.5 [60.3, 78.5]
OpenCLIP B/16	ImageNet	2 [1, 2]	3 [3, 3]	123 [17, 131]	133 [133, 176.3]	51.3 [40.7, 58.4]
OpenCLIP B/16	THINGS	1 [1, 1]	2 [2, 5]	130 [102, 146]	149 [123, 245]	57.2 [47.0, 87.3]
SigLIP B/16	ImageNet	1 [1, 2]	3 [3, 3]	134 [75, 135]	140 [138, 147]	54.8 [48.5, 56.9]
SigLIP B/16	THINGS	1 [1, 1]	6 [3, 6]	83 [71, 101]	161 [100, 223]	55.7 [37.2, 76.1]

Table S16: **MS-COCO retrieval gains with fixed external PCA bases.** Entries are routed-minus-full changes in percentage points, with paired 95% bootstrap intervals. Supercategory and category use image-to-text top-1 accuracy; caption retrieval uses image-to-text R@5. Ranks are selected on 2,500 validation images and evaluated on 2,500 disjoint test images; intervals use 1,000 paired query resamples.

Backbone	Basis	Supercategory Δ	Category Δ	Caption Δ
CLIP B/32	ImageNet	+13.19 [+11.61, +14.85]	+3.12 [+2.06, +4.17]	+0.00 [+0.00, +0.00]
CLIP B/32	THINGS	+15.46 [+13.64, +17.08]	+3.40 [+2.31, +4.49]	+0.00 [+0.00, +0.00]
CLIP L/14	ImageNet	+11.25 [+9.51, +12.91]	+1.38 [+0.73, +2.02]	−0.56 [-1.08, -0.04]
CLIP L/14	THINGS	+10.85 [+9.39, +12.34]	+1.38 [+0.73, +2.02]	−0.04 [-0.32, +0.24]
OpenCLIP B/16	ImageNet	+9.31 [+7.73, +10.85]	+1.82 [+0.97, +2.75]	+0.04 [+0.00, +0.12]
OpenCLIP B/16	THINGS	+9.15 [+7.65, +10.64]	+0.97 [+0.12, +1.94]	+0.04 [+0.00, +0.12]
SigLIP B/16	ImageNet	+6.35 [+4.65, +7.93]	+0.28 [-0.36, +0.97]	+0.00 [+0.00, +0.00]
SigLIP B/16	THINGS	+2.43 [+0.65, +4.25]	+0.16 [-0.36, +0.69]	+0.00 [+0.00, +0.00]

Table S17: **Cross-dataset odd-one-out transfer with frozen external PCA bases.** Full and compact columns give test accuracy (%); gains are percentage points with paired 95% bootstrap intervals. Ranks are selected on validation triplets before evaluation on disjoint test triplets. Intervals use 1,000 paired test-triplet resamples.

Backbone	Basis → benchmark	k^*	Full	Compact	Δ [95% CI]
CLIP B/32	THINGS → LEVELS	100	63.96	65.21	+1.25 [+0.67, +1.82]
CLIP B/32	ImageNet → THINGS	23	54.20	54.43	+0.24 [-0.14, +0.62]
CLIP L/14	THINGS → LEVELS	120	61.41	63.56	+2.15 [+1.55, +2.77]
CLIP L/14	ImageNet → THINGS	9	50.83	53.72	+2.88 [+2.30, +3.42]
OpenCLIP B/16	THINGS → LEVELS	152	65.37	66.60	+1.23 [+0.71, +1.76]
OpenCLIP B/16	ImageNet → THINGS	42	52.41	53.20	+0.79 [+0.41, +1.17]
SigLIP B/16	THINGS → LEVELS	239	64.68	66.16	+1.49 [+1.01, +1.97]
SigLIP B/16	ImageNet → THINGS	20	49.53	53.25	+3.72 [+3.21, +4.19]

A.12 Band-pass and leading-PC removal experiments

We isolate PC bands to distinguish where category information is concentrated from how many leading components are required to span it. Silhouette peaks move through successive bands: D1 at $[0, 1)$ (0.739), D2 at $[1, 3)$ (0.452), D3 at $[3, 8)$ (0.131), and D4 at $[8, 20)$ $(0.355$; Figure S6). Thus, the four depths have distinct spectral locations even when components outside the selected band are removed.

Within-band nearest-neighbor category retrieval provides a complementary image-only readout (Table S18). The first component already supports high animacy agreement, while bands extending into later components support much stronger intermediate and leaf-level agreement. For example, leaf retrieval rises from 0.023 in $[0, 1)$ to 0.785 in $[8, 20)$. Conversely, leading-PC removal preferentially disrupts coarse clustering (Table S19). Removing the first component reduces D1 silhouette from 0.172 to 0.024, while D4 remains approximately 0.314; removing the first three makes D1–D2 silhouettes near zero while retaining a D4 silhouette of 0.299. These complementary interventions support the interpretation that leading and later components contribute differently to coarse and fine category structure.

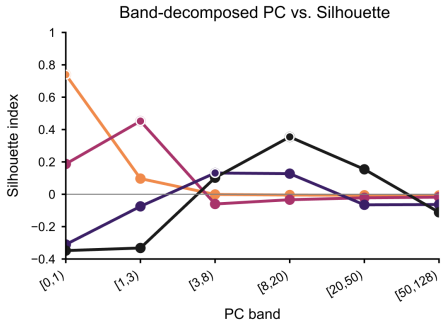


Figure S6: **Band-pass decomposition of taxonomic information.** Silhouette at D1–D4 in isolated PC bands $[a, b)$, using zero-based component indices. Outlined markers indicate the peak band for each depth. The analysis uses centered, L2-renormalized CLIP B/32 image embeddings; within-band category retrieval is reported separately in Table S18.

Table S18: **Within-band category retrieval on the 32-concept probing set (CLIP B/32).** For each query image, the nearest distinct image is retrieved by cosine distance; accuracy is agreement of their depth-specific category labels. The same-set analysis uses 256 images, with centered band projections L2-normalized without mean reintroduction.

PC band	D1	D2	D3	D4
$[0, 1)$	0.883	0.414	0.074	0.023
$[1, 3)$	0.793	0.777	0.348	0.176
$[3, 8)$	0.871	0.797	0.715	0.598
$[8, 20)$	0.938	0.922	0.867	0.785
$[20, 50)$	0.965	0.938	0.812	0.789
$[50, 128)$	0.754	0.621	0.344	0.273

Table S19: **Category structure after leading-PC removal on the probing set (CLIP B/32).** Silhouette is computed on the centered embedding after removing the first m PCs and L2-normalizing; $m = 0$ is the full centered representation. The images and PCA basis are the same as in the band analysis.

Removed PCs m	D1	D2	D3	D4
0	0.172	0.240	0.234	0.314
1	0.024	0.088	0.211	0.314
3	-0.005	-0.022	0.153	0.299
8	-0.006	-0.021	0.043	0.221

A.13 Taxonomic and human-similarity alignment in spectral versus random compact subspaces

We compare compact leading-PC subspaces with dimension-matched subsets of PCs sampled uniformly at random from the same basis. This holds the number of retained components fixed while discarding their variance ordering. For both taxonomic and human-similarity alignment, random subsets track or fall below the full-rank comparison, whereas leading subspaces preserve the corresponding distance structure at compact ranks (Figure S7). For CLIP B/32, the taxonomic correlation is 0.540 at rank 5 versus 0.537 at full rank, and human-similarity correlation is 0.762 at rank 6 versus 0.755 at full rank. These small differences show that compact leading subspaces preserve the structure captured by the full representation while removing most components.

The human-similarity result is consistent across backbones and nearby ranks (Table S20). Peak correlation gains are 0.0079, 0.0217, 0.0319, and 0.0325 for CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16. All four retain positive gains throughout ranks 3–8. The compact representation therefore preserves human-similarity structure across a neighborhood of ranks, rather than at a single isolated setting. Together with the random-PC comparison, this indicates that the retained variance ordering matters beyond dimensionality reduction alone.

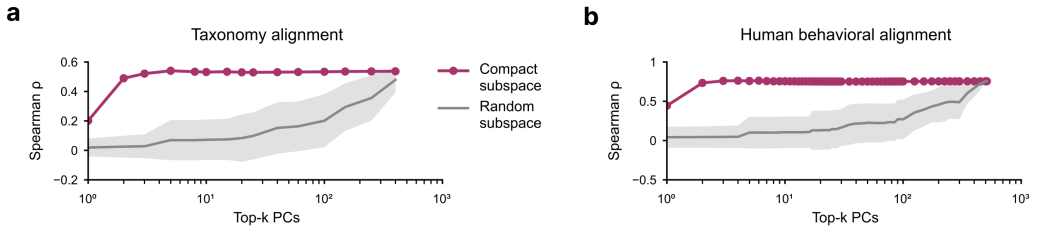


Figure S7: **Taxonomic and human-similarity alignment in spectral and random compact subspaces. a,** Spearman correlation with reference taxonomic distances as a function of retained rank. **b,** Correlation with human-similarity distances on THINGS. Leading top-k PCs are compared with k randomly sampled PCs from the same fitted basis and with the full-rank representation.

Table S20: **Human-similarity alignment across nearby retained ranks.** Entries are $\Delta\rho(k) = \rho_{\text{compact}}(k) - \rho_{\text{full}}$ on THINGS. Bold values mark the maximum gain for each backbone in the saved rank sweep. Compact and full-rank conditions both use centered representations.

Rank k	CLIP B/32	CLIP L/14	OpenCLIP B/16	SigLIP B/16
2	-0.0204	-0.0044	$+0.0088$	$+0.0127$
3	$+0.0051$	$+0.0156$	$+0.0141$	$+0.0080$
4	$+0.0076$	$\mathbf{+0.0217}$	$\mathbf{+0.0319}$	$\mathbf{+0.0325}$
5	$+0.0039$	$+0.0136$	$+0.0258$	$+0.0268$
6	$\mathbf{+0.0079}$	$+0.0131$	$+0.0289$	$+0.0299$
7	$+0.0022$	$+0.0124$	$+0.0202$	$+0.0213$
8	$+0.0001$	$+0.0081$	$+0.0175$	$+0.0205$

A.14 Odd-one-out alignment in spectral versus random compact subspaces

We apply the dimension-matched random-PC comparison to human odd-one-out judgments on LEVELS. For each backbone, 20 random PC subsets retain the same number of components as the compact spectral subspace. Across coarse-grained, class-boundary, and fine-grained trials, the compact spectral representation outperforms both the full-rank and random-subspace comparisons (Figure S8). For CLIP B/32, coarse-grained accuracy is 0.618 for the compact subspace, 0.568 for the full embedding, and 0.563 ± 0.054 for random subsets. The corresponding class-boundary values are 0.907, 0.881, and 0.886 ± 0.008; fine-grained values are 0.607, 0.589, and 0.593 ± 0.006.

The same ordering holds across the other backbones. For example, CLIP L/14 coarse accuracy is 0.596 in the compact subspace versus 0.482 ± 0.052 under random-PC selection; SigLIP B/16 yields 0.608 versus 0.508 ± 0.045. Thus, the behavioral benefit follows selection of particular variance-ranked components, rather than compression to an arbitrary subspace of the same dimensionality. Section A.11 reports the complementary validation-selected transfer comparison using fixed external bases.

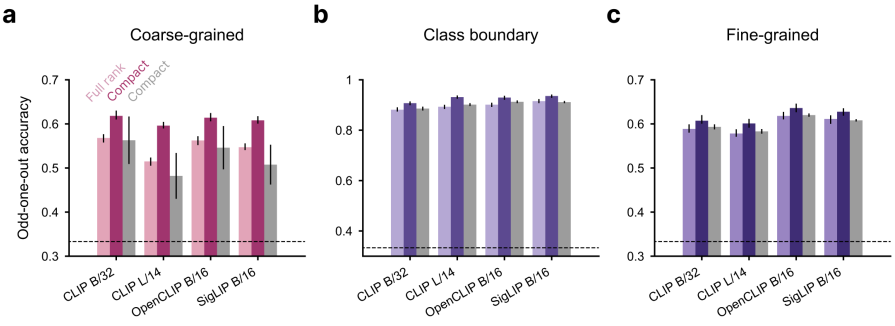


Figure S8: **Odd-one-out agreement in spectral and random compact subspaces. a–c,** LEVELS coarse-grained (a), class-boundary (b), and fine-grained (c) accuracy for the four VLM backbones. Bars compare the full embedding, compact spectral subspace, and dimension-matched random PC subsets. Random-subspace error bars show ± 1 SD over 20 subsets per backbone; dashed lines indicate chance. These comparisons use the retained ranks of the original main-result analysis.

B Numerical values for main results

Table S21: **Peak silhouette rank k^* per taxonomy depth on the four-level probing taxonomy**. Median over 100 bootstrap resamples; brackets give the 95% CI; depth-1 corresponds to $k^* = 1$ (95% CI [1, 2]) for CLIP B/32. All confidence intervals are 95% bootstrap CIs from $n = 100$ resamples.

Model	Depth 1	Depth 2	Depth 3	Depth 4
CLIP B/32	1 [1, 2]	3 [3, 3]	13 [10, 13]	29 [29, 33]
CLIP L/14	1 [1, 1]	3 [3, 3]	9 [9, 13]	31 [25, 32]
OpenCLIP B/16	2 [2, 2]	3 [3, 3]	13 [11, 13]	25 [24, 29]
SigLIP B/16	2 [1, 2]	3 [3, 3]	11 [11, 13]	30 [27, 31]

Table S22: **Peak silhouette rank k^* per taxonomy depth on four datasets** (CLIP B/32, complement to Figure 3c). Median bootstrap value; brackets give 95% CI. All confidence intervals are 95% bootstrap CIs from $n = 100$ resamples. A dagger marks intervals that reach the 500-component sweep cap.

Dataset	Depth (level)	k^*
THINGS	1 (animacy)	2 [2, 2]
	2 (super)	4 [4, 4]
	3 (mid)	131 [61, 248]
	4 (concept)	165 [110, 218]
LEVELS	1 (animacy)	1 [1, 1]
	2 (super)	3 [2, 5]
	3 (basic)	233 [124, 483]
	4 (subordinate)	494 [224, 500 †]
CIFAR-100	1 (coarse)	31 [20, 43]
	2 (fine)	131 [115, 198]
iNat21	1 (kingdom)	2 [2, 2]
	2 (phylum)	3 [2, 3]
	3 (class)	460 [46, 491]
	4 (order)	496 [455, 500 †]
	5 (family)	500 [490, 500 †]
	6 (genus)	499 [485, 500 †]
	7 (species)	500 [488, 500 †]

Table S23: **Image-to-text retrieval accuracy on the probing taxonomy as a function of retained top-k PCs (,** complement to Figure 4a). Saturation rank is the smallest k at which the curve is within 2 pp of its maximum.

Depth	Top-1 @ $k = 2$	Top-1 @ $k = 10$	Top-1 @ $k = 30$	Max top-1
1	0.955	0.988	0.994	0.997
2	0.668	0.941	0.969	0.980
3	0.180	0.604	0.792	0.882
4	0.007	0.245	0.537	0.742

Table S24: **MS-COCO retrieval, full-rank vs. rank-routed top-1 / R@5** (complement to Figure 4c). Routed values are reported at the validation-selected k^*.

Backbone	Level	Full	Routed	Δ	k^*
CLIP B/32	supercategory top-1	0.658	0.760	+0.103	346
	category top-1	0.851	0.871	+0.021	506
	caption R@5	0.749	0.749	+0.000	511
CLIP L/14	supercategory top-1	0.673	0.750	+0.077	560
	category top-1	0.897	0.914	+0.016	456
	caption R@5	0.795	0.797	+0.001	723
OpenCLIP B/16	supercategory top-1	0.676	0.751	+0.075	163
	category top-1	0.898	0.907	+0.009	485
	caption R@5	0.825	0.826	+0.001	503
SigLIP B/16	supercategory top-1	0.694	0.725	+0.030	455
	category top-1	0.915	0.918	+0.004	670
	caption R@5	0.862	0.862	+0.000	768

Table S25: **Pair-distance distributions and adjacent Cohen's d at full-rank vs. compact subspace** (CLIP B/32, complement to Figure 5d, e).

Pair relation (mean z of pairwise distance)	Full-rank	$k = 107$
same-subordinate	−0.75	−0.78
same-basic, different subordinate	−0.17	−0.21
same-superordinate, different basic	+0.51	+0.47
different superordinate	+0.83	+0.90
Cohen's d between adjacent relations	**Full-rank**	**$k = 107$**
subordinate → basic	0.69	0.76
basic → superordinate	0.81	0.85
superordinate → different-superordinate	0.53	0.68

Table S26: **Odd-one-out accuracy, full-rank vs. compact subspace** (complement to Figure 6). Brackets give 95% CI from 100 bootstrap resamples of test triplets.

Backbone	Trial type	Full	Compact	Δ	Δ 95% CI
CLIP B/32	LEVELS coarse	0.568 [0.558, 0.577]	0.618 [0.610, 0.625]	+0.050	[0.037, 0.065]
	LEVELS boundary	0.881 [0.874, 0.891]	0.907 [0.899, 0.914]	+0.026	[0.015, 0.038]
	LEVELS fine	0.589 [0.580, 0.599]	0.607 [0.600, 0.620]	+0.019	[0.004, 0.033]
	LEVELS overall	0.639 [0.632, 0.645]	0.671 [0.666, 0.677]	+0.033	[0.025, 0.040]
	THINGS overall	0.558 [0.553, 0.562]	0.598 [0.593, 0.602]	+0.040	[0.033, 0.046]
CLIP L/14	LEVELS coarse	0.515 [0.505, 0.524]	0.596 [0.589, 0.605]	+0.081	[0.071, 0.094]
	LEVELS boundary	0.893 [0.884, 0.901]	0.931 [0.925, 0.938]	+0.038	[0.029, 0.050]
	LEVELS fine	0.578 [0.570, 0.588]	0.601 [0.591, 0.611]	+0.023	[0.009, 0.035]
	LEVELS overall	0.615 [0.610, 0.621]	0.665 [0.660, 0.669]	+0.049	[0.042, 0.056]
	THINGS overall	0.543 [0.538, 0.548]	0.596 [0.592, 0.600]	+0.053	[0.047, 0.060]
OpenCLIP	LEVELS coarse	0.563 [0.552, 0.572]	0.614 [0.606, 0.625]	+0.052	[0.038, 0.066]
	LEVELS boundary	0.901 [0.892, 0.909]	0.929 [0.921, 0.937]	+0.028	[0.017, 0.039]
	LEVELS fine	0.618 [0.610, 0.627]	0.636 [0.626, 0.646]	+0.018	[0.006, 0.031]
	LEVELS overall	0.652 [0.645, 0.658]	0.685 [0.680, 0.692]	+0.033	[0.026, 0.041]
	THINGS overall	0.547 [0.542, 0.552]	0.592 [0.588, 0.597]	+0.044	[0.037, 0.052]
SigLIP	LEVELS coarse	0.547 [0.539, 0.556]	0.608 [0.601, 0.618]	+0.061	[0.050, 0.073]
	LEVELS boundary	0.915 [0.908, 0.923]	0.936 [0.928, 0.942]	+0.020	[0.010, 0.031]
	LEVELS fine	0.611 [0.600, 0.620]	0.628 [0.619, 0.636]	+0.016	[0.003, 0.027]
	LEVELS overall	0.646 [0.641, 0.651]	0.681 [0.676, 0.687]	+0.035	[0.030, 0.042]
	THINGS overall	0.545 [0.541, 0.550]	0.595 [0.592, 0.600]	+0.049	[0.043, 0.056]

C Methods

C.1 Overview of the spectral analysis pipeline

The analysis proceeds in four stages. First, image and text embeddings are extracted from a pre-trained, frozen vision-language model and L2-normalized in their native shared space. Second, a single principal-component basis is fit on a reference set of normalized image embeddings that is disjoint from any subsequent evaluation set. Third, rank-filtered embeddings are constructed by projecting both image and text embeddings onto the same top-k, remove-top-k, or band-pass principal-component subspace and re-normalizing. Fourth, downstream metrics (silhouette, representational similarity, multi-hierarchy retrieval, triplet odd-one-out) are evaluated as a function of the retained rank, and an optimal rank k^* is selected on a held-out validation split before being applied to a separate held-out test split. A central invariant of the pipeline is that image and text embeddings are always passed through the same basis, mean, and rank, so that cosine similarities remain comparable across all conditions.

C.2 Embedding extraction and preprocessing

Image preprocessing follows each backbone's released specification: `CLIPImageProcessor` for OpenAI CLIP variants, `AutoImageProcessor` for SigLIP, and the `open_clip` transform pipeline for OpenCLIP [1, 5, 6]. Pooled image features are cast to `float32` and L2-normalized along the embedding dimension. Text embeddings are produced by ensembling five canonical prompt templates per concept label,

> *"a photo of a {label}", "a photo of the {label}", "a cropped photo of a {label}", "an image of a {label}", "a picture of a {label}",*

with vowel-aware article assignment, encoding all five prompts independently, averaging the resulting embeddings across templates, and then L2-renormalizing the average. All extracted embeddings are cached to disk keyed on a stable per-sample identifier so that downstream stages reuse them deterministically and so that a single embedding extraction run supports every subsequent rank analysis.

C.3 Principal-component analysis and rank truncation

Let $\mathbf{u}_i \in \mathbb{R}^D$ denote a normalized image or text embedding, and let $\boldsymbol{\mu} \in \mathbb{R}^D$ be the empirical mean of a reference set of normalized image embeddings. Centered embeddings are written $\mathbf{z}_c = \mathbf{u} - \boldsymbol{\mu}$. A truncated singular value decomposition of the centered reference matrix yields the stacked principal directions $\mathbf{V} \in \mathbb{R}^{k_{\max} \times D}$ as right singular vectors and explained variances $s_j^2/(n-1)$ [31–33]. Three rank-filtered projections are used throughout the paper:

$$\begin{aligned} \mathbf{z}_{\text{top}k} &= \mathbf{z}_c \, \mathbf{V}_{:k}^\top \mathbf{V}_{:k}, \\ \mathbf{z}_{\setminus \text{top}k} &= \mathbf{z}_c - \mathbf{z}_{\text{top}k}, \\ \mathbf{z}_{[a,b)} &= \mathbf{z}_c \, \mathbf{V}_{a:b}^\top \mathbf{V}_{a:b}, \end{aligned}$$

corresponding to top-k retention, top-k ablation, and band-pass over principal components $[a, b)$. After projection, embeddings are returned to the ambient D-dimensional space, the empirical mean is reintroduced where downstream similarity scoring requires it, and the result is L2-renormalized before any cosine-based metric is computed. Image and text embeddings are always passed through the same basis, mean, and rank so that the two modalities remain matched. Rank sweeps span every integer from $k = 1$ to $k = D$, and tabular summaries report values at $\{1, 2, 5, 10, 25, 50, 100, 200, D\}$.

C.4 PCA fitting protocols and held-out evaluation

For each backbone and dataset, the principal-component basis is fit separately on a reference subset of normalized image embeddings from that dataset. Unless stated otherwise, the basis is fit on per-concept image-embedding centroids rather than individual image embeddings. The reference subset is disjoint from the samples used for rank selection and final evaluation, so that the PCA basis is not fit on the held-out evaluation items. Thus, when the dataset or backbone changes, the PCA

basis is refit; within a fixed backbone–dataset pair, the same fitted basis, mean, and rank-filtering procedure are reused across all metrics reported for that pair.

For every analysis that selects a rank from data, the relevant samples, images, triplets, or concepts depending on the dataset, are split into a held-out validation portion and a held-out test portion that share no items. The optimal rank k^* is chosen as the rank that maximizes the target metric on the validation portion, after which k^* is applied without further modification on the test portion. All peak silhouette values, peak RSA values, peak retrieval scores, and odd-one-out accuracies reported in the main text are computed from the test portion at the validation-selected k^*. Bootstrap 95% confidence intervals are estimated by resampling test items with replacement, using 100–1000 resamples depending on the analysis, and taking the 2.5th and 97.5th percentiles of the bootstrap distribution.

C.5 Probing taxonomic dataset

The probing taxonomy used for the rank–depth analyses in Figure 2 is a hand-curated four-level hierarchy designed so that every image carries a label at every depth, eliminating the confound that different depths might otherwise be evaluated on different image sets. Depth 1 separates the entire concept set into animate and inanimate. Depth 2 partitions instances into Human, Animal, Vehicle, and Tool. Depth 3 introduces 14 intermediate groupings (Adult, Child, Pet, Equid, Predator, Bird, Air, Rail, Road, Water, Kitchen, Garden, Hardware, Bristle), and depth 4 contains specific exemplars chosen to populate the depth-3 parents. The dataset is partitioned into a held-out validation portion used for rank selection (§ PCA fitting protocols and held-out evaluation) and a disjoint test portion used for all reported metrics, with the partitions stratified by depth-4 leaf so that every category is represented at both stages.

C.6 Vision-language models and vision-only control models

Four contrastive vision-language backbones serve as the primary models:

- CLIP B/32 (`openai/clip-vit-base-patch32`, $D = 512$) [1],
- CLIP L/14 (`openai/clip-vit-large-patch14`, $D = 768$) [1],
- OpenCLIP B/16 (ViT-B-16 with LAION-2B pretraining, $D = 512$) [5],
- SigLIP B/16 (`google/siglip-base-patch16-224`, $D = 768$) [6].

CLIP and SigLIP are loaded through the `transformers` library and OpenCLIP through `open_clip`; all four expose a paired image and text encoder that produces L2-normalized embeddings in a shared multimodal space.

Vision-only control models are used in the supplementary analyses to test whether spectral stratification depends on contrastive language supervision. The controls are

- DINOv2 ViT-B/14 (self-supervised) [56],
- MAE ViT-B/16 (masked-autoencoder pretraining) [57],
- ViT-B/16 (ImageNet supervised pretraining) [36].

C.7 Taxonomic benchmark datasets

Six taxonomic datasets are used to test cross-dataset generalization (Figure 3b) and to provide the multi-level retrieval targets in Figure 4. MS-COCO val2017 [21] contributes 5,000 images with three nested label sets: a binary animate–inanimate distinction, 12 supercategories, and 80 categories, together with ≈25,000 natural-language captions. THINGS [24, 25] provides object images organized under a curated 5-level taxonomic chain that links subordinate concepts to broader semantic groupings; concept-level metadata are taken from the public THINGS distribution. LEVELS [26] contributes ImageNet-derived odd-one-out triplets stratified into coarse-grained, class-boundary, and fine-grained types according to where the odd item sits in the synset hierarchy. WordNet [19] contributes hypernym chains for arbitrary leaf concepts. CIFAR-100 [23] provides a two-level coarse-fine hierarchy (20 coarse → 100 fine), and iNaturalist-21 [22] provides a Linnaean seven-level hierarchy (kingdom → phylum → class → order → family → genus → species). For datasets with

hierarchies of unequal length, comparisons across datasets normalize the depth coordinate to the unit interval (relative depth) so that the shallowest and deepest levels of every hierarchy align.

C.8 PC-wise category clustering analysis

Cluster structure at a given retained rank k is quantified by the silhouette index of taxonomic clusters at each hierarchy depth. For each k, image embeddings are projected onto the top-k subspace, returned to ambient space, and L2-renormalized, after which pairwise cosine distances are used to compute silhouette scores per depth. Rank-dependent silhouette curves yield a peak rank per depth, $k^*_{\text{depth}} = \arg\max_k \text{silhouette}_{\text{depth}}(k)$, selected on the held-out validation portion of the probing dataset; the silhouette value at k^*_{depth} is then re-evaluated on the held-out test portion and reported in the figures. The same rank-truncated embeddings are reused to compute cumulative explained variance and to produce the t-SNE visualizations shown in Figure 2d–g.

C.9 PC ablation and component removal task

To test whether leading principal components carry causal weight for coarse category structure, embeddings are passed through the remove-top-k projection $\mathbf{z}_{\setminus \text{top}k}$ defined above and re-normalized. Taxonomy alignment is then quantified by Spearman rank correlation between the upper-triangular vector of the embedding-derived cosine-distance matrix and the corresponding vector of the reference taxonomic distance matrix, computed at every k. The same procedure is applied with the band-pass projection $\mathbf{z}_{[a,b)}$ over selected component ranges to localize where in the spectrum a given depth's structure resides. Selectivity is read from the resulting curves: ablating a small number of leading components should disproportionately damage low-depth alignment while leaving higher-depth alignment relatively intact, in contrast to the all-but-the-top baseline where removing leading components is generally beneficial [34].

C.10 Multi-hierarchy MS-COCO retrieval task with rank-based routing

Three abstraction levels are evaluated on MS-COCO val2017: 12 supercategory prompts, 80 category prompts, and the full caption corpus, all encoded with the prompt-template ensemble described above. For supercategory and category prompts, retrieval is multi-label (each image carries a possibly-empty set of relevant labels) and is scored by image-to-text top-1 accuracy and mean average precision and by text-to-image mean average precision over images. Caption-level retrieval is scored by recall at 1, 5, and 10 in both directions; the image-to-caption ground truth is the union of an image's roughly five caption annotations. For each abstraction level, image and text embeddings are projected onto the top-k principal-component subspace using the THINGS-centroid basis and re-normalized before similarity is computed. The full-rank baseline is recovered as the special case $k = D$.

The rank-routed condition selects, separately for each abstraction level, the rank k^*_{level} that maximizes the target metric on a held-out validation slice of MS-COCO val2017, and reports retrieval at the selected k^* on a disjoint test slice. Rank routing therefore introduces no test-time supervision beyond the abstraction-level identity of the query. Bootstrap 95% confidence intervals are computed by resampling test images with replacement.

C.11 Random projection and random subspace controls

Two complementary random baselines are matched in dimensionality to top-k projection. The first projects embeddings onto a random k-dimensional orthonormal subspace constructed by QR decomposition of a Gaussian random matrix; the construction is independent of any data-derived basis. The second selects k principal-component indices uniformly at random without replacement from the same fitted basis used by top-k, so that the basis vectors are identical to the top-k vectors but the variance-ranked ordering is discarded. Both controls are run with multiple independent seeds and aggregated by mean and 95% bootstrap confidence interval. The first control isolates the contribution of having any k orthogonal directions of variance, while the second isolates the contribution of the variance-based ordering of components.

C.12 Randomly permuted taxonomy controls

To test that the rank-dependent organization is bound to the actual taxonomic structure rather than to any partition with the same marginal statistics, leaf-level taxonomic labels are permuted uniformly at random across the same set of images, and the reference distance matrix and downstream silhouette, RSA, and odd-one-out pipelines are rerun on the relabeled data. Image identities, embedding values, principal-component basis, and rank grid are held fixed; only the supervisory labels change. The procedure is repeated over at 20 independent permutation seeds, producing a null distribution against which the unshuffled effect can be tested.

C.13 Distance and separation measures across abstraction levels

The distance analyses in Figure 5 stratify pairs of categories by their lowest common ancestor in the reference hierarchy: same-subordinate, same-basic, same-superordinate, and different-superordinate. Pairwise distances are cosine distances on rank-filtered, L2-renormalized embeddings, z-scored across all pairs in a given condition before comparison. Separability between adjacent hierarchical relations (subordinate versus basic, basic versus superordinate, superordinate versus different-superordinate) is summarized by Cohen's d, computed both on the full-rank embedding and on the compact spectral subspace at the validation-selected k^* from § PC-wise category clustering analysis. Representational similarity between an embedding-derived distance matrix and the reference taxonomic distance matrix uses Spearman rank correlation on the upper-triangular vectors of the two matrices [28–30], with bootstrap confidence intervals computed by resampling categories with replacement.

C.14 Odd-one-out evaluation using LEVELS and THINGS

For each triplet of items, the predicted odd-one-out is the item that is most distant from the other two under cosine distance,

$$\hat{m} \;=\; \arg \max_{m \in \{i,j,l\}} \; \min_{m' \neq m} \left(1 - \mathbf{u}_m^\top \mathbf{u}_{m'}\right),$$

applied to L2-renormalized embeddings under the chosen rank-filtered projection. LEVELS triplets are stratified into coarse-grained, class-boundary, and fine-grained types based on where the odd item sits in the synset hierarchy [26], and accuracy is reported separately for each type as well as in aggregate. THINGS odd-one-out is evaluated on the publicly released human triplet trials [25]. Triplets are split into a held-out validation set and a disjoint test set; the compact-subspace rank k^* is selected as the rank that maximizes overall accuracy on the validation triplets, and accuracy at k^* is reported on the test triplets. Bootstrap 95% confidence intervals are estimated by resampling test triplets with replacement (1,000 resamples).

C.15 Taxonomy and visual controls

Alternative taxonomies omit each D2 branch in turn or sample 50 balanced subtrees with equal numbers of D3 children per D2 branch, leaves per selected D3 node, and images per leaf (Supplementary Results A.2). Rank summaries give medians and minimum–maximum ranges across variants.

Cluster-count controls pair direct siblings at each depth as binary classes, balancing descendant leaves between sides (up to four each) with 12 images per leaf (Supplementary Results A.4). Fifty sampling draws per pair are deduplicated. Assignment-level silhouette curves are aggregated by their pointwise median within each semantic task before peak selection; depth medians weight tasks equally. The 95% intervals use 1,000 semantic-task bootstrap resamples within depth. Separability controls restrict assignments to the intersection across depths of full-rank silhouette intervals, defined by either the 5th–95th percentiles or full empirical ranges, before task-level aggregation.

Visual controls isolate target objects using GroundingDINO and SAM 2.1 (Supplementary Results A.3). Foreground-only images replace backgrounds with a constant color, retaining 10 images per concept. Size-controlled images resize and center the foreground on a 512×512 canvas, preserving aspect ratio and targeting the median foreground-area fraction across valid masks. Images are retained at absolute area-fraction error ≤ 0.02, with at least two qualifying images per concept. Peak-rank search spans $k = 1, \ldots, 31$.

C.16 Synthetic hierarchies and semantic covariance

Synthetic embeddings use the probing taxonomy's 2/4/14/32 hierarchy (Supplementary Results A.5). Gaussian node directions are normalized and centered among siblings using descendant-leaf weights; leaf centroids sum their ancestral increments. Depth-specific covariance operators have equal leading eigenvalues, a descending $1{:}0.5{:}0.25{:}0.125$ schedule, or its reverse. A flat control draws leaf centroids independently. Centroids are rescaled to unit root-mean-square norm, preserving level-energy ratios. The nuisance condition adds taxonomy-independent Gaussian variation in 16 random directions with covariance leading eigenvalue 4. Isotropic Gaussian noise is added to 64 samples per leaf in each reference, validation, and test split. We evaluate 100 seeds for each $D \in \{512, 768\}$ and noise standard deviation $\sigma \in \{0.05, 0.10, 0.20\}$, reporting median ranks and interquartile ranges.

For VLM embeddings, semantic directions are $\boldsymbol{\beta}_c = \boldsymbol{\mu}_c - \boldsymbol{\mu}_{\mathrm{parent}(c)}$, with node centroids averaged equally over descendant leaves. Support-weighted energy and leading-subspace contribution are

$$\widehat{C}_c = S_c \|\boldsymbol{\beta}_c\|^2, \qquad A_c(k) = q_c \|P_k \boldsymbol{\beta}_c\|^2,$$

where S_c counts descendant leaves, q_c is their fraction of all leaves, and P_k is the reference PCA projector. Partial Spearman correlations control for depth. Descendant selectivity contrasts mean responses of held-out descendant and non-descendant centroids along the unit semantic direction. The null uses 1,000 global leaf permutations shared across nodes, preserving descendant-set sizes and overlap. The 95% intervals use 1,000 node bootstrap resamples stratified by depth.

C.17 Spectral analysis of language models

For the language-model analyses, we use the following models, loaded using the `transformers` library:

- Llama 3.1 8B (`meta-llama/Llama-3.1-8B`) [40],
- Llama 3.1 70B (`meta-llama/Llama-3.1-70B`) [40],
- Gemma 4 E4B (`google/gemma-4-E4B`) [41].

Each of the 32 concepts appears in five prompt families with four variants per family; prompts contain the concept label without its higher-level taxonomic labels. Hidden states are averaged over tokens overlapping the concept-label span at the middle, penultimate, and final layers. In each of five rotations, three families supply 384 reference vectors, one supplies 128 validation vectors, and one supplies 128 test vectors. Rank sweeps refine coarse-grid peaks at integer resolution, with maximum rank 383. Summaries report median ranks and minimum–maximum ranges across rotations.

Leading-component removal subtracts the first m PCs from mean-centered hidden states for $m \in \{0, 1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64\}$ [34]. Parent-conditioned D4 silhouette is computed among leaf concepts within each D3 parent and averaged equally across parents. Removal count m^* maximizes its validation improvement relative to the mean-centered $m = 0$ baseline, with ties resolved by the smaller count. Test-family changes are measured in global D1–D3 and parent-conditioned D4 silhouette. Matched controls remove 100 uniformly sampled PC subsets of the same size. Removal counts and test changes are summarized across the five rotations.

C.18 Statistical analysis

Unless otherwise noted, all reported confidence intervals are 95% bootstrap CIs computed from $n = 100$ resamples (over images, prompts, or odd-one-out triplets, depending on the analysis).

D Data availability

The image sets used in this study are available from the following sources:

- THINGS [24]: `https://osf.io/jum2f/`.
- MS-COCO 2017 [58]: `https://cocodataset.org`.
- ImageNet-1k [59]: `https://www.image-net.org/challenges/LSVRC/2012/2012-downloads.php`.

Four contrastive vision-language backbones serve as the primary models:

- CLIP B/32 [1]: `https://huggingface.co/openai/clip-vit-base-patch32`.
- CLIP L/14 [1]: `https://huggingface.co/openai/clip-vit-large-patch14`.
- OpenCLIP B/16 [5]: `https://huggingface.co/laion/CLIP-ViT-B-16-laion2B-s34B-b88K`.
- SigLIP B/16 [6]: `https://huggingface.co/google/siglip-base-patch16-224`.

We use the following vision-only control models:

- DINOv2 ViT-B/14 [56]: `https://huggingface.co/facebook/dinov2-base`.
- MAE ViT-B/16 [57]: `https://huggingface.co/facebook/vit-mae-base`.
- ViT-B/16 [36]: `https://huggingface.co/google/vit-base-patch16-224`.

For the language-model analyses, we use the following models:

- Llama 3.1 8B [40]: `https://huggingface.co/meta-llama/Llama-3.1-8B`.
- Llama 3.1 70B [40]: `https://huggingface.co/meta-llama/Llama-3.1-70B`.
- Gemma 4 E4B [41]: `https://huggingface.co/google/gemma-4-E4B`.

The human similarity reference was derived from the publicly released 66-dimensional SPoSE behavioral embedding [25]. The embedding and the THINGS human triplet odd-one-out judgments are available at `https://osf.io/f5rn6/`, and the Levels human triplet odd-one-out judgments are available at `https://doi.org/10.12751/g-node.hg4tdz`.

E Computing resources

Model training and GPU-based evaluation were conducted on the MIT ORCD Engaging computing cluster using NVIDIA A100 GPUs. The models were implemented in PyTorch, with batched computations performed on the GPUs and input loading parallelized across CPU worker processes through PyTorch.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the paper's main claims, including its contributions, assumptions, and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We include a separate "Limitations" section in the paper, where we discuss the scope of our claims and assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: We describe detailed methods and instructions for reproducing the experimental results in the corresponding sections of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: We provide open access to the data and code via a GitHub repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: We provide experimental details in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: We report error bars for all results shown in the figures, as well as detailed numerical statistical values.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We provide information on the compute resources used for the experiments in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We reviewed the NeurIPS Code of Ethics, and our research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: We discuss both the potential positive and negative societal impacts of the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: Since our paper is basic academic research, it does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: We cite the original papers that provide the publicly available assets and describe how to access them in the Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: We provide detailed documentation, such as README files, alongside the assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.