Spectral Stratification of Semantic Abstraction in Vision-Language Models

Jeonghwan Cheon [1] **Marin Vogelsang** [2] **Lukas Vogelsang** [2] **Pawan Sinha** [2]

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, low-rank principal-component subspaces encode broad conceptual structure, while higher-rank components carry progressively finer-grained distinctions. This pattern holds across multiple contrastive vision-language models and taxonomic datasets. Consistent with this stratification, rank-based modulation produces predictable behavioral shifts: retaining only the leading components preserves coarse retrieval but degrades fine-grained retrieval, while removing them selectively disrupts abstract category structure. Furthermore, rank-selective routing improves retrieval when the selected subspace matches the target level of abstraction. Notably, compact spectral subspaces, which preserve taxonomic structure while truncating high-rank residuals, exhibit substantially better alignment with human abstraction behavior than the full-rank model. Together, these results reveal that vision-language embeddings are not semantically homogeneous. Instead, hierarchical abstraction is stratified along the spectral geometry, and the structure most aligned with human cognition is concentrated in a compact subspace, suggesting that human-aligned semantics are a recoverable substructure of, rather than a property of, the full embedding.

[1]Independent Researcher, Seoul, Republic of Korea [2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA. Correspondence to: Jeonghwan Cheon <jeonghwan518@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Vision-language models (VLMs) have become a standard foundation 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 (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Schuhmann et al., 2022; Cherti et al., 2023; Zhai et al., 2023). Their embeddings are now widely regarded as general-purpose semantic spaces, used to retrieve images, classify novel categories, and evaluate model–human representational alignment (Frome et al., 2013; Peterson et al., 2018; Muttenthaler et al., 2023a;b). However, it remains poorly understood how conceptual abstraction emerges and is organized within these spaces. Specifically, are broad conceptual and fine-grained perceptual distinctions uniformly distributed, or are they organized along a latent geometric structure?

This question connects to a long-standing issue in cognitive science, where conceptual knowledge is structured across levels of abstraction, with concepts embedded in taxonomic hierarchies in which abstract categories subsume more specific ones (Collins & Quillian, 1969; Smith & Medin, 1981; Murphy, 2002). These levels are not equally salient: basic-level categories such as *dog* or *chair* occupy a privileged position, balancing within-category similarity against between-category distinctiveness, while superordinate (*animal*) and subordinate (*poodle*) categories are placed above and below (Rosch et al., 1976; Rosch, 1978). A central question is how such hierarchical organization is reflected in representations that jointly encode perceptual regularities and linguistic naming (Rogers & McClelland, 2004; Lupyan, 2012).

Contrastive VLMs offer a direct setting in which to examine this question, since they are trained on aligned image–text pairs at scale (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). Modern taxonomies, fine-grained recognition benchmarks, human similarity judgments, and neural response datasets enable us to test this structure across multiple notions of abstraction (Miller, 1995; Deng et al., 2009; Lin et al., 2014; van Horn et al., 2018; Krizhevsky, 2009; Hebart et al., 2019; 2023; Muttenthaler et al., 2025). However, most analyses evaluate VLM embeddings as single, full-rank representations, leaving open whether different

components of the embedding space support different levels of abstraction.

Analyses of representational geometry, together with the spectral perspective developed here, enable us to address this question. Beyond analyses of simple prediction accuracies, comparing pairwise embedding distances with taxonomic, behavioral, or neural dissimilarities can reveal how the model's internal similarity structure is organized (Kriegeskorte et al., 2008a; Kriegeskorte & Kievit, 2013; Nili et al., 2014). Principal components order directions of variation by explained variance, so rank-filtered subspaces let us ask which parts of the spectrum support what kinds of semantic structure (Pearson, 1901; Hotelling, 1933; Jolliffe, 1986; Mu & Viswanath, 2018). If broad distinctions concentrate in leading components while fine-grained distinctions require higher-rank residuals, the PC spectrum itself becomes an operational axis for probing and controlling semantic granularity.

In this work, we show that conceptual abstraction in pretrained VLM embeddings is spectrally stratified. Specifically, low-rank subspaces encode broad conceptual structure, while higher-rank components carry progressively finer-grained distinctions. This pattern generalizes across multiple contrastive VLMs and taxonomic datasets. Rank-based modulation produces predictable behavioral shifts: retaining leading components preserves coarse retrieval while degrading 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. Finally, the same spectral structure identifies the part of the embedding most aligned with humans. We define *compact spectral subspaces* as low-rank subspaces that preserve taxonomic structure while truncating high-rank residuals. Notably, these subspaces align better with human similarity judgments than the full-rank embedding. Together, these results suggest that VLM embeddings are spectrally stratified by abstraction level, with human-aligned structure recoverable from a compact subspace rather than distributed across the full embedding.

2. Preliminaries

2.1. Contrastive vision-language learning

We study 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 (Radford et al., 2021; Jia et al., 2021), 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 written as

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

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 (van den Oord et al., 2018). 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 (Frome et al., 2013; Radford et al., 2021; Jia et al., 2021).

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 (Dosovitskiy et al., 2021; Radford et al., 2021). OpenCLIP follows the same training paradigm but is trained in an open-source setting, often using public datasets such as LAION (Schuhmann et al., 2022; Cherti et al., 2023). SigLIP-B/16 uses a ViT-B/16 backbone and replaces the softmax contrastive objective with a pairwise sigmoid loss for improved scalability (Dosovitskiy et al., 2021; Zhai et al., 2023). 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 accounts, 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 (Smith & Medin, 1981; Murphy, 2002). 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

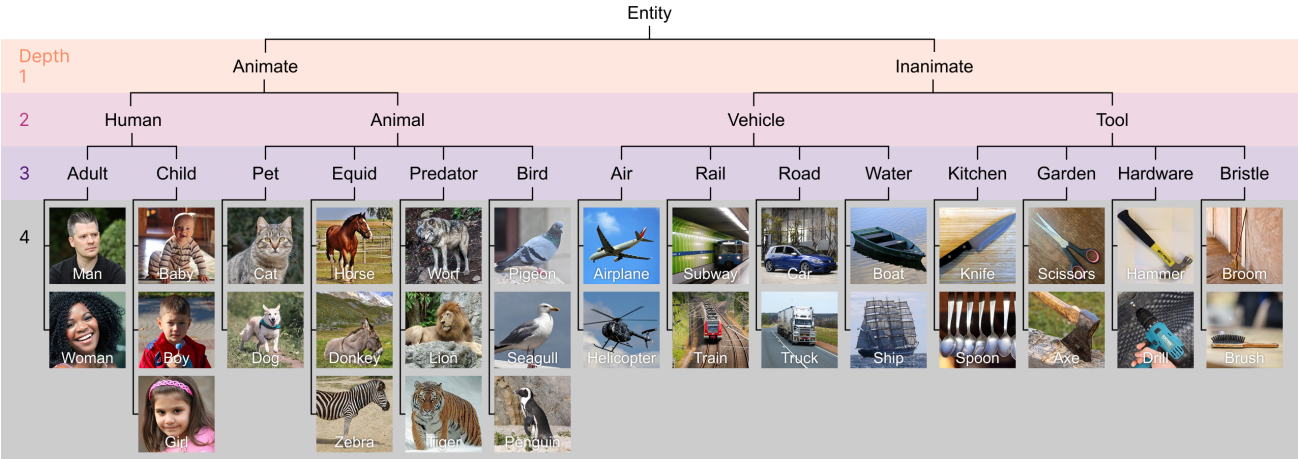

Figure 1. **A four-level probing taxonomy for measuring abstraction in VLM embeddings.** The taxonomy spans four hierarchical levels, from a broad animate–inanimate split at depth 1 to fine-grained leaf-level categories at depth 4. Intermediate levels separate entities into major categories and subcategories,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

nested under more abstract superordinate concepts such as *mammal* or *animal*. Thus, abstraction constitutes a graded relation over category structure, ranging from broad conceptual groupings to fine-grained distinctions among visually and semantically similar instances (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 to support informative perceptual and functional expectations, yet general enough to group together many similar instances (Rosch et al., 1976; Rosch, 1978; Murphy, 2002). Superordinate categories provide broader semantic coverage but often discard many diagnostic visual details, whereas subordinate categories preserve fine distinctions but require more specific evidence. Categorization is also 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 (Tversky, 1977; Medin & Schaffer, 1978; Nosofsky, 1986; Murphy, 2002). This graded structure motivates evaluating representations by how they organize similarities across levels of abstraction.

3. Results

3.1. Semantic abstraction is spectrally stratified in vision-language models

We first constructed a controlled probing taxonomy to measure abstraction across multiple semantic depths. The taxonomy contains four levels, ranging from the broad animate–inanimate distinction to fine-grained object categories, and groups visual concepts under nested superordinate, intermediate, basic-level, and subordinate labels (Figure 1). Because each instance carries a label at every depth, the same set of embeddings can be evaluated against four reference partitions of varying granularity. This design provides a reference hierarchy for asking whether VLM embeddings encode different levels of abstraction at different locations along the principal-component spectrum, while holding the image set fixed across depths.

We then evaluated category structure across rank-truncated subspaces of the CLIP B/32 image embedding. For each retained rank k, we projected embeddings onto the top-k PCs and computed 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. The reference taxonomic distance matrix at depth 2 already exhibits clear block structure across Human, Animal, Tool, and Vehicle (Figure 2a), which serves as the target geometry for the lowest superordinate partition. Although cumulative explained variance grew smoothly with k and reached roughly 80% only after several tens of components, silhouette (Rousseeuw, 1987) curves for different depths peaked at sharply distinct ranks (Figure 2b). The shallowest distinction was recovered by the very first PC (Figure 2b, $k^* = 1$, 95% CI [1, 2]), depth-2 superordinates peaked near the top three PCs (Figure 2b, $k^* = 3$, 95% CI [3, 3]), depth-3 intermediate groupings near rank 13 (Figure 2b, $k^* = 13$, 95% CI [10, 13]), and depth-4 fine-grained categories required roughly 30 components (Figure 2b, $k^* = 29$, 95% CI [29, 33]). The peak rank thus increased monotonically with taxonomy depth (Figure 2c), indicating that abstraction is sequentially organized along the spectrum rather than uniformly distributed across the embedding.

This rank-dependent organization was also visible in the geometry of the embeddings. At the lowest retained ranks, representational distance matrices were dominated by a sin-

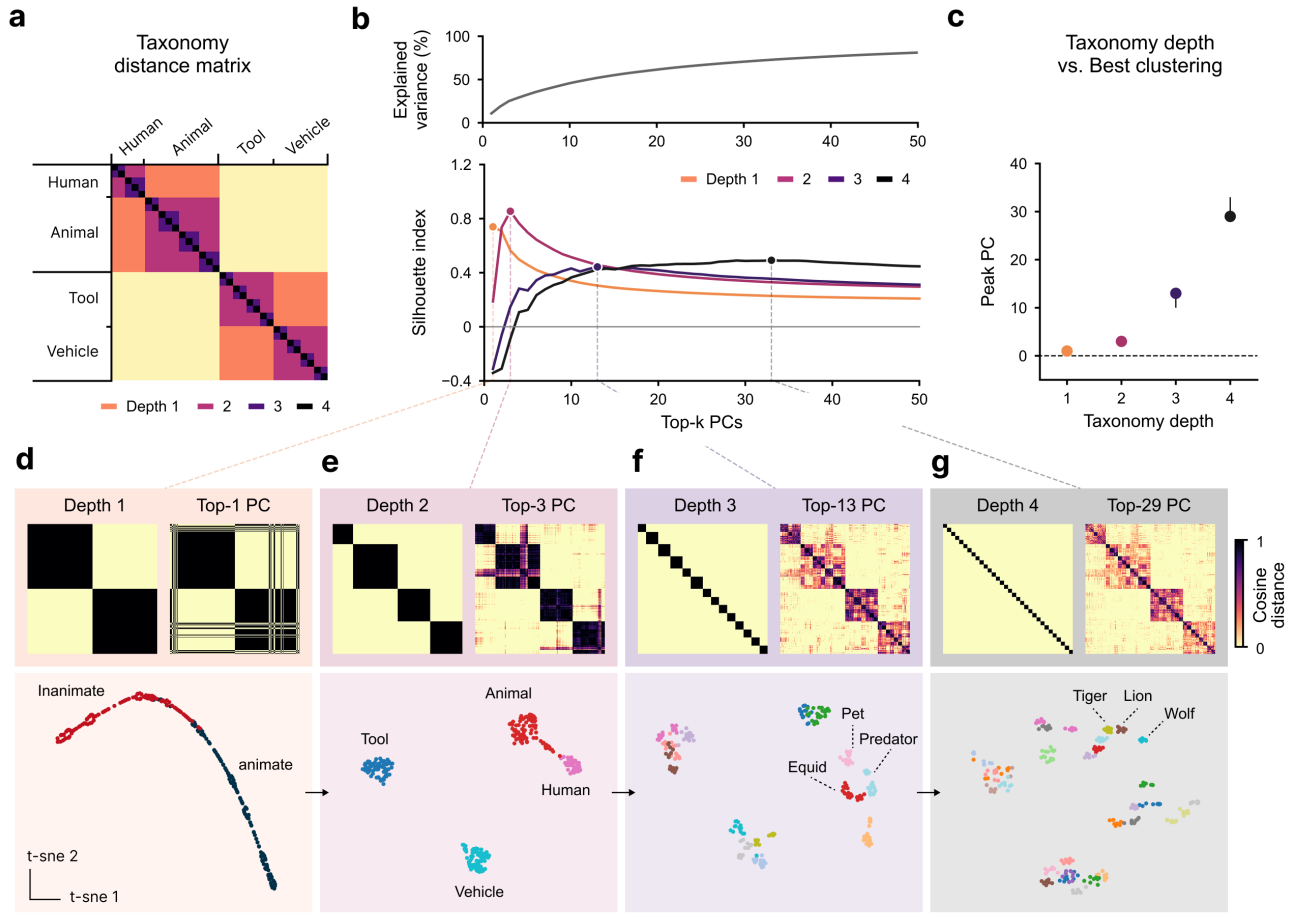

Figure 2. **Category abstraction shifts systematically across the principal-component spectrum. a,** Reference taxonomic distance matrix derived from the four-level probing hierarchy. **b,** Top: cumulative explained variance as a function of retained principal components in the CLIP B/32 image embedding. 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 fine-grained category structure.

gle coarse partition: cosine distances were small within the animate and inanimate halves and large between them, mirroring the depth-1 reference matrix. As k increased, finer block structure emerged within these halves, first separating the four superordinate categories at rank 3, then resolving intermediate groupings such as Pet, Predator, and Equid at rank 13, and finally exposing instance-level clusters at rank 29 (Figure 2d–g, top). Low-dimensional t-SNE projections (van der Maaten & Hinton, 2008) of the same rank-truncated embeddings showed a parallel transition (Figure 2d–g, bottom). At depth 1, points formed a single curve with animate and inanimate exemplars at opposite ends; at depth 2, four well-separated clusters emerged; at depth 3, these clusters fragmented into nested sub-clusters; and at depth 4, fine-grained categories such as Wolf, Lion, and Tiger became distinguishable.

To validate that spectrally stratified semantic abstraction is not merely a consequence of arbitrary grouping but instead emerges from the semantic hierarchy, we tested a shuffled taxonomy that preserves the tree structure while breaking the concept–label correspondence (Supplementary Results A.1). The original strictly monotonic peak-k^* ordering collapsed under shuffled labels into near-flat or non-monotonic trajectories, implying that our results are not simply an artifact of arbitrary grouping. These results suggest that pretrained VLM embeddings are not semantically homogeneous. Coarse conceptual structure is concentrated in a small set of leading directions that account for a modest fraction of variance, whereas progressively finer category distinctions emerge from higher-rank components. Abstraction is therefore spectrally stratified within the embedding, rather than encoded uniformly across all dimensions, and the principal-component spectrum can be interpreted as a graded axis from broad to fine semantic structure.

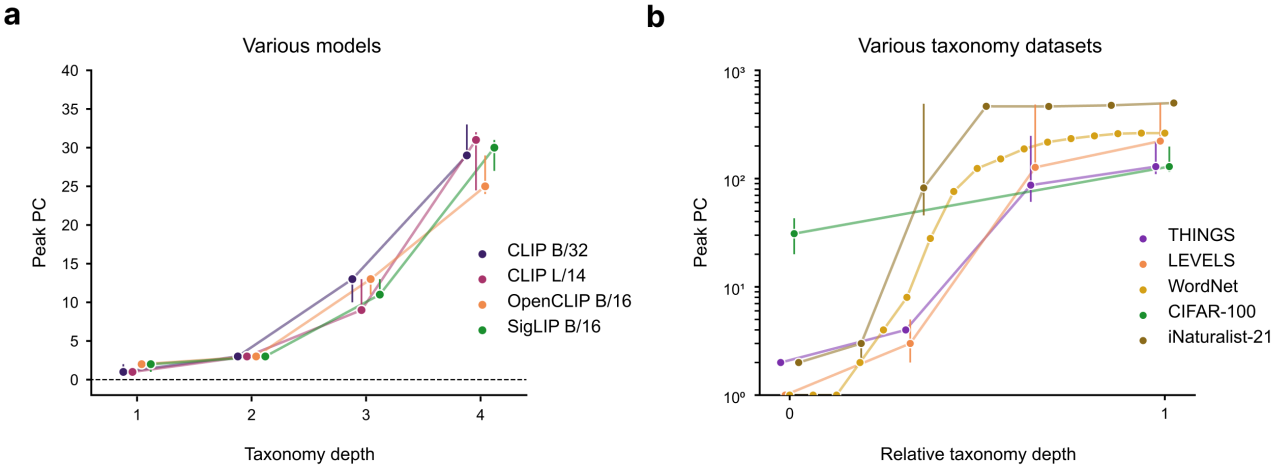

Figure 3. **Spectral abstraction generalizes across VLM architectures and taxonomic datasets. a,** Peak silhouette rank as a function of taxonomy depth across CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16, evaluated on the four-level probing taxonomy. Although absolute peak ranks differ across architectures, all models exhibit a monotonic increase in peak rank with taxonomy depth. **b,** Peak rank as a function of relative taxonomy depth across THINGS, LEVELS, WordNet, CIFAR-100, and iNaturalist-21, which differ in category coverage, hierarchy depth, and granularity. Relative depth normalizes hierarchies of unequal length to a common scale. Across datasets, deeper levels are consistently associated with higher peak ranks.

3.2. Spectral abstraction generalizes across models and taxonomies

We next asked whether this spectral organization was specific to a single VLM or a single taxonomy. Repeating the analysis across CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16, all four backbones produced the same pattern on the probing taxonomy: peak rank increased monotonically with taxonomy depth (Figure 3a; Table S1). Differences in architecture and training details modulated the absolute peak ranks but did not alter the rank ordering across depths.

The same pattern also generalized across taxonomic datasets. We repeated the analysis on THINGS (Hebart et al., 2019; 2023), LEVELS (Muttenthaler et al., 2025), WordNet (Miller, 1995), CIFAR-100 (Krizhevsky, 2009), and iNaturalist-21 (van Horn et al., 2018), which differ in category coverage, hierarchy depth, and granularity. Because these datasets have different numbers of levels, we compared peak rank as a function of relative taxonomy depth, normalizing each hierarchy to the unit interval. Across all five datasets, deeper relative depths were associated with substantially higher peak ranks (Figure 3b). The coarsest levels were typically recovered at single-digit ranks, whereas the finest levels required hundreds of components in iNaturalist-21 and WordNet, where leaf categories are most numerous and visually similar.

A natural follow-up question is whether this spectrally stratified abstraction is a characteristic of vision-language models or can emerge through vision-only training. We therefore repeated the analysis on self-supervised image models (DINOv2 ViT-B/14, MAE ViT-B/16) and a supervised ViT-B/16 (Supplementary Results A.2). These vision-only embeddings preserved fine-grained instance clustering but failed to recover the cleanly stratified higher-level hierarchy seen in contrastive backbones, suggesting that hierarchical abstraction depends on multi-modal learning. Together, these results support the view that rank-ordered spectral structure provides a robust operational axis for indexing abstraction across evaluated vision-language models and diverse taxonomic structures.

3.3. Spectral interventions control abstraction-dependent behavior

The preceding analyses establish that abstraction is organized across the PC spectrum. To examine whether this organization governs model behavior, we asked whether rank-based interventions on the embedding produce predictable shifts in retrieval. On the probing taxonomy, image-to-text retrieval was evaluated separately at each of the four depths while varying the retained top-k PCs. Retrieval curves were strongly depth-dependent: depth-1 retrieval saturated near ceiling with only two or three components (Figure 4a, top-1 $= 0.955$ at $k = 2$ vs. max 0.997), depth-2 retrieval reached its plateau by roughly the top ten PCs, and depth-3 and depth-4 retrieval improved gradually and continued to benefit from additional components well beyond rank 30 (Figure 4a, depth-4 max 0.742 at $k = 406$; Table S3). In addition, to localize where in the spectrum each abstraction level resides rather than how many leading components are required to span it, we evaluated each PC band $[k_{\mathrm{lo}}, k_{\mathrm{hi}})$ in isolation (Supplementary Results A.3). The silhouette

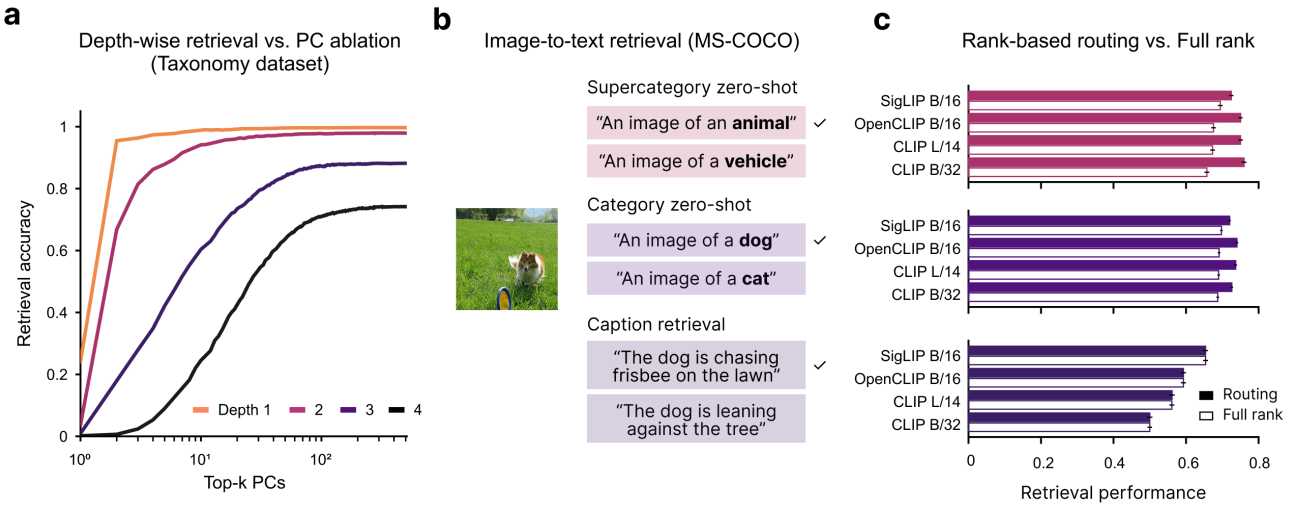

Figure 4. **Rank-based spectral interventions selectively control abstraction-dependent retrieval. a,** Image-to-text retrieval accuracy on the probing taxonomy as a function of retained top-k PCs, evaluated separately at depths 1–4. Coarse retrieval saturates with very few leading components, whereas 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.

index at each depth peaked in a strictly later band than its predecessor, confirming that distinct rank ranges support distinct abstraction levels.

To test whether rank selection generalizes beyond our taxonomy, we evaluated retrieval on MS-COCO with prompts constructed 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 fine-grained caption semantics. For each abstraction level, we compared the full-rank embedding with a rank-routed embedding whose retained subspace was selected to match the target depth, using a held-out validation split to choose k. Across all four backbones, rank routing produced clear gains for supercategory and category retrieval, whereas caption-level retrieval showed 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 (Table S4).

These results suggest that the spectral stratification observed in the previous sections is not merely a descriptive property of the embedding geometry but a functional axis for controlling behavior. Selecting a subspace matched to the target abstraction level recovers structure that is relevant for that level while suppressing higher-rank residuals that introduce noise into coarse retrieval. The PC spectrum therefore serves as both a diagnostic and an intervention: rank truncation acts as a tunable filter on semantic granularity without retraining or modifying model parameters.

3.4. Compact spectral 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 asked whether taxonomic and human-similarity structure can be preserved after removing high-rank residual components. Spearman correlation between embedding distances and the reference taxonomic distance matrix peaked at a low retained rank and declined slightly thereafter, showing that compact subspaces preserve hierarchical category structure while truncating high-rank residuals (Figure 5a, $k^* = 5$, $\rho = 0.540$ vs. full-rank $\rho = 0.537$). Alignment with human behavioral similarity on THINGS showed a similar profile, with peak correlation at a comparable rank and only marginal change at higher ranks (Figure 5b, $k^* = 6$, $\rho = 0.762$ vs. full-rank $\rho = 0.755$). These results indicate that high-rank residuals 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 constructed a dimensionality-matched random-subspace control by sampling k PCs uniformly at random (Supplementary Results A.4). Random subspaces of the same dimensionality fell be-

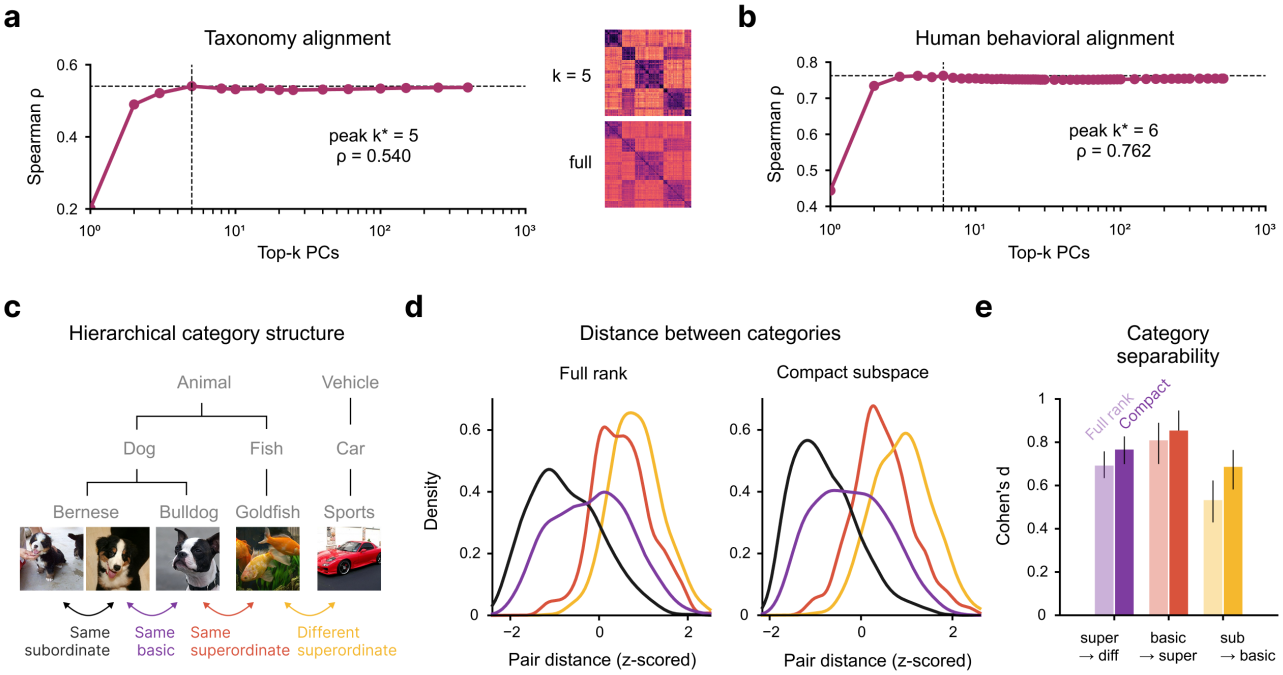

Figure 5. **Compact spectral subspaces preserve hierarchical category structure and human-aligned similarity. a,** Spearman correlation between embedding-derived pairwise distances and the reference taxonomic distance matrix as a function of retained top-k PCs. Alignment peaks at a compact rank ($k^* = 5$, $\rho = 0.540$) rather than at the full rank. 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. Behavioral alignment also peaks at a compact rank ($k^* = 6$, $\rho = 0.762$). **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). The compact subspace produces sharper separation between pair types. **e,** Category separability quantified as Cohen's d between adjacent hierarchical relations: superordinate vs. different-superordinate, basic vs. superordinate, and subordinate vs. basic. Compact subspaces yield larger separability than the full-rank embedding across all three comparisons.

low the full-rank value at every rank on both taxonomy and human-similarity readouts, whereas the leading compact spectral subspace preserved the alignment structure.

To understand what this compact subspace preserves, we examined pairwise category distances stratified by hierarchical relation. Using a small reference hierarchy (Animal vs. Vehicle at the superordinate level; Dog, Fish, and Car at the basic level; Bernese, Bulldog, Goldfish, and Sports car at the subordinate level), we classified 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 were broad and substantially overlapping, particularly between same-basic and same-superordinate pairs. In the compact spectral subspace, the distributions shifted apart and narrowed, producing sharper separation organized strictly by hierarchical distance (Figure 5d). Quantifying separability with Cohen's d between adjacent hierarchical relations confirmed this pattern: compact subspaces increased separability across superordinate, basic-level, and subordinate comparisons (Figure 5e, superordinate→different-superordinate Cohen's d $0.53 \rightarrow 0.68$; Table S5).

We then evaluated compact subspaces directly against human odd-one-out judgments. On LEVELS, compact subspaces improved 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]). The same improvement held on class-boundary trials, which require deciding whether two of three concepts cross a category boundary (Figure 6b), and on fine-grained trials, which require subordinate-level discrimination (Figure 6c). The benefit was therefore not restricted to the coarsest distinctions but extended across the full range of LEVELS difficulty. Aggregating across both LEVELS and THINGS, every backbone achieves better odd-one-out performance in a compact subspace than in the full embedding space (Figure 6d; Table S6), indicating that spectral compression consistently increased behavioral alignment relative to the full-rank embedding. To confirm that the behavioral gain is not simply due to dimensionality reduction, we again compared the compact spectral subspace against a dimensionality-matched random-PC base-

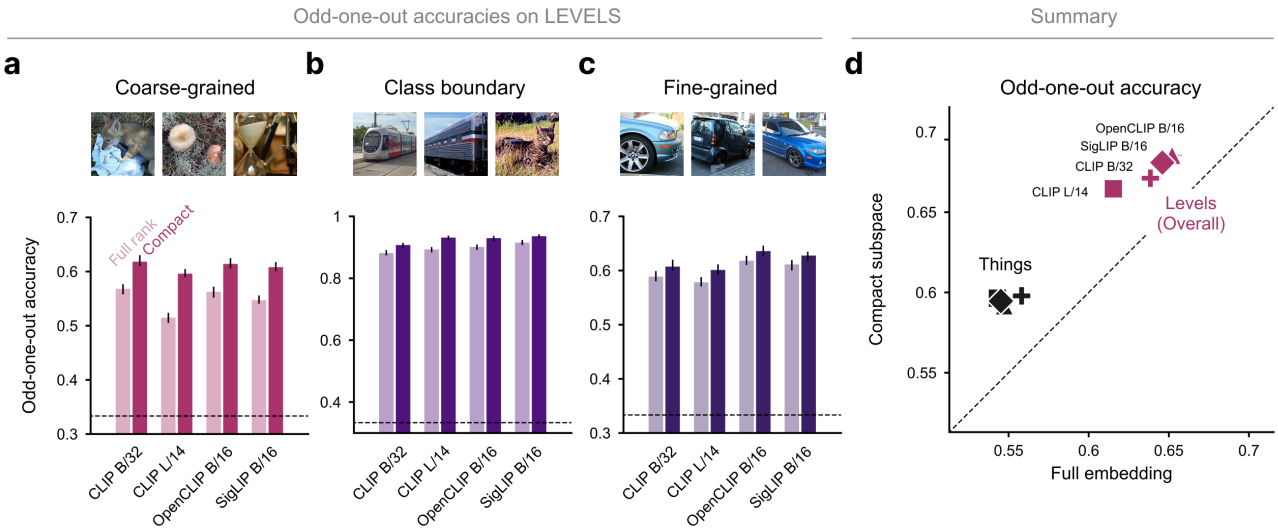

Figure 6. **Compact spectral subspaces improve alignment with human similarity judgments. a–c,** Odd-one-out accuracy on the LEVELS benchmark across coarse-grained (a), class-boundary (b), and fine-grained (c) trials, evaluated for CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16. Compact spectral 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.

line (Supplementary Results A.4). Random subspaces with the same number of components performed 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.

These results suggest that human-aligned semantic structure is not a property of the full embedding but a recoverable substructure of it. A compact spectral subspace, identified without any human supervision and without modifying model parameters, captures hierarchical category structure more faithfully than the full-rank representation and aligns more closely with human similarity judgments across multiple datasets and backbones. The remaining high-rank components carry information that is useful for fine-grained perceptual matching but appears to dilute the coarse, taxonomically organized structure that humans rely on when judging conceptual similarity.

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 models and taxonomies, and it has behavioral consequences: retaining leading components preserves coarse retrieval while degrading 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 spectral subspaces also improve alignment with human similarity judgments relative to the full embedding. 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 such as CLIP, ALIGN, OpenCLIP, and SigLIP have been studied primarily as transferable zero-shot representations (Radford et al., 2021; Jia et al., 2021; Cherti et al., 2023; Zhai et al., 2023), while interpretability work has often focused on identifying human-interpretable units or concepts inside these models. For example, prior work identified multimodal neurons in CLIP (Goh et al., 2021), and language supervision has been used to automatically describe hidden units in vision networks (Oikarinen & Weng, 2023). Other work evaluated full-representation alignment between models and human judgments or neural responses (Peterson et al., 2018; Muttenthaler et al., 2023a; Schrimpf et al., 2018). Our contribution differs in that 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 (Mu &

Viswanath, 2018); in VLMs, leading components can carry meaningful abstract category structure, and removing them selectively damages coarse semantic organization.

The spectral hierarchy we observe also connects to cognitive and neuroscientific accounts of category structure. Cognitive theories have long argued that concepts are organized across levels of abstraction, from superordinate to basic-level and subordinate categories (Collins & Quillian, 1969; Rosch et al., 1976; Rosch, 1978; Murphy, 2002). Similarity-based theories further emphasize that categorization depends on graded relations among objects rather than discrete labels alone (Tversky, 1977; Nosofsky, 1986). In neuroscience, representational similarity analyses comparing models with biological systems (Kriegeskorte et al., 2008a; Nili et al., 2014) have shown that human and primate IT representations contain both categorical and continuous structure (Kriegeskorte et al., 2008b; Konkle & Caramazza, 2013). Our results suggest that pretrained VLMs contain a spectral organization that can be compared with these behavioral and neural structures. We do not claim that VLMs implement human conceptual mechanisms, nor that principal components correspond directly to neural dimensions. Rather, the PC spectrum provides an operational axis along which machine representations can be related to taxonomic depth and human similarity judgments.

Finally, our results contribute to the growing literature on human-aligned machine learning. Recent work has shown that model representations can be made more human-aligned by using human behavioral data as an explicit alignment target. A recent study trained a teacher model to imitate human judgments and then fine-tuned foundation models through a human-aligned distillation objective, obtaining large improvements across multi-level similarity tasks (Muttenthaler et al., 2025). In contrast, our method does not introduce a human-alignment objective, alter model parameters, or fine-tune on human judgments; it simply truncates the spectrum of a 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: human-aligned structure is already present in pretrained VLMs as a recoverable spectral substructure, even when it may be obscured in the full-rank embedding.

Acknowledgements

Marin Vogelsang is supported by the Japan Society for the Promotion of Science (JSPS) Overseas Research Fellowship and the Yamada Science Foundation, and Lukas Vogelsang is supported by a grant from the Simons Foundation International to the Simons Center for the Social Brain at MIT.

Impact Statement

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

References

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Collins, A. M. and Quillian, M. R. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247, 1969.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26, 2013.

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. Multimodal neurons in artificial neural networks. *Distill*, 2021.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., van Wicklin, C., and Baker, C. I. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS one*, 14(10):e0223792, 2019.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C. I. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Jolliffe, I. T. Principal components in regression analysis. In *Principal component analysis*, pp. 129–155. Springer, 1986.

Konkle, T. and Caramazza, A. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

Kriegeskorte, N. and Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008a.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

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

Medin, D. L. and Schaffer, M. M. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

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

Murphy, G. *The big book of concepts*. MIT press, 2002.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., and Kornblith, S. Human alignment of neural network representations. In *International Conference on Learning Representations*, 2023a.

Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A., and Kornblith, S. Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems*, 36:50978–51007, 2023b.

Muttenthaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M. C., Müller, K.-R., Unterthiner, T., and Lampinen, A. K. Aligning machine and human visual representations across abstraction levels. *Nature*, 647: 349–355, 2025.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4): e1003553, 2014.

Nosofsky, R. M. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57, 1986.

Oikarinen, T. and Weng, T.-W. CLIP-dissect: Automatic eescription of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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(11): 559–572, 1901.

Peterson, J. C., Abbott, J. T., and Griffiths, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Rogers, T. T. and McClelland, J. L. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

Rosch, E. Principles of categorization. In *Cognition and Categorization*, pp. 27–48. Lawrence Erlbaum, 1978.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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.

Smith, E. E. and Medin, D. L. *Categories and concepts*. Harvard University Press, 1981.

Tversky, A. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11941–11952, 2023.

A. Supplementary results

A.1. Control experiment with randomly permuted taxonomic structure

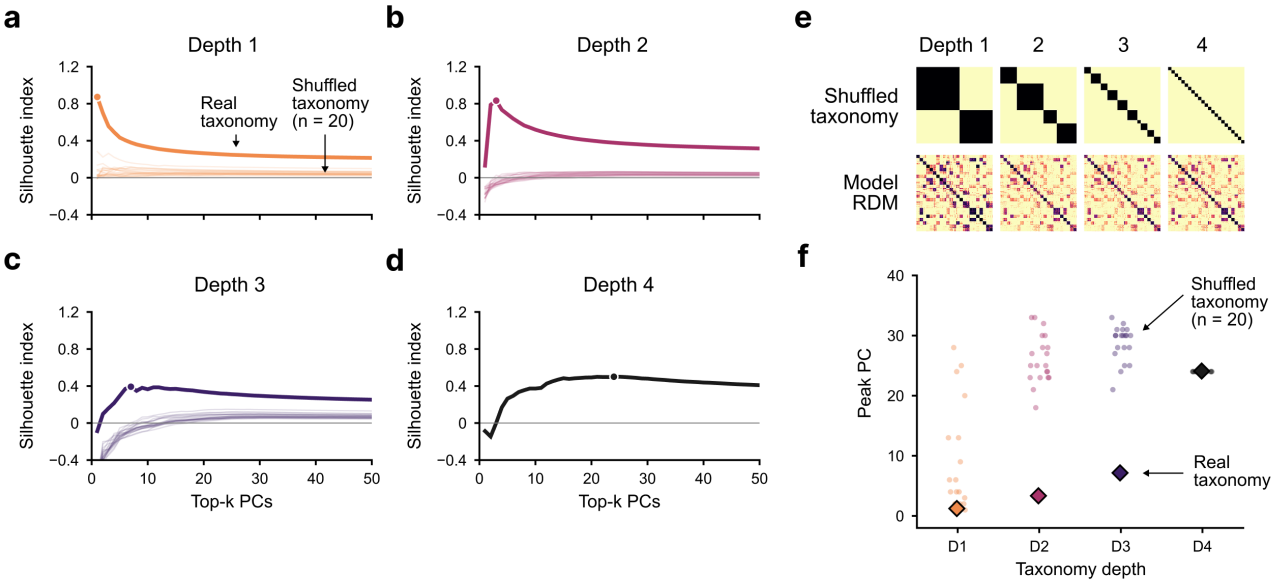

Figure S1. **Spectral abstraction does not emerge under randomly permuted taxonomic labels. a-d,** Silhouette index versus retained top-k PCs at depths 1–4 on CLIP B/32 with the original taxonomy (thick lines) and 20 shuffled-label repeats (thin lines, one per seed). **e,** Depth-wise representational distance matrices ordered lexicographically by the original depth-1–4 labels: target taxonomy (top) and rank-truncated cosine RDM at the original k^* (bottom). **f,** Per-depth peak silhouette rank (k^*) across the 20 shuffles (dots) with the original k^* overlaid as a diamond.

We tested whether the rank-stratified silhouette pattern depends on the actual taxonomic labels by randomly permuting concept-to-path mappings, which preserves the tree structure (same set of paths, same per-depth label histogram) but breaks the semantic correspondence between images and labels. Across 20 shuffle seeds, the peak-k^* ordering broke down: the original taxonomy yielded a strictly monotonic $k^* = 1, 3, 13, 29$ across depths 1–4, whereas shuffled labels produced near-flat or non-monotonic trajectories (Figure S1a–d, mean shuffled $k^* = 8.6 \pm 8.6$, 26.1 ± 3.9, 28.7 ± 2.9, 24.0 ± 0.0 at D1–D4 respectively, with the monotonic depth ordering preserved in only 2/20 seeds). The cosine distance matrix at the original k^* recovered clean depth-wise block structure, but the same projection at a shuffled k^*, ordered by the shuffled labels, exhibited no block organization (Figure S1e). Spectral stratification therefore tracks the actual taxonomic content of the labels rather than any partition with the same marginal statistics (Figure S1f).

A.2. PC-wise category clustering across vision-language models and vision-only control models

We extended the PC-wise silhouette analysis from the main-text CLIP B/32 backbone to three additional contrastive vision-language models (CLIP L/14, OpenCLIP B/16, SigLIP B/16) and to three vision-only control models (DINOv2 ViT-B/14, MAE ViT-B/16, supervised ViT-B/16). All three additional contrastive backbones reproduced the strict monotonic peak-k^* pattern on the four-depth probing taxonomy (Figure S2a–c, depth-1 $k^* \in \{1, 2\}$ and depth-4 $k^* \in [25, 31]$ across the four contrastive backbones; full per-backbone CIs in Table S1), and their rank-truncated cosine distance matrices progressively recovered the reference hierarchy from coarse animate–inanimate structure at low ranks to instance-level structure at high ranks (Figure S3b–d).

The vision-only controls displayed a qualitatively different pattern: although peak-k^* values still tended to grow with depth, the depth-1 silhouette was substantially lower than in any contrastive model and the rank-truncated cosine distance matrices failed to recover a clean animate–inanimate split at any retained rank (Figure S2d–f, Figure S3e–g, e.g. DINOv2 best $k = 1, 6, 28, 29$ and MAE best $k = 3, 4, 32, 62$ across D1–D4). Fine-grained instance clustering is therefore preserved by visual statistics alone, but the cleanly stratified higher-level conceptual hierarchy that emerges in contrastive models is largely absent in vision-only embeddings, suggesting that hierarchical abstraction depends on multimodal contrastive supervision.

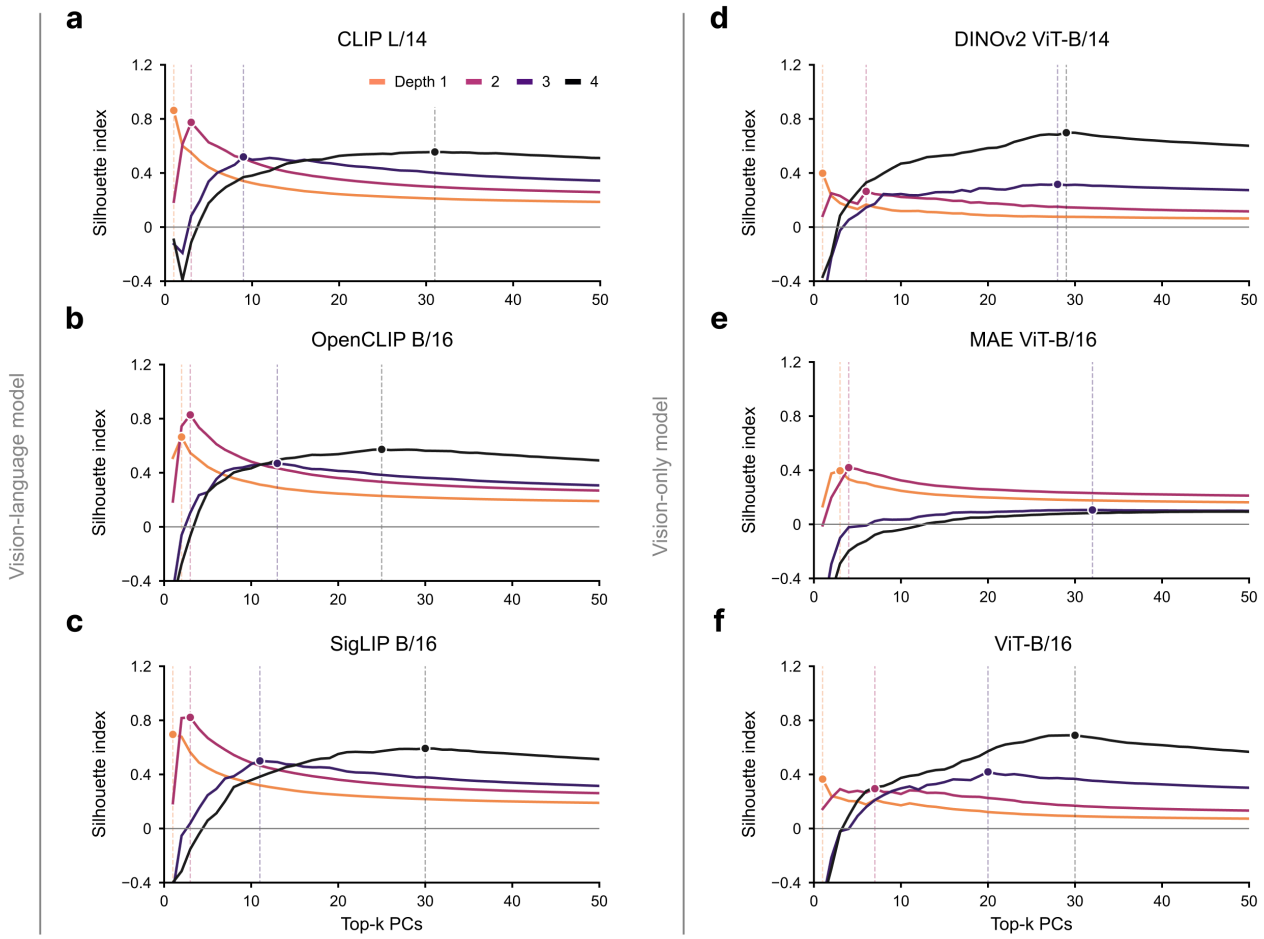

Figure S2. **PC-wise category clustering across vision-language and vision-only models.** Silhouette index of taxonomic clusters at depths 1–4 as a function of retained top-k PCs. **a–c,** Contrastive vision-language models: CLIP L/14, OpenCLIP B/16, and SigLIP B/16. **d–f,** Vision-only control models: DINOv2 ViT-B/14, MAE ViT-B/16, and supervised ViT-B/16. Colored dots and dashed vertical lines mark the peak silhouette rank per depth.

A.3. Band-decomposed PC experiments

To localize where in the spectrum each abstraction level resides, rather than how many leading components are required to span it, we evaluated each PC band $[k_{\text{lo}}, k_{\text{hi}})$ in isolation by zeroing all components outside the band. With band widths growing roughly with the depth of the target partition, the silhouette index at each depth peaked in a strictly later band than its predecessor: D1 at $[0, 1)$ (sil $= 0.74$), D2 at $[1, 3)$ (0.45), D3 at $[3, 8)$ (0.13), and D4 at $[8, 20)$ (0.36) (Figure S4). The same picture holds when bands are evaluated under the cumulative-retention complement, confirming that the spectral-stratification picture from the main text is not specific to a single readout.

A.4. Comparison between compact spectral subspaces and random compact subspaces

To verify that the gain from compact spectral subspaces in the main text (Figure 5, Figure 6) is specific to the spectral identification rather than a generic effect of dimensionality reduction, we constructed a dimensionality-matched random-subspace baseline by sampling k principal components uniformly at random from the same fitted basis (20 trials per backbone), and applied this control to two readouts in turn: alignment between rank-truncated embedding distances and the reference structures (taxonomic and behavioral), and odd-one-out accuracy on LEVELS triplets.

For both alignment readouts, the random-PC baseline tracked or fell below the full-rank value at every retained rank, while the spectral compact subspace exceeded both at and around the validation-selected k^* (Figure S5). Removing the variance-based ordering by sampling components at random therefore removes the alignment peak that characterizes the

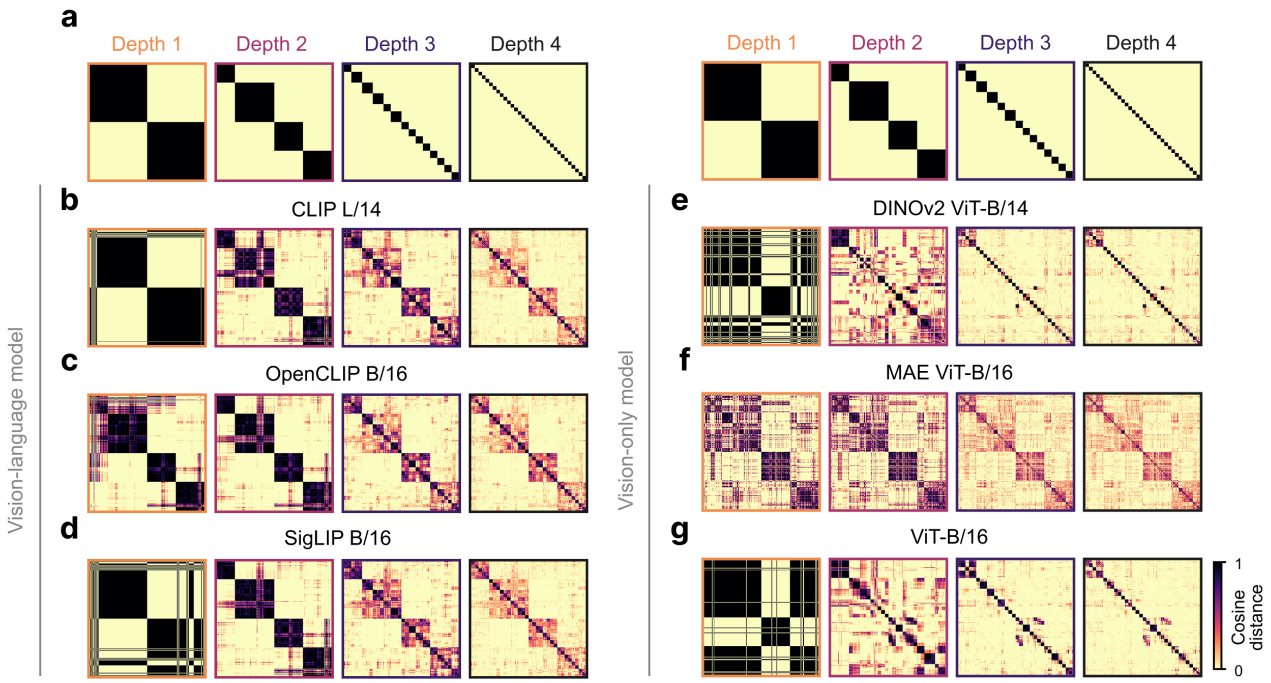

Figure S3. **Rank-truncated representational distance structure across model families. a,** Reference taxonomic distance matrices for depths 1–4. **b–d,** Cosine distance matrices computed from rank-truncated embeddings of contrastive vision-language models (CLIP L/14, OpenCLIP B/16, and SigLIP B/16); each column is evaluated at the silhouette-optimal retained rank for the corresponding depth. **e–g,** Same analysis for vision-only control models (DINOv2 ViT-B/14, MAE ViT-B/16, and supervised ViT-B/16).

spectral curve, even when the number of retained components is held fixed.

The same conclusion held for the LEVELS odd-one-out task: random subspaces with the same number of components as the compact spectral subspace performed essentially at the full-rank level across all four backbones, while the spectral compact subspace remained above both (Figure S6, CLIP B/32 coarse-grained: full 0.568, compact 0.618, random 0.563 ± 0.054; class-boundary: full 0.881, compact 0.907, random 0.886 ± 0.008; fine-grained: full 0.589, compact 0.607, random 0.593 ± 0.006; the same ordering held for the other backbones, e.g. CLIP L/14 coarse-grained random 0.482 ± 0.052 vs. compact 0.596, and SigLIP B/16 coarse-grained random 0.508 ± 0.045 vs. compact 0.608). The behavioral gain from compact subspaces is therefore not produced by reducing dimensionality alone, but by retaining the specific variance-ranked components that carry the relevant abstraction.

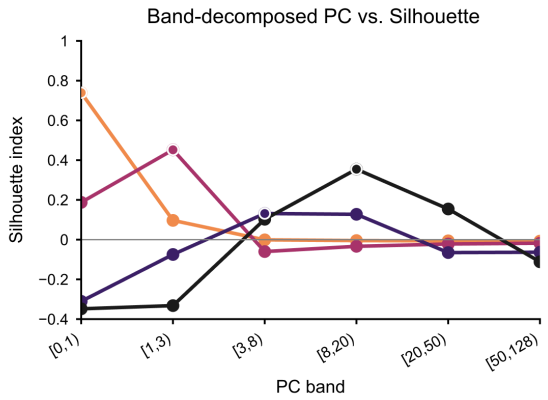

Figure S4. **Band-pass decomposition of taxonomic information across the PC spectrum.** Silhouette index at depths 1–4 as a function of PC band $[k_{\mathrm{lo}}, k_{\mathrm{hi}})$ evaluated in isolation. Filled markers indicate the peak band per depth.

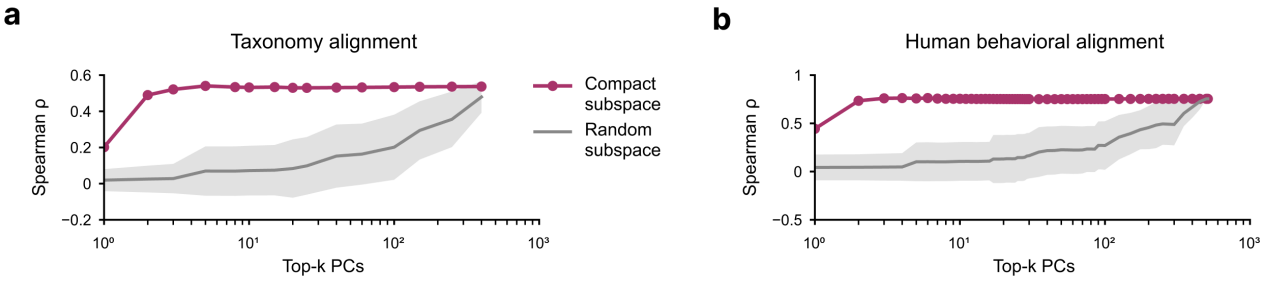

Figure S5. **Compact spectral subspaces vs. dimensionality-matched random subspaces on taxonomy and human alignment. a,** Spearman correlation between embedding-derived pairwise distances and the reference taxonomic distance matrix as a function of retained rank, comparing top-k PCs (compact spectral subspace) with k principal components sampled uniformly at random from the same fitted basis (matched-dimensionality random control). **b,** Spearman correlation with human similarity judgments on THINGS as a function of retained rank, with the same compact-vs-random comparison.

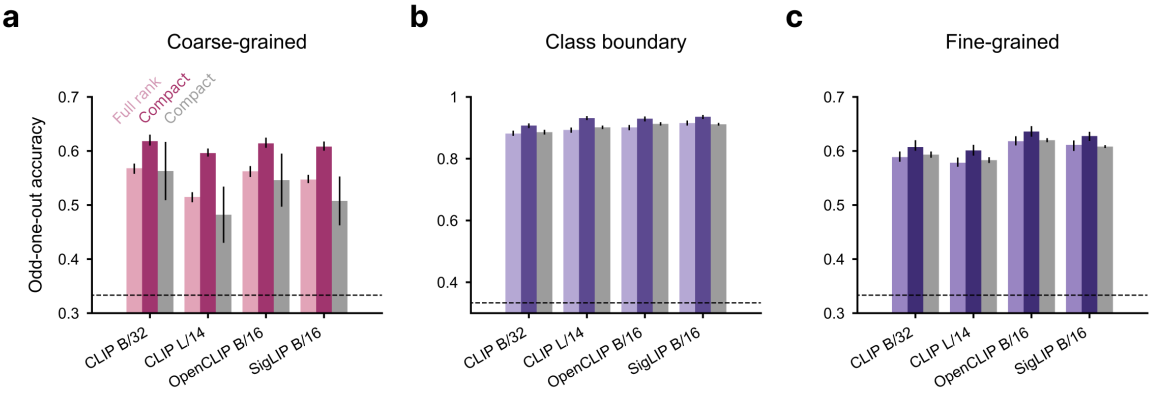

Figure S6. **Compact spectral subspaces vs. dimensionality-matched random subspaces on LEVELS odd-one-out. a–c,** Odd-one-out accuracy on coarse-grained (a), class-boundary (b), and fine-grained (c) trials for CLIP B/32, CLIP L/14, OpenCLIP B/16, and SigLIP B/16. Bars compare the full-rank embedding (light), the validation-selected compact spectral subspace (dark), and a random subspace of the same dimensionality (gray; error bars are ± 1 SD over 20 random PC subsets per backbone). Dashed lines indicate chance.

B. Numerical values for main results

Table S1. Peak silhouette rank k^* per taxonomy depth on the four-level probing taxonomy (complement to Figure 3a). 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 S2. Peak silhouette rank k^* per taxonomy depth on various datasets, CLIP B/32 (complement to Figure 3b). Median bootstrap value; brackets give 95% CI. All confidence intervals are 95% bootstrap CIs from $n = 100$ resamples.

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 †]

† The bootstrap CI saturates at the 500-component sweep cap.

Table S3. Image-to-text retrieval accuracy on the probing taxonomy as a function of retained top-k PCs, CLIP B/32 on the four-level probing taxonomy (complement to Figure 4a).

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 S4. 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 S5. Pair-distance distributions and adjacent Cohen's d at full-rank vs. compact subspace ($k = 107$), 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 S6. Odd-one-out accuracy, full-rank vs. compact subspace at validation-selected k^* (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 pretrained, 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 (Radford et al., 2021; Zhai et al., 2023; Cherti et al., 2023). 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 $s_j^2/(n-1)$ (Pearson, 1901; Hotelling, 1933; Jolliffe, 1986). 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 $k = D$.

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. 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^*.

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.

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`, $D = 512$) (Radford et al., 2021),
- CLIP L/14 (`openai/clip-vit-large-patch14`, $D = 768$) (Radford et al., 2021),
- OpenCLIP B/16 (ViT-B-16 with LAION-2B pretraining, $D = 512$) (Cherti et al., 2023),
- SigLIP B/16 (`google/siglip-base-patch16-224`, $D = 768$) (Zhai et al., 2023).

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) (Oquab et al., 2023),
- MAE ViT-B/16 (masked-autoencoder pretraining) (He et al., 2022),
- ViT-B/16 (ImageNet supervised pretraining) (Dosovitskiy et al., 2021).

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 (Lin et al., 2014) 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 (Hebart et al., 2019; 2023) 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 (Muttenthaler et al., 2025) 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 (Miller, 1995) contributes hypernym chains for arbitrary leaf concepts. CIFAR-100 (Krizhevsky, 2009) provides a two-level coarse-fine hierarchy (20 coarse → 100 fine), and iNaturalist-21 (van Horn et al., 2018) 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.

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.

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. Supercategory and category labels were encoded using the five-template prompt ensemble. Natural-language MS-COCO captions were encoded directly, without additional prompt templating. 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. 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 subspace controls

A random baseline is matched in dimensionality to the top-k projection. It selects k principal-component indices uniformly at random without replacement from the same fitted basis used by top-k. Thus, the basis vectors are identical to those used in the top-k projection, but the variance-based ordering is discarded. These controls are run with multiple independent seeds, and the results are aggregated using the mean and a 95% bootstrap confidence interval. This baseline 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. Image identities, embedding values, principal-component basis, and rank grid are held fixed; only the supervisory labels change. The procedure is repeated over 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 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.

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,

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

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 (Muttenthaler et al., 2025), 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 (Hebart et al., 2023). Bootstrap 95% confidence intervals are estimated by resampling test triplets with replacement.

C.15. 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).