Coarse-to-Fine Category Structure in Child's-View Model Representations

Changho Shin[1] **Jeonghwan Cheon**[2] **Brenden M. Lake**[1,3]
[1]Department of Computer Science, Princeton University, Princeton, NJ, USA
[2]Independent Researcher, Seoul, Republic of Korea,
[3]Department of Psychology, Princeton University, Princeton, NJ, USA
cs1095@princeton.edu, jeonghwan518@gmail.com, brenden@princeton.edu

Abstract

Humans can categorize the same object at different taxonomic levels, from broad kinds to fine-grained classes, supporting generalizations that vary in scope and specificity. Classical connectionist models show how these distinctions can emerge from shared patterns of object properties. Here, we examine how hierarchical category organization develops in self-supervised vision models trained on naturalistic child's-view video. We find that broad category distinctions emerge earlier than finer ones during training. In the learned representations, broad distinctions are most strongly expressed in leading principal components, while finer distinctions become apparent as additional components are retained. Compared with conventionally trained models, child's-view models show stronger separation at the broadest category level but weaker separation at finer levels, consistent with the distinctive statistics of children's visual environments. Their representations align more strongly with human perceptual categorization, while language grounding modestly improves conceptual and contextual categorization. Together, these findings illustrate how hierarchical category organization emerges from naturalistic child's-view learning and how visual and linguistic experience may contribute complementary aspects of category organization to learned representations.

1 Introduction

Categorization allows knowledge about one object to be generalized to others, but the scope of that generalization depends on the level at which the object is categorized. A terrier is also a dog, a mammal, and an animal. Broader categories apply across more objects, whereas finer categories provide more specific information but require distinguishing closely related objects. Representations for categorization must therefore support distinctions at several levels of specificity [Rosch et al., 1976, Mervis et al., 1981, Murphy and Brownell, 1985, Murphy, 2002].

Connectionist models of semantic development learn to predict objects' properties from their identities, using shared hidden representations [McClelland and Rogers, 2003, Rogers and McClelland, 2004]. Although no category tree is explicitly supplied, objects with shared properties develop similar representations. During training, broad groups separate before more closely related concepts become differentiated [McClelland and Rogers, 2003, Rogers and McClelland, 2004, Saxe et al., 2019]. These findings suggest a possible account of how children acquire category knowledge, raising the question of whether similar organization can emerge from their everyday visual experience.

Recent work shows related organization in modern vision–language models: broad distinctions are accessible from dominant representational directions, while finer distinctions become apparent as additional dimensions are retained [Cheon et al., 2026]. However, the training data used in these studies consist of structured object–property associations, category-curated images, or web-scale

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Developmental Perspectives on AI.

image collections [Deng et al., 2009, Radford et al., 2021, Cherti et al., 2023]. Whether comparable organization emerges from the egocentric visual input available to children remains less understood.

Child's-view datasets such as SAYCam and BabyView provide naturalistic first-person recordings of children's everyday experience [Sullivan et al., 2021, Long et al., 2025]. Previous work suggests that the distinctive visual statistics of children's environments may provide a natural curriculum for early object and word learning [Clerkin et al., 2017, Smith et al., 2018]. More recently, objects extracted from BabyView recordings were found to exhibit stronger superordinate category grouping than canonical object images, despite substantial visual variability [Yang et al., 2026]. Models trained on child's-view recordings learn rich visual representations [Orhan and Lake, 2024], while pairing visual input with caregiver speech supports grounded word learning [Vong et al., 2024]. Yet whether and how these properties of naturalistic visual experience are reflected in the hierarchical organization of learned representations remains unclear.

Following Cheon et al. [2026], we examine how category distinctions are distributed across representations learned from child's-view data. We find coarse-to-fine spectral organization, with broad distinctions emerging before finer ones during training. Consistent with findings on children's visual environments [Yang et al., 2026], child's-view models show stronger separation at the broadest level but weaker separation at finer levels than conventionally trained models. Their representations align more closely with human perceptual than conceptual or contextual judgments, while language grounding improves conceptual and contextual alignment. Together, these results characterize the emergence and nature of hierarchical category structure learned from child's-view experience.

2 Related Work

We provide background on the development of category organization, its explanation in connectionist models, and its analysis in learned representations. Additional background appears in Appendix A.

Development of category organization. Infants form categories before they can name them, generalizing across perceptually variable exemplars within the first months of life [Quinn et al., 1993, Murphy, 2002]. Early categorization spans both relatively specific kinds and broad domains, although the categories inferred from infant behavior depend on the exemplars and contrasts used and need not correspond directly to adult basic or superordinate levels [Mandler et al., 1991, Mandler, 2000, Murphy, 2002]. Broad divisions such as animals and vehicles may themselves be supported by systematic perceptual regularities, making it difficult to distinguish perceptual from conceptual organization based on categorization alone. Evidence from induction provides a stronger indication that categories organize knowledge: by late infancy, children generalize non-obvious properties across category members [Baldwin et al., 1993]. These findings support a gradual account in which perceptually grounded categories become increasingly integrated with functional, relational, and linguistic knowledge, rather than a sharp transition from perceptual to conceptual representations [Quinn and Eimas, 1997, Murphy, 2002].

Category organization in learned representations. Connectionist models show how taxonomic organization can emerge from shared patterns of properties rather than from an explicitly encoded category tree. Coherent covariation among properties leads to progressive differentiation from broad domains to more specific categories [McClelland and Rogers, 2003, Rogers and McClelland, 2004], with theoretical analyses relating this progression to statistical modes learned at different rates [Saxe et al., 2019, Benjamin et al., 2026]. Related spectral analyses of modern vision and vision–language models find that broad distinctions are expressed in dominant representational directions, while finer distinctions emerge as additional dimensions are retained [Cheon et al., 2026]. More broadly, learned language and vision representations recover aspects of taxonomic structure and human category organization [Fu et al., 2014, Huebner and Willits, 2018, Lovón-Melgarejo et al., 2024, Jozwik et al., 2017, Muttenthaler et al., 2023, Mahner et al., 2025, Muttenthaler et al., 2025].

Learning from child's-view experience. Longitudinal datasets such as SAYCam and BabyView provide naturalistic, first-person recordings of infants' and young children's everyday experience [Sullivan et al., 2021, Long et al., 2025]. Unlike conventionally curated image collections, this input is temporally structured, unevenly distributed across objects and activities, and lacks explicit category labels or deliberate category coverage [Clerkin et al., 2017, Smith et al., 2018]. Self-supervised

models trained on child's-view video nevertheless learn transferable high-level visual representations [Orhan et al., 2020, Orhan and Lake, 2024], while pairing child's-view images with caregiver speech supports grounded word learning and vision–language representations [Vong et al., 2024]. These studies establish that naturalistic developmental input can support rich learned representations, but how category structure is organized across multiple taxonomic levels, how it develops during training, and how it is reshaped by language remain less understood.

3 Characterizing Category Structure in Learned Representations

We study how representations learned from child's-view video support category distinctions at different levels of specificity. Broad distinctions may be expressed in the dominant variation of a representation, while finer distinctions may depend on additional directions. We adopt the cumulative PCA reconstruction and silhouette analysis of Cheon et al. [2026], who observed such a coarse-to-fine pattern in conventionally trained vision and vision–language models.

The analysis uses an image pool with category labels at several taxonomic depths. A depth may use the full pool or a subset selected by the hierarchy. For each model, PCA is fit once on the common pool without category labels. We then reconstruct the representations using progressively more principal components (PCs) and measure category separation at every retained rank. For each depth of the hierarchy, the analysis produces a curve relating category separation to the number of components retained. We summarize this curve by its peak rank and peak value. Peak rank is the number of retained PCs that maximizes mean silhouette, while peak value is the corresponding maximum (Section 3.3). Keeping these quantities distinct allows two models to exhibit the same coarse-to-fine ordering while differing substantially in how clearly they represent finer categories.

3.1 Rank-truncated representations

Let $z_i \in \mathbb{R}^d$ be the representation of image i, and let μ be the mean representation. PCA gives directions $V = [v_1, \ldots, v_d]$, ordered by the amount of variance they explain. For a retained rank k, we reconstruct the representation using the first k directions:

$$\hat{z}_i^{(k)} = \mu + V_k V_k^\top (z_i - \mu), \qquad V_k = [v_1, \ldots, v_k].$$

Retaining only a few components preserves the dominant variation in the representation. Increasing k gradually restores lower-variance variation. The reconstructions are cumulative, so the rank-k representation contains all directions retained at smaller ranks together with the newly added directions. These cumulative reconstructions show how category structure changes as additional representational directions are introduced. Separation at a small k indicates that a distinction is already expressed by the leading, high-variance directions. If separation improves at larger k, additional lower-variance directions contribute to that distinction. We compare peak ranks across depths to test whether broader categories reach their strongest separation with fewer PCs than finer categories.

3.2 Measuring category separation

At depth ℓ, we evaluate a category partition C_ℓ over an image set I_ℓ. The partitions are ordered from broad to fine. Most benchmarks use the same images at every depth. When a depth applies a stricter hierarchy filter, I_ℓ is a subset of a common image pool, and PCA is still fit once on that pool. In either case, the encoder and PCA basis remain fixed across depths.

At each retained rank, we measure category separation using the silhouette index [Rousseeuw, 1987]. For each image, silhouette compares its mean distance to other images in the same category with its mean distance to the nearest competing category. Averaging across images gives a measure of separation for the partition. Silhouette therefore indicates whether images are, on average, closer to members of their assigned category than to members of the nearest competing category.

We compute cosine distance between the rank-k reconstructions as $d_k(i,j) = 1 - \left(\bar{z}_i^{(k)}\right)^\top \bar{z}_j^{(k)}$, where $\bar{z}_i^{(k)} = \hat{z}_i^{(k)} / \|\hat{z}_i^{(k)}\|_2$. For image i at hierarchy depth ℓ, let $A_i^{(\ell)}(k)$ be its mean distance to other images in the same category. Let $B_i^{(\ell)}(k)$ be its mean distance to the nearest alternative category. The

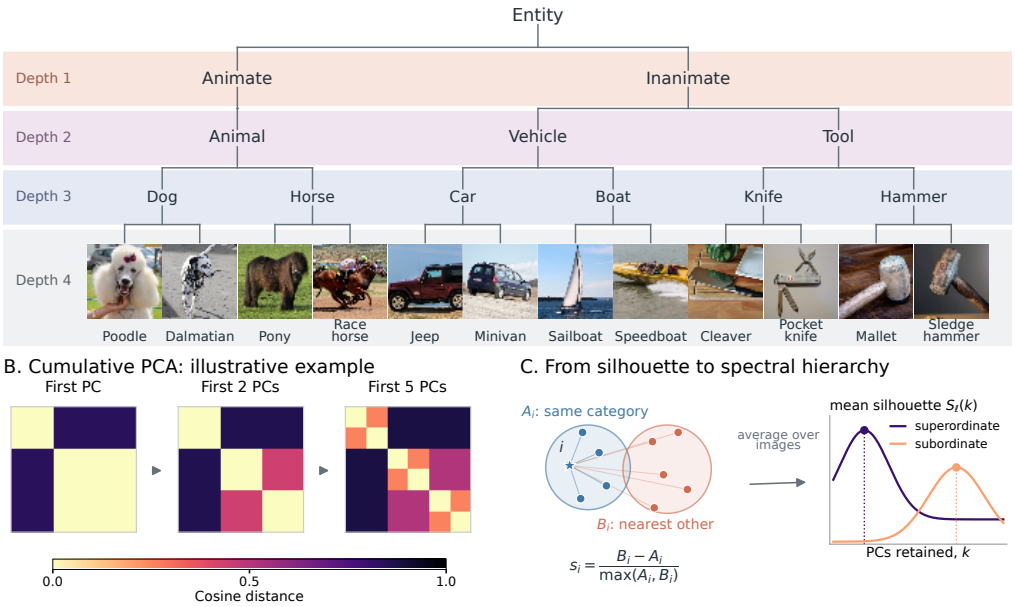


Figure 1: **Characterizing category structure using rank-truncated representations. A:** An image pool can be grouped at several hierarchy depths, from broad partitions to basic categories at Depth 3 and finer subtypes at Depth 4. **B:** Cosine dissimilarity matrices from one illustrative set of synthetic representations, ordered as in A. PCA is fit once; reconstructions retain the first 1, 2, or 5 PCs, restore the mean, and are normalized before computing distances. Darker cells indicate greater dissimilarity. **C:** At each rank k, silhouette compares an image's mean distance to members of its category, A_i, with its mean distance to the nearest competing category, B_i. Averaging across images produces a silhouette curve $S_\ell(k)$ for each hierarchy depth ℓ.

silhouette of image i is

$$s_i^{(\ell)}(k) = \frac{B_i^{(\ell)}(k) - A_i^{(\ell)}(k)}{\max\{A_i^{(\ell)}(k), B_i^{(\ell)}(k)\}},$$

and the mean silhouette at depth ℓ is $S_\ell(k) = n^{-1} \sum_{i=1}^{n} s_i^{(\ell)}(k)$. The index lies in $[-1, 1]$. Positive values indicate that images are closer to members of their assigned category than to the nearest alternative. Values near zero indicate overlapping groups, while negative values indicate that images are closer to another category.

3.3 Peak rank and peak separation

We summarize the silhouette curve for each hierarchy depth by

$$k_\ell^* = \arg\max_k S_\ell(k), \qquad S_\ell^* = S_\ell(k_\ell^*).$$

The peak rank k_ℓ^* identifies where the partition reaches its clearest separation. A small value means that the distinction is already prominent among the leading directions. A larger value means that more dimensions are needed before the distinction is expressed most clearly. The peak does not mark the first rank at which category information is present, nor does category information necessarily disappear after the peak. Silhouette can decline when later dimensions restore variation among members of the same category.

The peak silhouette S_ℓ^* measures the strength of separation at the best rank. Because the number and size of categories differ across hierarchy depths, we use peak silhouette primarily to compare the same depth across models or training conditions. Peak rank cannot be interpreted on its own because even a flat or negative curve has an argmax. We therefore interpret k_ℓ^* only when the peak silhouette is reliably positive and exceeds a label-permutation null. If the maximum occurs at the largest rank examined, we treat the location as a lower bound: the partition requires at least that many dimensions, but its true peak may occur later.

We identify coarse-to-fine ordering when reliable peak ranks increase with hierarchy depth. Peak separation is reported separately to indicate how clearly each depth is represented. We report raw peak rank when representations share the same available rank range. Details of analysis procedure are provided in Appendix B.3.

4 Experiments

The primary objective of our experiments is to characterize how multilevel category organization emerges in representations learned from child's-view experience. First, we test whether child's-view visual models exhibit coarse-to-fine category organization and compare them with conventionally trained models (Section 4.1). Next, we trace category geometry throughout training to examine how broad and fine distinctions develop, inspired by Rogers and McClelland [2004] (Section 4.2). Finally, we compare the learned geometry with human perceptual, conceptual, and contextual similarity judgments and ask how this alignment changes after language grounding (Section 4.3). We first describe the evaluation setup shared across these experiments.

Data. Our main taxonomy analyses use THINGS images from 32 everyday-object concepts, with 12 images per concept unless otherwise noted [Hebart et al., 2019]. These images define four nested partitions containing 2, 4, 14, and 32 groups from Depth 1 to Depth 4. Depth 1 separates animate from inanimate objects, and Depth 2 separates animals, humans, tools, and vehicles. Depth 3 contains intermediate groups such as birds, pets, and road vehicles, while Depth 4 distinguishes individual concepts such as penguin, dog, and car. Appendix B.1 provides THINGS sampling and hierarchy details; Appendix C.1 describes the additional evaluation datasets.

Models. Our primary models are self-supervised vision models trained on SAYCam recordings from all three children [Sullivan et al., 2021, Orhan and Lake, 2024]. Specifically, we use DINO [Caron et al., 2021], which learns through self-distillation without category labels or paired language. For the taxonomy comparison, we evaluate SAYCam-trained ViT-B/14 and ResNeXt-50 models alongside ImageNet-1k DINO ViT-B/16 and ResNet-50 models. Raw pixels provide a baseline for category structure present in the images before learning.

Metric. For each representation and hierarchy depth ℓ, we reconstruct the representation from its leading principal components and compute the mean silhouette curve $S_\ell(k)$ using cosine distance. We report its peak location k^*_ℓ and peak strength S^*_ℓ, as defined in Section 3.3. The main THINGS analyses evaluate every available rank from 1 to 383 and report the raw peak rank. Later peaks at greater depths indicate coarse-to-fine spectral ordering; peak strength measures how clearly the corresponding categories separate. Section 4.3 separately defines the representational-similarity analysis used to compare model and human dissimilarity matrices.

4.1 Child's-view representations show coarse-to-fine organization

We test whether self-supervised vision models trained on SAYCam exhibit coarse-to-fine organization across the four THINGS hierarchy depths.

Setup. We measure category separation across retained PC ranks at each of the four THINGS hierarchy depths. We compare peak ranks to test for coarse-to-fine ordering and peak silhouettes to assess how strongly SAYCam-trained models, ImageNet-trained models, and raw pixels separate categories at each depth.

Results. Both SAYCam models exhibit coarse-to-fine spectral organization similar to conventionally trained models (Figure 2). Broad category distinctions are most strongly expressed in earlier principal components, while finer distinctions peak as additional components are retained. Specifically, peak ranks increase from $k^{=}1$ for animate versus inanimate objects (Depth 1) to $k^{=}46$–47 for individual concepts (Depth 4). However, the models differ in which levels of distinction they separate most strongly. SAYCam-trained models show higher peak silhouettes at the broadest level, whereas conventionally trained models exhibit substantially stronger separation at finer levels. This contrast is consistent with recent findings that naturalistic child's-view environments exhibit distinctive superordinate category structure [Yang et al., 2026].

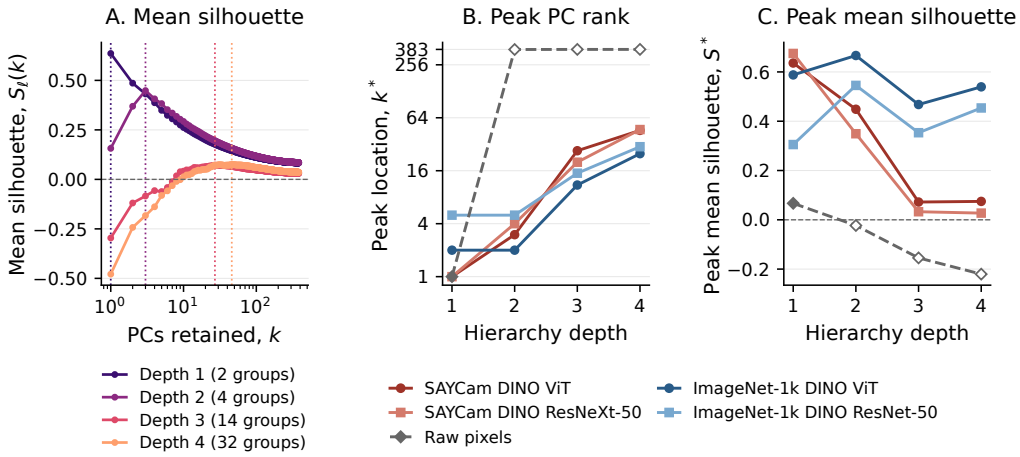


Figure 2: **Child's-view representations show coarse-to-fine organization. A:** Mean silhouette across retained PC rank for DINO-SAYCam ViT-B/14; dotted lines mark k^*_ℓ. **B:** Peak PC rank for child's-view DINO, ImageNet-1k DINO, and raw pixels. All representations share the same 383-PC sample-rank ceiling. **C:** Peak mean silhouette for the same representations. Hollow markers denote non-positive peaks.

4.2 Broad category geometry emerges first

Connectionist theory predicts that broad structure emerges before finer distinctions as stronger statistical regularities are learned earlier [Saxe et al., 2019]. We examine whether this predicted ordering also emerges during child's-view training by tracing category organization across checkpoints.

Setup. We train a ViT-B/14 with DINOv2 on SAYCam videos [Sullivan et al., 2021, Oquab et al., 2024]. Every 200 updates from step 200 to 9,800, we refit PCA and measure peak rank and peak silhouette on THINGS. To visualize the learned hierarchy, we apply average-linkage clustering to distances between category means in the full representation at steps 2,500, 5,000, 7,500, and 10,000 [Sokal et al., 1958]. We cut each tree into 2, 4, and 14 groups, corresponding to Depths 1–3, and measure agreement with the reference taxonomy using the adjusted Rand index (ARI) [Hubert and Arabie, 1985]. An ARI of 1 indicates an exact match with the reference taxonomy at the evaluated depth, while 0 indicates chance-level agreement.

Results. Broad category distinctions emerge before finer ones during training (Figure 3). In the spectral analysis, the four taxonomic depths become reliably expressed in order at 200, 800, 4,000, and 7,400 updates, with final peaks at $k^* = (1, 4, 25, 40)$. This sequential emergence is consistent with progressive differentiation accounts of category development [Rogers and McClelland, 2004, Saxe et al., 2019]. While broad distinctions become strongly separated, finer distinctions remain comparatively weak, even as their spectral organization emerges. This coarse-to-fine structure appears before the hierarchy fully stabilizes in the full representation: broad category groupings eventually become stable, whereas finer distinctions remain less consistently organized.

4.3 Perceptual, conceptual, and contextual structure in child's-view representations

Category structure can be supported by perceptual similarity alone or by broader conceptual knowledge about objects, including their functions, relations, and typical contexts [Murphy, 2002]. We study which aspects of human similarity structure are captured by visual learning from child's-view experience and how language grounding changes this structure.

Setup. We use the human similarity dataset from Kim et al. [2026], which measures perceptual, conceptual, and contextual similarity across 190 familiar THINGS concepts. Participants judged object triplets according to one of these three criteria, and we convert their choices into pairwise dissimilarity matrices. For each model, we compute cosine distances between concept-mean represen-

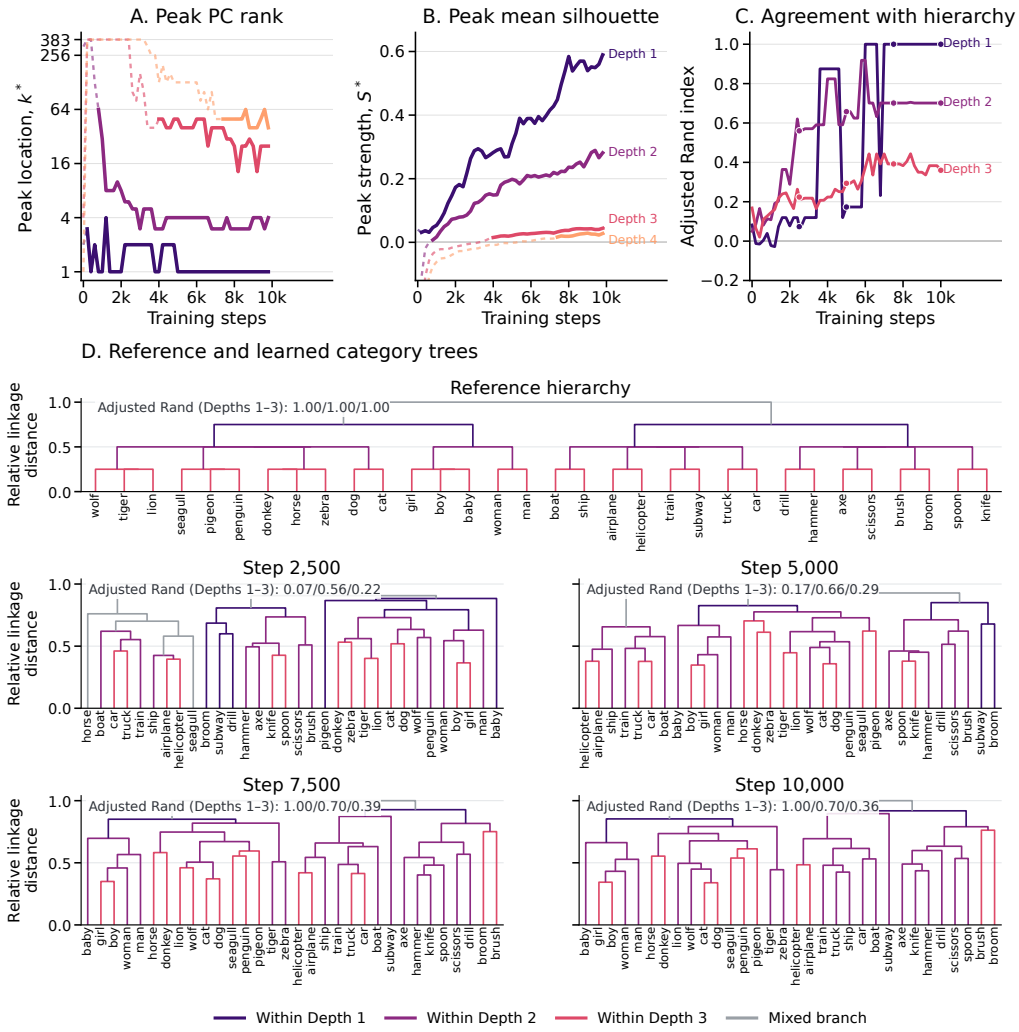


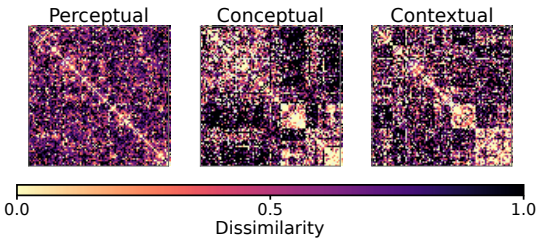
Figure 3: **Broad category geometry emerges first. A–B:** Peak location and strength during early training; dashed segments indicate peaks that are not yet reliably positive. **C:** Adjusted Rand index (ARI) between learned and reference category partitions; circles mark the checkpoints shown in D. **D:** Reference hierarchy and dendrograms obtained by average-linkage clustering of category means in the full representations. Branch colors indicate the deepest shared reference category, and annotations report ARI for Depths 1–3.

tations and correlate the resulting matrix with each human reference using Spearman correlation.[1] We compare SAYCam DINO with a random ViT and an architecture-matched ImageNet-trained DINO. To examine the effect of language grounding on the same child's-view representation, we compare the SAYCam DINO ViT-B/14 with its CVCL grounded projection [Vong et al., 2024]. Additionally, we use matched DINO and CLIP models trained on the same DataComp images [Liu et al., 2025] to test whether a similar shift toward conceptual and contextual alignment occurs in web-scale models. Appendix B.8 provides additional analysis results.

Results. Figure 4 shows that SAYCam DINO already aligns relatively strongly with human perceptual similarity, while its conceptual and contextual alignment is lower. Applying the CVCL projection yields slightly larger gains in conceptual and contextual alignment than in perceptual alignment. The matched DataComp models show a similar trend: CLIP's gains over DINO are slightly larger for conceptual and contextual alignment than for perceptual alignment.

[1]Because within-condition interrater reliability is not available, we report image–word human RDM correlations as a reference: $\rho = .273$, .425, and .345 for perceptual, conceptual, and contextual judgments, respectively.

A. Human judgments of object dissimilarity



B. Spearman correlation

	Perceptual	Conceptual	Contextual
Random init.	.155	.124	.107
ImageNet-1k DINO	.215	.250	.151
SAYCam DINO	.283	.245	.105
+ CVCL	.291	.283	.134
DataComp DINO	.318	.323	.205
DataComp CLIP	.340	.398	.301

Figure 4: **Child's-view representations reflect perceptual and conceptual similarity. A:** Human perceptual, conceptual, and contextual dissimilarity matrices. Darker cells denote more dissimilar concepts. The matrices share an ordering obtained by average-linkage clustering of their mean for display only; the original ordering is used for the correlations. **B:** Spearman correlation between each model dissimilarity matrix and the three human matrices. SAYCam DINO denotes the frozen ViT-B/14 backbone used by CVCL; the next row applies its learned projection. Below the rule, DINO and the final projected CLIP representation use ViT-B/16 models trained on the same DataComp.

4.4 Alignment with infant visual cortex across the representation spectrum

We next ask whether the category structure expressed by leading model components resembles category relationships in infant visual cortex.

Setup. We use the released group-average similarity matrices from ventral visual cortex (VVC) in two-month-old infants and adults in FOUNDCOG [O'Doherty et al., 2026], with adults providing a comparison for the infant results. The original stimulus set contains 36 images from 12 categories, with three exemplars per category. To estimate model spectral structure from more examples, we assemble a separate pool of 1,200 images from Open Images [Kuznetsova et al., 2020], with 100 images per category. These categories define three nested partitions: animate versus inanimate objects (Depth 1); animals, small inanimate objects, and large inanimate objects (Depth 2); and individual categories (Depth 3). We compare SAYCam DINO trained on all three children with an architecture-matched ImageNet-1k DINO, both using ViT-B/14.

For each model, we fit PCA to unit-normalized features from the added images and measure their spectral peaks. We then reconstruct the original stimuli using this fixed basis and its restored mean, following Section 3.1. For both models and VVC, category distances average the nine cross-category image-pair distances: cosine distance for models and one minus the released similarity for VVC. We apply average-linkage clustering to these distances and compare each model tree with the fixed VVC tree for each age group across a common PC-rank grid and the full representation. We cut each pair of trees into the same number of groups and measure their agreement using ARI, averaging across cuts into 2–11 groups.

Results. Peak ranks increase with category depth for both SAYCam and ImageNet DINO, at $k^* = (1, 2, 8)$ and $(1, 3, 12)$, respectively (Figure 5A–B). ImageNet yields stronger category separation, particularly at the finest depth. In contrast, SAYCam has higher mean ARI with infant VVC at every evaluated rank through 128 (C). At PC1, mean ARI is .260 for SAYCam versus .151 for ImageNet. ImageNet has higher mean ARI at ranks 256 and 512 and in the full representation (.123 versus .074).

The adult comparison shows a different rank profile (D). ImageNet has higher mean ARI at PC1 (.265 versus .208) and PC2. SAYCam has higher agreement at the evaluated ranks from 3 through 64, while ImageNet has higher agreement from rank 128 onward and in the full representation (.251 versus .158).

These descriptive results suggest that category relationships resembling infant VVC are more prominent in the leading components of SAYCam DINO. The relative advantage varies with both retained rank and reference age. These model differences have not been tested for statistical reliability or infant specificity, and agreement with group-average category trees does not establish a shared neural spectral hierarchy.

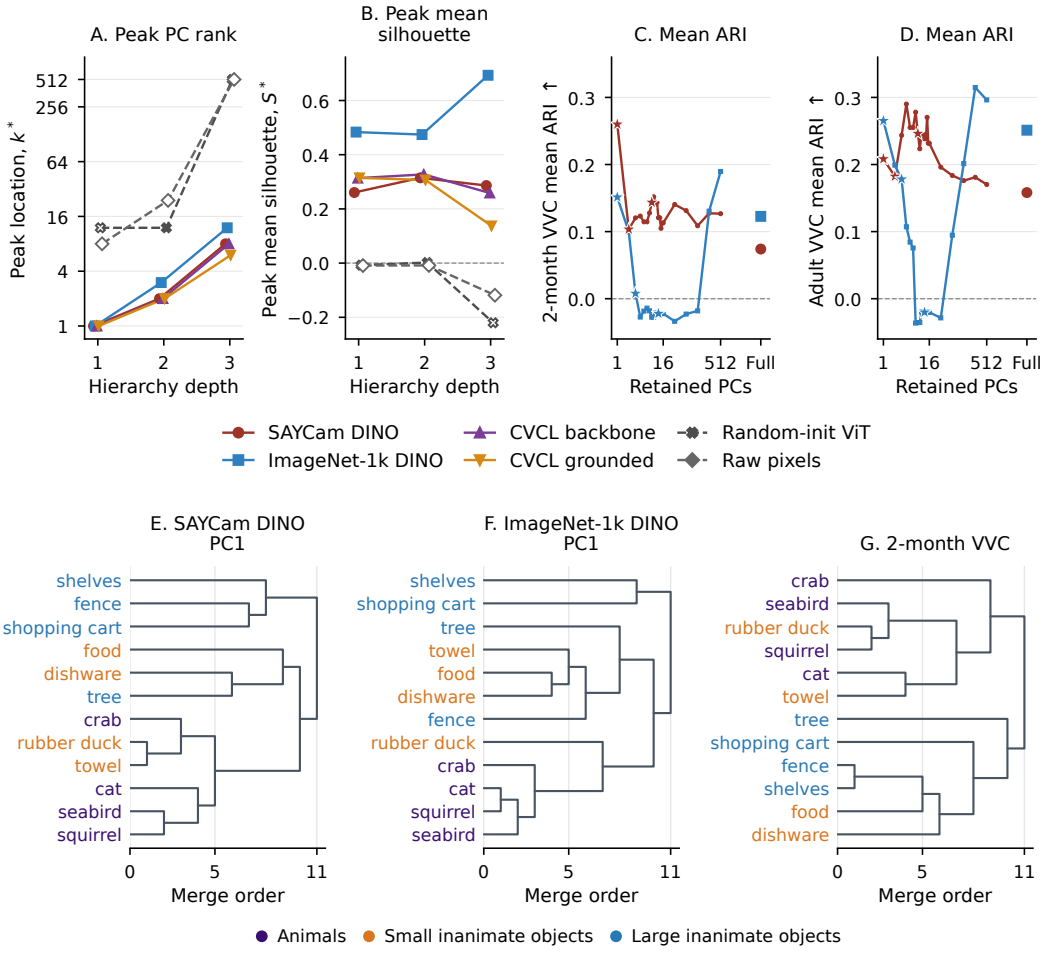


Figure 5: **Spectral organization and category-tree agreement with infant and adult VVC. A–B:** Peak PC rank and peak mean silhouette on the 1,200 added images, with a common 512-PC search ceiling. Hollow markers denote non-positive peaks. CVCL uses child S; the other learned models use all three SAYCam children or ImageNet-1k. **C–D:** Mean ARI with two-month and adult VVC, respectively, on the original 36 stimuli, averaged across cuts into 2–11 groups. Both models use ranks 1–16, 32, 64, 128, 256, and 512 on a logarithmic axis; full features are shown separately. Stars mark each model's spectral peak ranks from A. Higher values indicate greater agreement. **E–G:** Model trees retaining PC1 and the fixed two-month VVC tree. Leaf colors indicate the three broad groups without constraining clustering. Horizontal axes show merge order.

4.5 Routine and role structure in naturalistic child's-view scenes

The preceding category probes use images selected to represent individual objects. We next ask whether a hierarchy organized around everyday routines and object roles is also expressed in naturalistic child's-view scenes.

Setup. We use an exploratory subset of automatically labeled child-S SAYCam frames [Sullivan et al., 2021]. From an augmented pool of 852 frames spanning 71 nouns, we searched 7,627 eight-noun subsets within eight fixed semantic templates using SAYCam discovery features. The meals/play candidate defines three nested partitions—two routines, four roles, and eight nouns—based on typical object uses (Figure 6A). Discovery and evaluation each contain 48 images: six images from three sessions per noun, with no recording sessions shared between splits.

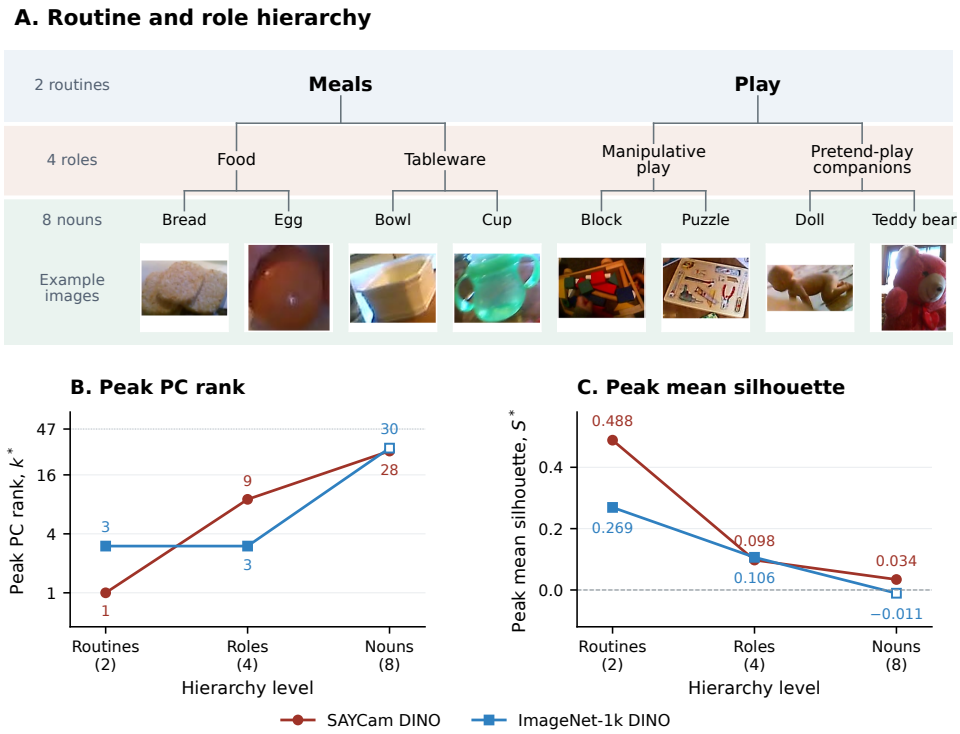


Figure 6: **Routine and role structure in naturalistic child's-view scenes. A:** Reference hierarchy and example crops from the selected image pool. Meals divides into food and tableware; play divides into manipulative play and pretend-play companions. Crops are used only for illustration; the analysis uses complete, unmasked frames. **B–C:** Peak PC rank and peak mean silhouette for SAYCam and ImageNet-1k DINO on the same 48 evaluation images. Both models use local PCA and a 47-PC search ceiling. Hollow markers denote non-positive peaks. These are descriptive estimates for a hierarchy selected using SAYCam discovery features.

We compare SAYCam DINO and architecture-matched ImageNet-1k DINO, both using ViT-B/14. Each model receives complete frames, resized and padded without object masking. We fit PCA to unit-normalized features from the 48 evaluation images and measure mean cosine silhouette at every retained rank from 1 to 47, following Section 3.1.

Results. SAYCam shows increasing peak rank and decreasing, positive peak silhouette across the three levels, at $k^* = (1, 9, 28)$ and $S^* = (.488, .098, .034)$ (Figure 6B–C). ImageNet peaks at $k^* = (3, 3, 30)$, with $S^* = (.269, .106, -.011)$. SAYCam has a larger peak for the broad routine distinction, while noun-level separation remains weak in both models. The SAYCam ordering persists in 20 of 21 checks that remove one evaluation session and refit PCA.

These results are exploratory. The hierarchy was selected using SAYCam features, and the existing evaluation split had been inspected in earlier analyses. A conditional comparison with alternative balanced hierarchies does not survive correction across the eight templates ($p_{\text{Holm}} = .178$). Automatically labeled scenes include depictions and multiple objects, so label errors and shared scene context may contribute to the observed geometry. The analysis does not establish a statistically reliable model difference or context-independent coding of routines and roles.

5 Discussion

We examined how children's naturalistic egocentric experience is reflected in learned hierarchical representations. Child's-view models exhibit coarse-to-fine organization, with broad distinctions emerging earlier during training. Notably, child's-view-trained models show stronger separation at the broadest category level than conventionally trained models, consistent with the distinctive statistics

of children's visual environments. Language grounding further increases alignment with human conceptual and contextual judgments. Together, these findings suggest complementary contributions of visual and linguistic experience to category organization.

The emergence of coarse-to-fine organization connects child's-view learning with progressive differentiation accounts, in which shared statistical regularities support broad distinctions before finer ones [Rogers and McClelland, 2004, Saxe et al., 2019]. A comparable ordering can arise from self-supervised learning on naturalistic visual input, rather than explicitly constructed concept–property datasets. Building on spectral analyses of conventionally trained models [Cheon et al., 2026], we distinguish where category structure is expressed across retained ranks from when it emerges during learning.

Beyond this shared coarse-to-fine ordering, the strength of category separation differs across training environments. SAYCam-trained models show stronger separation at the broadest category level, whereas conventionally trained models more clearly differentiate finer categories. This contrast is consistent with recent findings that objects encountered in naturalistic child's-view environments exhibit stronger superordinate grouping than canonical object images [Yang et al., 2026]. Together, these results raise the possibility that the distinctive visual statistics of children's environments may particularly support broad category organization in learned representations.

Language grounding may contribute a complementary source of organization. While representations learned from child's-view vision align most strongly with human perceptual similarity, language grounding modestly increases alignment with conceptual and contextual judgments. This pattern suggests that linguistic information may enrich visually grounded representations with broader object relationships, consistent with evidence that words influence early category formation [Waxman and Markow, 1995]. Thus, visual experience and language grounding may contribute different, complementary aspects of category organization, rather than simply strengthening the same distinctions.

Several limitations constrain the developmental interpretation. The taxonomies are externally defined and need not match infants' psychological categories. Child's-view recordings capture only part of children's experience, and the models lack important sources of learning, including action, interaction, and broader social context. Differences in training data and model configurations limit causal conclusions about environmental statistics, while changes under language grounding do not isolate linguistic information from the effects of the learned projection. Future work should test whether these representations predict children's categorization and generalization under more closely matched learning conditions, including richer multimodal and interactive experience.

References

D. A. Baldwin, E. M. Markman, and R. L. Melartin. Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child development*, 64(3):711–728, 1993.

A. S. Benjamin, A.-L. Beyer, M. D. H. Kloots, J. Hwang, H. Karoui, M. Ostrow, J. Rubruck, K. J. Sandbrink, S. Grant, A. M. Saxe, et al. An introduction to connectionist theories of semantic cognition. In *Analytical Connectionism School*, pages 42–67. PMLR, 2026.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pages 9630–9640. IEEE, 2021.

J. Cheon, M. Vogelsang, L. Vogelsang, and P. Sinha. Spectral stratification of semantic abstraction in vision-language models. In *Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning*, 2026.

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

E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160055, 2017.

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

R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, 2014.

S. A. Gelman and E. M. Markman. Categories and induction in young children. *Cognition*, 23(3): 183–209, 1986.

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

M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4 (11):1173–1185, 2020.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

P. A. Huebner and J. A. Willits. Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, 9:133, 2018.

K. M. Jozwik, N. Kriegeskorte, K. R. Storrs, and M. Mur. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, 8:1726, 2017.

A. H. Kim, G. L. Quek, D. Moerel, O. Gorton, and T. A. Carlson. Disentangling objects' contextual associations from perceptual and conceptual attributes using time-resolved neural decoding. *Journal of Vision*, 26(8):2–2, 2026.

N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. doi: 10.1007/s11263-020-01316-z.

Y. Liu, Y. Zhang, D. Ghosh, L. Schmidt, and S. Yeung-Levy. Data or language supervision: What makes CLIP better than DINO? In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1868–1874, Suzhou, China, Nov. 2025. Association for Computational Linguistics.

B. L. Long, R. Z. Sparks, V. Xiang, S. Stojanov, ZiYin, G. Keene, A. W. M. Tan, S. Y. Feng, A. Nag, C. Zhuang, V. A. Marchman, D. L. Yamins, and M. Frank. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. In *8th Annual Conference on Cognitive Computational Neuroscience*, 2025.

J. Lovón-Melgarejo, J. G. Moreno, R. Besançon, O. Ferret, and L. Tamine. Probing pretrained language models with hierarchy properties. In *European Conference on Information Retrieval*, pages 126–142. Springer, 2024.

F. P. Mahner, L. Muttenthaler, U. Güçlü, and M. N. Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

J. M. Mandler. Perceptual and conceptual processes in infancy. *Journal of cognition and development*, 1(1):3–36, 2000.

J. M. Mandler, P. J. Bauer, and L. McDonough. Separating the sheep from the goats: Differentiating global categories. *Cognitive psychology*, 23(2):263–298, 1991.

J. L. McClelland and T. T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, 4(4):310–322, 2003.

C. B. Mervis, E. Rosch, et al. Categorization of natural objects. *Annual review of psychology*, 32(1): 89–115, 1981.

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

G. L. Murphy and H. H. Brownell. Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of experimental psychology: Learning, memory, and cognition*, 11(1):70, 1985.

L. Muttenthaler, J. Dippel, L. Linhardt, R. A. Vandermeulen, and S. Kornblith. Human alignment of neural network representations. In *The 11th International Conference on Learning Representations*, 2023.

L. Muttenthaler, K. Greff, F. Born, B. Spitzer, S. Kornblith, M. C. Mozer, K.-R. Müller, T. Unterthiner, and A. K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

C. O'Doherty, Á. T. Dineen, A. Truzzi, G. King, L. Zaadnoordijk, K. Harrison, E.-L. D'Arcy, J. White, C. Caldinelli, T. Holloway, A. Kravchenko, J. Diedrichsen, A. Tarrant, A. T. Byrne, A. Foran, E. J. Molloy, and R. Cusack. Infants have rich visual categories in ventrotemporal cortex at 2 months of age. *Nature Neuroscience*, 29(3):693–702, 2026. doi: 10.1038/s41593-025-02187-8.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

A. E. Orhan and B. M. Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3):271–283, 2024.

E. Orhan, V. Gupta, and B. M. Lake. Self-supervised learning through the eyes of a child. *Advances in neural information processing systems*, 33:9960–9971, 2020.

P. C. Quinn and P. D. Eimas. A reexamination of the perceptual-to-conceptual shift in mental representations. *Review of General Psychology*, 1(3):271–287, 1997.

P. C. Quinn, P. D. Eimas, and S. L. Rosenkrantz. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):463–475, 1993.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

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

T. T. Rogers, M. A. Lambon Ralph, P. Garrard, S. Bozeat, J. L. McClelland, J. R. Hodges, and K. Patterson. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, 111(1):205, 2004.

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

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

A. M. Saxe, J. L. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

L. B. Smith, S. Jayaraman, E. Clerkin, and C. Yu. The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, 22(4):325–336, 2018.

R. R. Sokal, C. D. Michener, et al. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38:1409–1438, 1958.

J. Sullivan, M. Mei, A. Perfors, E. Wojcik, and M. C. Frank. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5:20–29, 2021.

J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, 23(3):457–482, 1991.

G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021.

W. K. Vong, W. Wang, A. E. Orhan, and B. M. Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

S. R. Waxman and D. B. Markow. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, 29(3):257–302, 1995.

J. Yang, T. Sepuri, A. W. M. Tan, K. L. Aw, M. C. Frank, and B. Long. Characterizing the visual representation of objects from the child's view. *arXiv preprint arXiv:2605.14990*, 2026.

A Extended Related Work

We provide additional background to complement the related work in Section 2.

A.1 Development of Category Organization

Rosch et al. [1976] identified the basic level of categorization through shared attributes, shape similarity, and associated motor actions. Intermediate categories, such as chairs within furniture, preserved substantial similarity among their members while remaining differentiated from other categories. The preferred level of categorization also depends on experience. Tanaka and Taylor [1991] found that experts could name and categorize objects at subordinate levels as readily as at the basic level within their domain of expertise. The developmental ordering of taxonomic distinctions is less straightforward. Mandler et al. [1991] found that 18-month-olds differentiated global categories such as animals and vehicles more consistently than categories within those domains in an object-manipulation task. However, visual familiarization studies provide evidence of more specific category representations at younger ages. Quinn et al. [1993] found that 3- and 4-month-olds formed cat and dog categories that excluded birds. The dog category initially included cats, but reducing variability among the dog exemplars removed this asymmetry. Differences in tasks, exemplars, and category contrasts complicate comparisons across these findings and have informed debates over perceptual and conceptual accounts of infant categorization [Mandler, 2000, Quinn and Eimas, 1997, Murphy, 2002].

Category-based induction provides evidence about how these representations support generalization beyond observable similarity. Baldwin et al. [1993] found that infants expected perceptually similar objects to share non-obvious properties, as measured through exploratory play. Gelman and Markman [1986] tested preschool children using objects whose category membership conflicted with their perceptual resemblance. Children often generalized non-obvious properties according to the object's named category, even when it more closely resembled an object from another category. Labels also influence category formation itself: Waxman and Markow [1995] found that novel words facilitated categorization in 12- to 13-month-olds, particularly at the superordinate level. Both nouns and adjectives supported this effect, suggesting an early association between words and categories before more specific grammatical expectations develop.

A.2 Connectionist Models of Category Learning

Connectionist models explain the acquisition of conceptual knowledge through learning the statistical relationships among object properties [McClelland and Rogers, 2003, Rogers and McClelland, 2004]. In the semantic networks of Rogers and McClelland [2004], an object and a relation are used to predict a set of properties. Objects sharing many properties acquire similar internal representations, and training progressively differentiates broad domains, such as plants and animals, followed by more specific categories and individual concepts. This organization develops without an explicitly specified category structure. A related account explains aspects of semantic deterioration. Rogers et al. [2004] trained a recurrent network to associate names, verbal descriptions, and visual attributes through shared representations, then compared the effects of network damage with semantic dementia. Across naming, sorting, and drawing tasks, distinctive properties were more vulnerable than properties shared across related concepts, in both the model and patient data.

Saxe et al. [2019] provided an analytical account of progressive differentiation in deep linear networks. Their learning dynamics relate the acquisition of item–property associations to singular modes of the input–output correlation matrix. For the hierarchical datasets they studied, broad distinctions corresponded to larger singular values and were learned earlier. Learning these modes at different rates can cause predictions for some object properties to become temporarily less accurate before improving. With one-hot item inputs, these modes are derived from the object–property matrix in the training data, whereas our principal components are computed from the models' learned representations.

Huebner and Willits [2018] investigated whether categorical and hierarchical structure could emerge from predicting word sequences in naturalistic child-directed speech. They trained recurrent networks on speech directed to children aged 0–3 years and examined the learned word representations using PCA and hierarchical clustering. Individual principal components reflected grammatical and activity-

related distinctions, while clustering revealed organization within and across semantic categories. We examine a related question in visual representations learned from child's-view video. Rather than interpreting individual components, we use cumulative PCA reconstructions to measure how category separation at different taxonomic depths changes as more components are retained. We also track this separation across training checkpoints, allowing us to examine both where broad and fine distinctions are expressed in the variance spectrum and when they emerge during learning.

A.3 Category Structure in Learned Representations

Representational similarity analysis compares models, neural measurements, and behavior through their pairwise dissimilarity structure [Kriegeskorte et al., 2008]. Using human similarity judgments for object images, Jozwik et al. [2017] found that the final layers of AlexNet and VGG-16 outperformed most models constructed from annotated visual features. Models based on category annotations nevertheless outperformed every network layer, indicating that object-recognition performance alone did not account for human similarity judgments. Hebart et al. [2020] examined this structure at a larger scale, recovering 49 interpretable dimensions from judgments for 1,854 objects. The dimensions included both perceptual and conceptual properties and predicted additional categorization and typicality judgments.

Human judgments can also be used to modify model representations. Muttenthaler et al. [2025] trained a teacher model to predict human similarity judgments and transferred the resulting structure to pretrained vision models. The adapted models more closely matched human behavior across levels of abstraction. This approach directly incorporates behavioral supervision, while our comparisons evaluate the alignment acquired through visual learning and language grounding.

The form of supervision also affects which object properties are emphasized in visual representations. DINO learns through self-distillation without category labels [Caron et al., 2021], whereas CLIP learns to associate images with accompanying text [Radford et al., 2021]. Their original implementations differ in training data as well as supervision. To distinguish these factors, Liu et al. [2025] trained the two models with matched architectures, images, and training configurations. Despite similar ImageNet accuracy, CLIP representations were more sensitive to semantic attributes, while DINO representations were more sensitive to visual attributes such as color and style.

Cheon et al. [2026] examined taxonomic structure in conventionally trained vision–language models by retaining or removing principal components. Broad distinctions were accessible in leading subspaces, while finer distinctions depended on additional components. Manipulating these subspaces also changed retrieval at the corresponding taxonomic levels. These findings raise the question of whether similar organization emerges from children's everyday visual experience.

Learning from child's-view recordings has yielded transferable visual representations [Orhan et al., 2020, Orhan and Lake, 2024] and grounded word–referent knowledge [Vong et al., 2024]. We examine the taxonomic organization acquired from this experience and its development during training. We also assess alignment with human perceptual, conceptual, and contextual similarity judgments [Kim et al., 2026], including how this alignment changes after language grounding.

B Experiment Details

B.1 Datasets and hierarchy construction

The taxonomy analyses use 32 THINGS concepts arranged in the hierarchy of Cheon et al. [2026]. We select the first 12 image files in lexicographic order from each concept directory, yielding 384 images, and use the same images across models, ranks, and training checkpoints. The four depths contain 2, 4, 14, and 32 groups (Table 1). This hierarchy is hand-curated and strictly nested.

The human-similarity analysis in Section 4.3 uses the 190 THINGS concepts studied by Kim et al. [2026]. We select the first ten images per concept in lexicographic order from the THINGS release [Hebart et al., 2019], yielding 1,900 images. The released behavioral data identify concepts but do not include an exact stimulus-image manifest, so the model and human evaluations are matched at the concept level; an image-level match cannot be verified. Construction of the model dissimilarity matrices is described in Appendix B.5.

Table 1: The complete THINGS hierarchy used for the taxonomy analyses. Each row lists one path through Depths 1–3 and its constituent concepts at Depth 4.

Depth 1	Depth 2	Depth 3	Depth 4
Animate	Animal	Bird	penguin, pigeon, seagull
Animate	Animal	Equid	donkey, horse, zebra
Animate	Animal	Pet	cat, dog
Animate	Animal	Predator	lion, tiger, wolf
Animate	Human	Adult	man, woman
Animate	Human	Child	baby, boy, girl
Inanimate	Tool	Bristle	broom, brush
Inanimate	Tool	Garden tool	axe, scissors
Inanimate	Tool	Hardware	drill, hammer
Inanimate	Tool	Kitchen tool	knife, spoon
Inanimate	Vehicle	Air vehicle	airplane, helicopter
Inanimate	Vehicle	Rail vehicle	subway, train
Inanimate	Vehicle	Road vehicle	car, truck
Inanimate	Vehicle	Water vehicle	boat, ship

Complete human-similarity concept inventory

The Section 4.3 probe uses this 190-concept inventory, distinct from the 32-concept taxonomy in Table 1. Names retain the released identifiers, including numeric suffixes.
airplane, ant, apple, arm, arrow, avocado, axe, ball, banana, baseball_bat, bassoon, beachball, bed, bedpan, bee, beetle, bell, bench, berry, bike, binder, blazer, boat, bongo, bookshelf, bottle, bowl, bread, broccoli, burner, bus, bush, butterfly, cabbage, cabinet, candlestick, cane, cannonball, car, cardigan, carrot, cat, catapult, cello, cellphone, chainsaw, cheeseburger, chicken2, chisel, closet, coaster, cockroach, coconut, coffeemaker, computer_screen, crow, crumb, cup, cupcake, cutting_board, desk, doily, doll, donut, drill, duck, eagle, ear, enchilada, engine, exhaust_pipe, eye_patch, fan, figurine, finger, flamethrower, flipper, flower, flute, food_processor, foot, fox, french_fries, frisbee, garbage_truck, gasket, gazelle, gearshift, glove, goalpost, goat, goose, grapevine, grass, grater, gun, hand, headlight, headrest, hedge, honey, horse, hula_hoop, inhaler, jar, kettle, keyboard, kite, knife, lab_coat, ladybug, leek, leg, lettuce, machete, mailbox, mallet, mandolin, marble, melon, microwave, milkshake, mirror, missile, mosquito, mouse2, nose, notebook, oilcan, oven, pantyhose, paperclip, parfait, peach, pear, pen, pesto, piano, pie, pigeon, pillow, plant, plaster_cast, pocketknife, pool_table, popsicle, printer, punch2, rabbit, radish, rearview_mirror, recliner, refrigerator, roadsweeper, rollerblade, rooster, saddle, scarf, scissors, scoop, seagull, seaweed, shark, shoe, shopping_basket, shortbread, shovel, sifter, skateboard, sling, squirrel, squirt_gun, stereo, stethoscope, stretcher, subway, teddy_bear, television, toaster, tomato, tongs, tongue, tree, trumpet, tupperware, vase, wasp, wheel, whipped_cream, wire_cutters.

B.2 Models and representation extraction

We use released SAYCam models from Orhan and Lake [2024][2] and the CVCL checkpoint from Vong et al. [2024].[3] All encoders are frozen during evaluation. Transformer backbone features are post-normalization class tokens; convolutional features are spatially pooled before the classifier (Table 2). For CVCL, we extract the backbone features and learned linear projection from the same checkpoint. We use ImageNet-1k ViT-B/16 and ResNet-50 models for the taxonomy comparison. The human-similarity analysis uses an ImageNet-1k ViT-B/14 to match the SAYCam architecture.

The matched DINO and CLIP [Radford et al., 2021] checkpoints are from Liu et al. [2025]. Both models were trained for 20 epochs on the same 10M-image subset of DataComp [Liu et al., 2025].

B.3 Spectral analysis details

We L2-normalize each image representation before fitting centered, unwhitened PCA. The PCA basis is shared across hierarchy depths and refit at each developmental checkpoint. We restore the mean in each rank-truncated reconstruction and normalize it again before computing cosine distances, as

[2] https://github.com/eminorhan/silicon-menagerie
[3] https://github.com/wkvong/multimodal-baby

Table 2: Models and evaluation features. SAYCam rows use pooled recordings from all three children; CVCL uses recordings from child S.

Model / training data	Architecture	Feature	Dimensions
DINO / SAYCam	ViT-B/14	Teacher class token	768
DINO / SAYCam	ResNeXt-50	Pooled visual features	2,048
DINO / ImageNet-1k	ViT-B/16	Class token	768
DINO / ImageNet-1k	ResNet-50	Pooled visual features	2,048
DINO / ImageNet-1k	ViT-B/14	Teacher class token	768
CVCL / child S	ViT-B/14	Class token / projection	768 / 512
DINOv2 / SAYCam	ViT-B/14	Teacher class token	768
DINO / DataComp	ViT-B/16	Class token	768
CLIP / DataComp	ViT-B/16	Projected image features	512
Random, seed 0	ViT-B/14	Class token	768
Raw pixels	—	Downsampled color pixels	3,072

defined in Section 3.1. Mean silhouette weights images equally, excludes self-distances, and assigns zero to singleton groups.

The spectra in Section 4.1 evaluate every rank from 1 to $\min(d, N-1) = 383$. The developmental analysis uses

$$\mathcal{K}_{\mathrm{T}} = \{1, 2, 3, 4, 5, 6, 8, 10, 13, 16, 20, 25, 30, 40, 50, 64, \\ 80, 100, 128, 160, 200, 256, 320, 383\}.$$

Ties are resolved by selecting the smallest maximizing rank. At rank 383, all available PCA directions are retained. Table 3 reports the main endpoint peaks; these are descriptive point estimates, not permutation-test results.

Table 3: Peak mean silhouette and retained rank (in parentheses) on the 384-image THINGS pool. Entries are point estimates from the exhaustive rank search.

Training / architecture	Depth 1	Depth 2	Depth 3	Depth 4
SAY DINO ViT-B/14	.636 (1)	.449 (3)	.072 (27)	.075 (46)
SAY DINO ResNeXt-50	.675 (1)	.349 (4)	.033 (20)	.027 (47)
ImageNet-1k DINO ViT-B/16	.588 (2)	.667 (2)	.468 (11)	.539 (25)
ImageNet-1k DINO ResNet-50	.305 (5)	.545 (5)	.353 (15)	.454 (30)

B.4 Details of Section 4.2

Training and spectral evaluation. We train a ViT-B/14 using DINOv2 on SAYCam frames sampled at 5 frames per second from children S, A, and Y. We use AdamW with a global batch size of 1,024 images. The learning rate warms up from zero to .004 over 6,000 updates, then follows cosine decay to 10^{-6}. The optimizer and teacher schedules span 250,000 updates, although this run stops at 10,000 updates. We evaluate teacher features on the 384 THINGS images every 200 updates from 200 to 10,000.

Hierarchical clustering. At each checkpoint, we average L2-normalized full representations within each concept, normalize the resulting 32 centroids, and apply average-linkage clustering to their cosine distances. We cut each tree into 2, 4, and 14 groups and compare the assignments with the corresponding reference partitions at Depths 1–3 using adjusted Rand index [Hubert and Arabie, 1985].

For a given depth, let n_{ab} denote the number of concepts assigned to learned group a and reference group b, with marginal counts $n_{a\cdot} = \sum_b n_{ab}$ and $n_{\cdot b} = \sum_a n_{ab}$. For $n = 32$ concepts,

$$\mathrm{ARI} = \frac{\sum_{a,b} \binom{n_{ab}}{2} - \frac{\left(\sum_a \binom{n_{a\cdot}}{2}\right)\left(\sum_b \binom{n_{\cdot b}}{2}\right)}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_a \binom{n_{a\cdot}}{2} + \sum_b \binom{n_{\cdot b}}{2}\right] - \frac{\left(\sum_a \binom{n_{a\cdot}}{2}\right)\left(\sum_b \binom{n_{\cdot b}}{2}\right)}{\binom{n}{2}}}. \tag{1}$$

The dendrograms in Section 4.2 show the results from checkpoints at 2,500, 5,000, 7,500, and 10,000 updates.

B.5 Human similarity and grounding evaluation

We use the released image-judgment matrices of Kim et al. [2026]. Participants selected the most similar pair in each object triplet based on appearance, conceptual properties, or real-world context. Human dissimilarity is one minus the proportion of joint presentations in which a pair was selected. Unobserved pairs remain missing.

For each concept, we average ten L2-normalized image representations and normalize the mean. Cosine distances between concept means form the model dissimilarity matrix (Figure 7). These distances, $1 - \cos(u, v)$, range from 0 to 2, whereas human dissimilarities range from 0 to 1. We compute Spearman correlations over the 14,964 concept pairs observed in all six human matrices: image and word judgments for each of the three criteria. Spearman correlation depends on the ordering of dissimilarities, so positive rescaling does not affect it; heatmap limits affect only the display, not the values used for correlation. The primary comparisons use image judgments.

We estimate 95% percentile intervals for model differences using 2,000 paired bootstrap draws. Each draw resamples 190 concepts with replacement and ten images within each sampled concept, also with replacement. Models share the same draws, and we recompute their concept means, dissimilarities, and correlations. Repeated concepts retain their multiplicity, while comparisons of a concept with itself and pairs missing from the human matrices are excluded. The intervals reflect variation over concepts and image exemplars, conditional on the released human judgments and trained models.

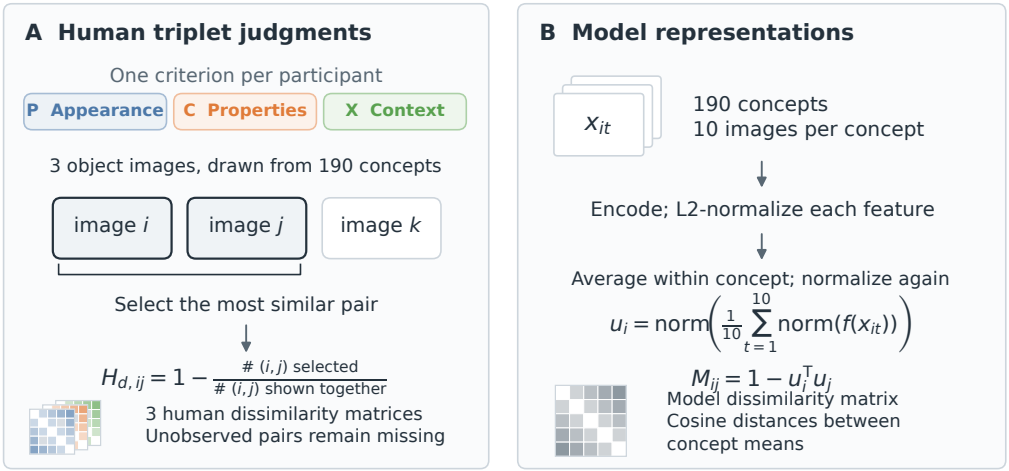


Figure 7: Schematic of the human and model similarity analyses. (A) Triplet judgments from Kim et al. [2026] yield separate perceptual, conceptual, and contextual dissimilarity matrices. (B) Model dissimilarities are cosine distances between normalized concept means, with $\mathrm{norm}(v) = v/\|v\|_2$. Comparisons are matched at the concept level. Gray entries indicate missing human judgments.

B.6 Representational dissimilarity matrices for Section 4.1

Figure 8 shows representational dissimilarity matrices for the models and raw-pixel baseline in Section 4.1, evaluated on the THINGS taxonomy probe at each hierarchy depth's peak rank.

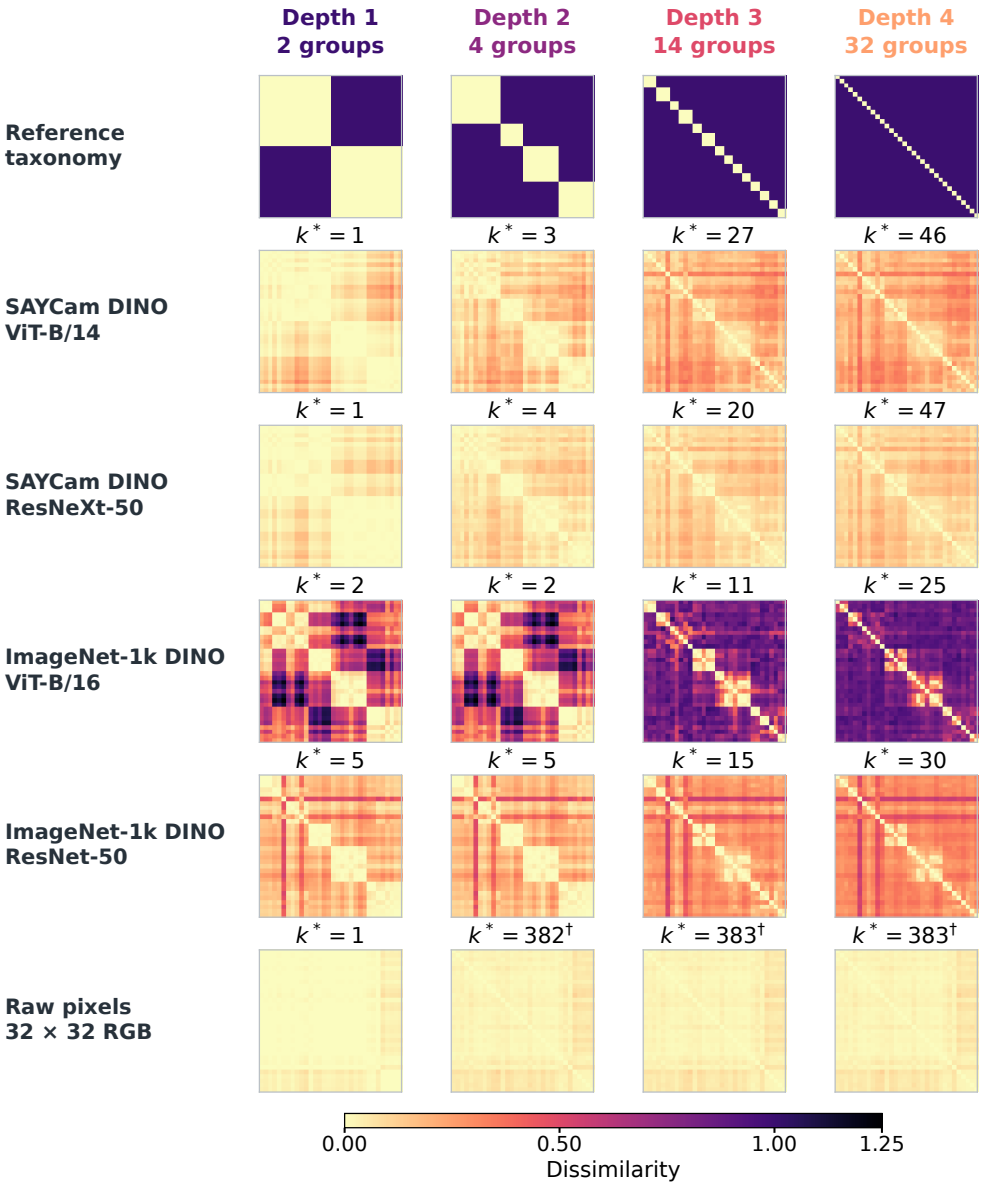


Figure 8: Representational dissimilarity matrices at the THINGS silhouette peaks. Columns correspond to hierarchy depths containing 2, 4, 14, and 32 groups. Each matrix compares the same 32 concepts in a common taxonomy-based order. Reference entries are zero for concepts sharing a category at that depth and one otherwise. For each model, we reconstruct the 384 image representations at the annotated peak rank, restore the PCA mean, normalize each reconstruction, and average the 12 images within each concept. Model entries are cosine distances between normalized concept means. Peak ranks are selected from image-level silhouettes, not from these concept-level matrices.

B.7 Representational dissimilarity matrices for Section 4.3

Figures 9 and 10 compare the Section 4.3 models with human references on the 190-concept THINGS human-similarity probe.

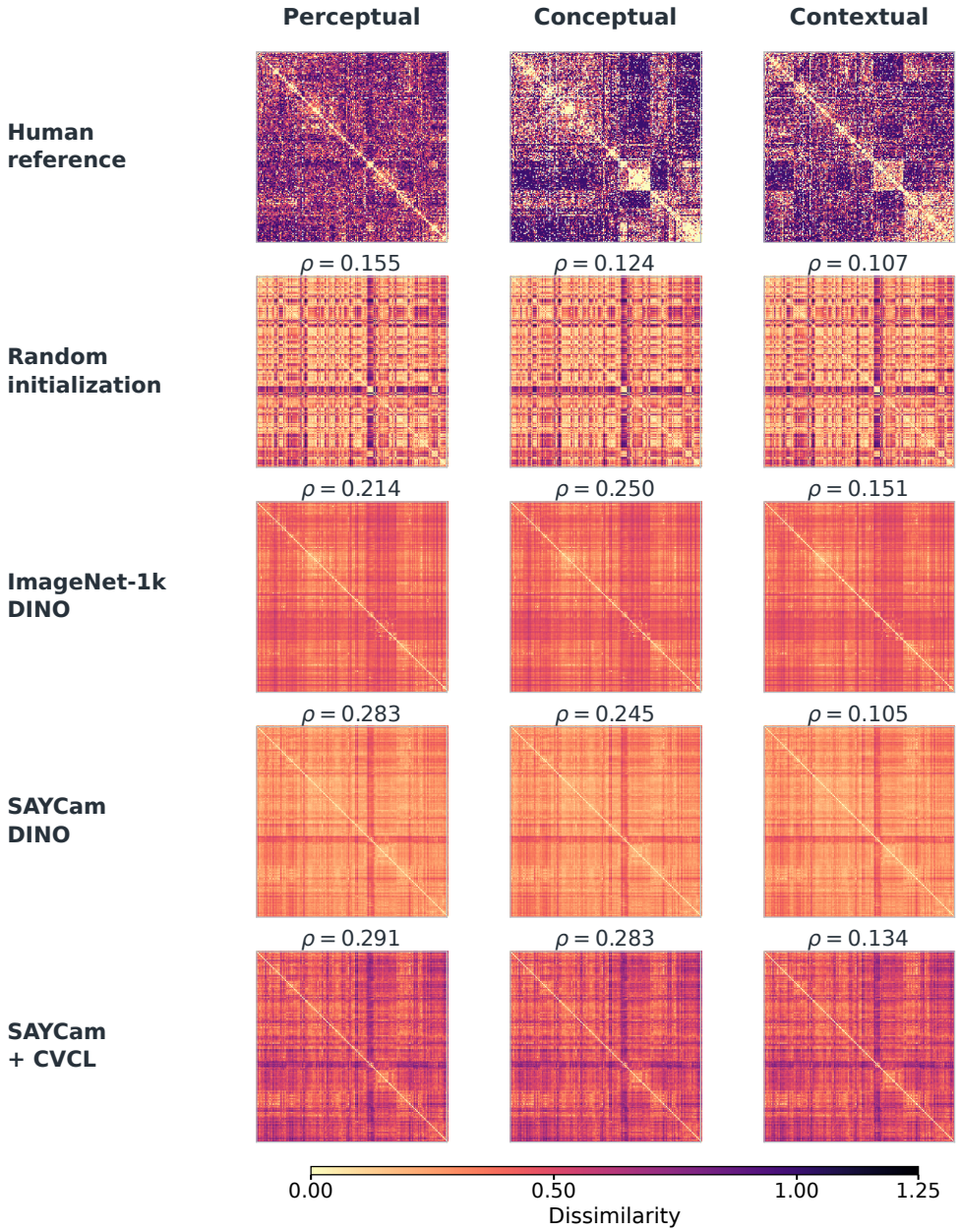


Figure 9: Human and model RDMs for Section 4.3. The top row shows the same human matrices as Figure 4; all matrices compare 190 concepts in its shared order. Here, all panels use a 0–1.25 color scale to accommodate model distances above 1; Figure 4 uses 0–1 for human judgments alone. Each model's full RDM repeats across columns, with Spearman ρ against the column's human reference. Model entries are cosine distances between normalized means of ten L2-normalized image representations per concept. Models use ViT-B/14.

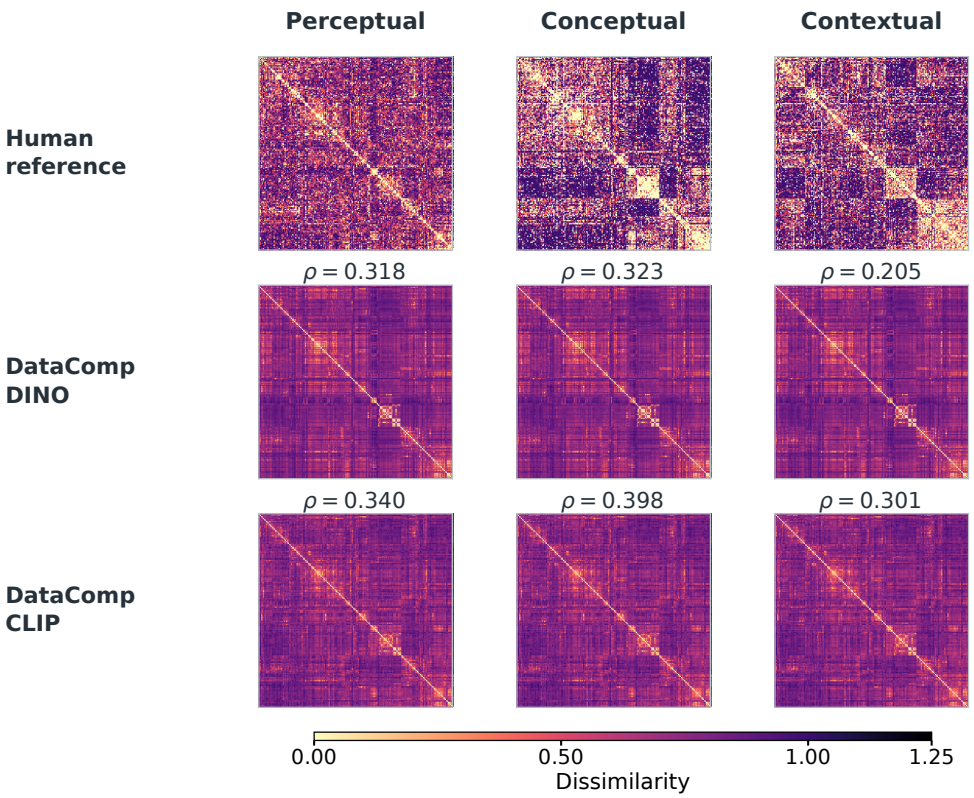


Figure 10: Human and DataComp model dissimilarities for Section 4.3. The top row repeats the human image-judgment references; the following rows show DINO and the CLIP representation from ViT-B/16 models trained on the same DataComp images. All panels use the 0–1.25 color scale of Figure 9. Each model's full RDM is identical across its three cells; the annotated Spearman correlations compare it with different human references.

B.8 Perceptual, conceptual, and contextual alignment across rank

To supplement Section 4.3, we examine how alignment with human perceptual, conceptual, and contextual similarity changes across the representation spectrum.

Setup. We analyze the SAYCam DINO ViT-B/14 from Section 4.3, using rank-truncated representations of the same 1,900 images from 190 concepts. At each retained rank, we compute cosine distances between concept-mean representations and correlate the resulting dissimilarity matrix with each human reference using Spearman correlation, following Appendix B.5. We evaluate 19 ranks from 1 to 768.

Results. Figure 11 shows the results. Although the full representation aligns more strongly with perceptual judgments (Section 4.3), conceptual alignment is highest at $k = 1$ ($\rho = .303$), while perceptual alignment peaks at $k = 24$ ($\rho = .301$). This pattern suggests that fewer leading components suffice to capture the model's strongest alignment with conceptual judgments, whereas perceptual alignment benefits from additional dimensions.

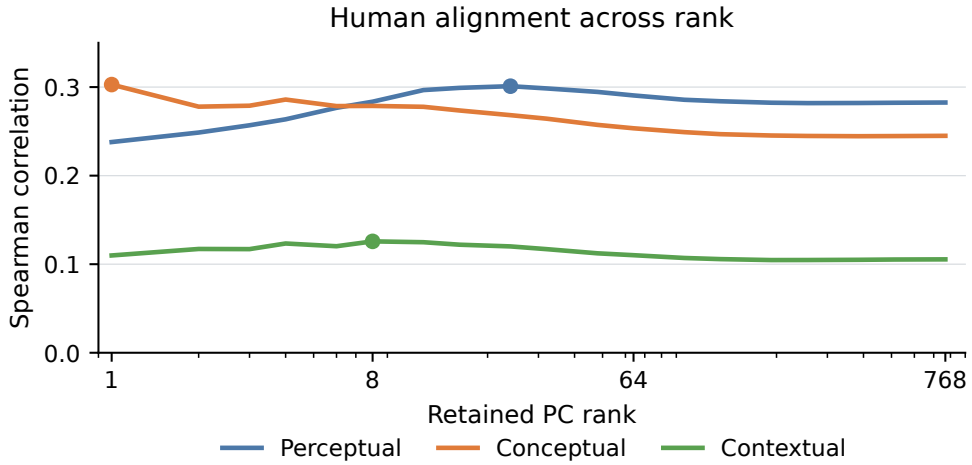


Figure 11: Perceptual, conceptual, and contextual alignment across retained rank for SAYCam DINO ViT-B/14 (child S). Curves show Spearman correlations between model and human dissimilarity matrices computed from rank-truncated representations. Circles mark the highest correlation for each criterion among the 19 evaluated ranks.

C Additional Experiments

The main experiments examine category structure in 32 THINGS concepts. Additionally, we extend these analyses to ImageNet-1k and iNaturalist [Van Horn et al., 2021], which provide larger or more finely differentiated category hierarchies. We also examine changes in taxonomic separation associated with language grounding.

C.1 Category organization on ImageNet and iNaturalist

We test whether the coarse-to-fine organization observed on THINGS in Section 4.1 extends to larger object and biological hierarchies.

Data. For ImageNet-1k, we use the hierarchy of Muttenthaler et al. [2025], retaining 385 classes grouped into 96 basic and seven superordinate categories. An additional animate–inanimate partition gives four depths containing 2, 7, 96, and 385 groups. We sample 30 validation images per class, giving 11,550 images shared across depths [Deng et al., 2009].

For iNaturalist [Van Horn et al., 2021], we select 64 species and use all ten validation images per species, giving 640 images. We construct a balanced binary hierarchy within Animalia–Chordata and Plantae–Tracheophyta, retaining two child taxa at each level. We select taxa that support binary branching down to species, preferring those with more descendant species and breaking ties by name. The six depths are kingdom, class, order, family, genus, and species, with 2, 4, 8, 16, 32, and 64 groups. Phylum coincides with kingdom in this subset and is omitted. Taxon labels retain their ancestral paths to distinguish repeated names.

The evaluated hierarchies are listed below; the THINGS hierarchy appears in Appendix B.1.

ImageNet-1k: 385 classes.

Depths: 1 = animate/inanimate; 2 = superordinate; 3 = basic; 4 = subordinate.

Headings give Depths 1–2; bold terms give Depth 3, followed by all Depth 4 classes. Synset IDs distinguish the two classes named *maillot*.

Animate → animals
antelope: gazelle, hartebeest, impala. **bear**: american black bear, brown bear, ice bear, sloth bear. **bird**: african grey, albatross, american coot, american egret, bald eagle, bee eater, bittern, black grouse, black stork, black swan, brambling, bulbul, bustard, chickadee, cock, coucal, crane, dowitcher, drake, european gallinule, flamingo, goldfinch, goose, great grey owl, hen, hornbill, house finch, hummingbird, indigo bunting, jacamar, jay, junco, king penguin, kite, limpkin, little blue heron, lorikeet, macaw, magpie, ostrich, oystercatcher, partridge, peacock, pelican, prairie chicken, ptarmigan, quail, red-backed sandpiper, red-breasted merganser, redshank, robin, ruddy turnstone, ruffed grouse, spoonbill, sulphur-crested cockatoo, toucan, vulture, water ouzel, white stork. **butterfly**: admiral, cabbage butterfly, lycaenid, monarch, ringlet, sulphur butterfly. **camel**: arabian camel. **cattle**: ox. **dog**: affenpinscher, afghan hound, airedale, american staffordshire terrier, appenzeller, australian terrier, basenji, basset, beagle, bedlington terrier, bernese mountain dog, black-and-tan coonhound, blenheim spaniel, bloodhound, bluetick, border collie, border terrier, borzoi, boston bull, bouvier des flandres, boxer, brabancon griffon, briard, brittany spaniel, bull mastiff, cairn, cardigan, chesapeake bay retriever, chihuahua, chow, clumber, cocker spaniel, collie, curly-coated retriever, dalmatian, dandie dinmont, doberman, english foxhound, english setter, english springer, entlebucher, eskimo dog, flat-coated retriever, french bulldog, german shepherd, german short-haired pointer, giant schnauzer, golden retriever, gordon setter, great dane, great pyrenees, greater swiss mountain dog, groenendael, ibizan hound, irish setter, irish terrier, irish water spaniel, irish wolfhound, italian greyhound, japanese spaniel, keeshond, kelpie, kerry blue terrier, komondor, kuvasz, labrador retriever, lakeland terrier, leonberg, lhasa, malamute, malinois, maltese dog, mexican hairless, miniature pinscher, miniature poodle, miniature schnauzer, newfoundland, norfolk terrier, norwegian elkhound, norwich terrier, old english sheepdog, otterhound, papillon, pekinese, pembroke, pomeranian, pug, redbone, rhodesian ridgeback, rottweiler, saint bernard, saluki, samoyed, schipperke, scotch terrier, scottish deerhound, sealyham terrier, shetland sheepdog, shih-tzu, siberian husky, silky terrier, soft-coated wheaten terrier, staffordshire bullterrier, standard poodle, standard schnauzer, sussex spaniel, tibetan mastiff, tibetan terrier, toy poodle, toy terrier, vizsla, walker hound, weimaraner, welsh springer spaniel, west highland white terrier, whippet, wire-haired fox terrier, yorkshire terrier. **domestic cat**: egyptian cat, persian cat, siamese cat, tabby, tiger cat. **dragonfly**: dragonfly. **elephant**: african elephant, indian elephant. **fox**: arctic fox, grey fox, kit fox, red fox. **frog**: bullfrog, tailed frog, tree frog. **giant panda**: giant panda. **hippopotamus**: hippopotamus. **horse**: sorrel. **koala bear**: koala. **lizard**: african chameleon, agama, alligator lizard, american chameleon, banded gecko, common iguana, frilled lizard, gila monster, green lizard, komodo dragon, whiptail. **lobster**: american lobster, spiny lobster. **monkey**: baboon, capuchin, colobus, guenon, howler monkey, langur, macaque, marmoset, patas, proboscis monkey, spider monkey, squirrel monkey, titi. **person**: ballplayer, groom, scuba diver. **rabbit**: angora, wood rabbit. **ray**: electric ray, stingray. **red panda**: lesser panda. **scorpion**: scorpion. **seal**: sea lion. **sheep**: ram. **snake**: boa constrictor, diamondback, garter snake, green mamba, green snake, hognose snake, horned viper, indian cobra, king snake, night snake, ringneck snake, rock python, sea snake, sidewinder, thunder snake, vine snake, water snake. **squirrel**: fox squirrel. **starfish**: starfish. **swine**: hog, warthog, wild boar. **turtle**: box turtle, leatherback turtle, loggerhead, mud turtle, terrapin. **whale**: grey whale, killer whale.

Inanimate → clothes
bathing cap: bathing cap. **diaper**: diaper. **hat with a wide brim**: sombrero. **helmet**: crash helmet, football helmet, pickelhaube. **maillot**: maillot [`n03710637`], maillot [`n03710721`]. **miniskirt**: miniskirt. **sunglasses**: sunglasses. **swimming trunks**: swimming trunks. **tie**: bolo tie, bow tie, windsor tie.

Inanimate → food
apple: granny smith. **artichoke**: artichoke. **banana**: banana. **bell pepper**: bell pepper. **cream**: ice cream. **fig**: fig. **guacamole**: guacamole. **hamburger**: cheeseburger. **lemon**: lemon. **mushroom**: mushroom. **pineapple**: pineapple. **pomegranate**: pomegranate. **popsicle**: ice lolly. **salt or pepper shaker**: saltshaker. **strawberry**: strawberry.

Inanimate → furniture
baby bed: bassinet, cradle, crib. **bench**: park bench. **bookshelf**: bookcase. **chair**: barber chair, folding chair, rocking chair, throne. **sofa**: studio couch. **table**: desk, dining table, pool table.

Inanimate → musical instruments
accordion: accordion. **banjo**: banjo. **cello**: cello. **french horn**: french horn. **guitar**: acoustic guitar, electric guitar. **harmonica**: harmonica. **maraca**: maraca. **piano**: grand piano, upright. **saxophone**: sax.

Inanimate → sports equipment
balance beam: balance beam. **baseball**: baseball. **basketball**: basketball. **croquet ball**: croquet ball. **dumbbell**: dumbbell. **golf ball**: golf ball. **horizontal bar**: horizontal bar. **ping-pong ball**: ping-pong ball. **puck**: puck. **punching bag**: punching bag. **rugby ball**: rugby ball. **soccer ball**: soccer ball. **tennis ball**: tennis ball. **volleyball**: volleyball.

Inanimate → vehicles
airplane: airliner. **bicycle**: bicycle-built-for-two, mountain bike. **bus**: minibus, school bus, trolleybus. **car**: ambulance, beach wagon, cab, convertible, jeep, limousine, minivan, model t, racer, sports car. **cart**: barrow, horse cart, jinrikisha, oxcart, shopping cart. **golfcart**: golfcart. **motorcycle**: moped. **snowplow**: snowplow. **train**: bullet train. **unicycle**: unicycle. **watercraft**: aircraft carrier, canoe, catamaran, container ship, fireboat, gondola, lifeboat, liner, pirate, schooner, speedboat, submarine, trimaran, wreck, yawl.

iNaturalist: 64 species.

Depths: 1 = kingdom; 2 = class; 3 = order; 4 = family; 5 = genus; 6 = species.

Headings give kingdom → class → order. Bold terms are families; each genus is followed by its two species epithets in parentheses. A full species name combines its genus and epithet.

Animalia → Aves → Charadriiformes
Laridae: *Chroicocephalus* (*bulleri*, *cirrocephalus*); *Larus* (*argentatus*, *cachinnans*). **Scolopacidae**: *Calidris* (*acuminata*, *alba*); *Tringa* (*brevipes*, *erythropus*).

Animalia → Aves → Passeriformes
Parulidae: *Geothlypis* (*formosa*, *philadelphia*); *Setophaga* (*americana*, *caerulescens*). **Tyrannidae**: *Empidonax* (*alnorum*, *difficilis*); *Tyrannus* (*couchii*, *crassirostris*).

Animalia → Reptilia → Squamata
Colubridae: *Lampropeltis* (*californiae*, *calligaster*); *Thamnophis* (*atratus*, *cyrtopsis*). **Phrynosomatidae**: *Phrynosoma* (*blainvillii*, *cornutum*); *Sceloporus* (*clarkii*, *consobrinus*).

Animalia → Reptilia → Testudines
Emydidae: *Graptemys* (*geographica*, *ouachitensis*); *Pseudemys* (*concinna*, *nelsoni*). **Testudinidae**: *Gopherus* (*agassizii*, *berlandieri*); *Testudo* (*graeca*, *hermanni*).

Plantae → Liliopsida → Asparagales
Asparagaceae: *Agave* (*americana*, *deserti*); *Yucca* (*arkansana*, *baccata*). **Orchidaceae**: *Ophrys* (*apifera*, *fuciflora*); *Platanthera* (*aquilonis*, *bifolia*).

Plantae → Liliopsida → Poales
Cyperaceae: *Carex* (*albursina*, *aurea*); *Cyperus* (*brevifolius*, *echinatus*). **Poaceae**: *Bouteloua* (*curtipendula*, *dactyloides*); *Bromus* (*catharticus*, *diandrus*).

Plantae → Magnoliopsida → Asterales
Asteraceae: *Cirsium* (*acaule*, *altissimum*); *Solidago* (*altissima*, *bicolor*). **Campanulaceae**: *Campanula* (*aparinoides*, *barbata*); *Lobelia* (*anceps*, *angulata*).

Plantae → Magnoliopsida → Lamiales
Lamiaceae: *Salvia* (*apiana*, *aurea*); *Stachys* (*bullata*, *byzantina*). **Plantaginaceae**: *Penstemon* (*barbatus*, *centranthifolius*); *Veronica* (*americana*, *anagallis-aquatica*).

Setup. We use the same five representations and spectral analysis as in Section 4.1.

Results. Both SAYCam models show their strongest separation of animate versus inanimate categories on ImageNet and animals versus plants on iNaturalist at $k = 1$ (Figure 12). However, finer categories are less clearly separated, with non-positive peak silhouettes at several deeper levels. Across both benchmarks, SAYCam-trained models show separation profiles more concentrated at broad category levels, whereas ImageNet-trained models exhibit relatively stronger fine-grained separation, consistent with the contrasting profiles observed on THINGS. Nevertheless, even ImageNet-trained models show limited species-level separation on iNaturalist.

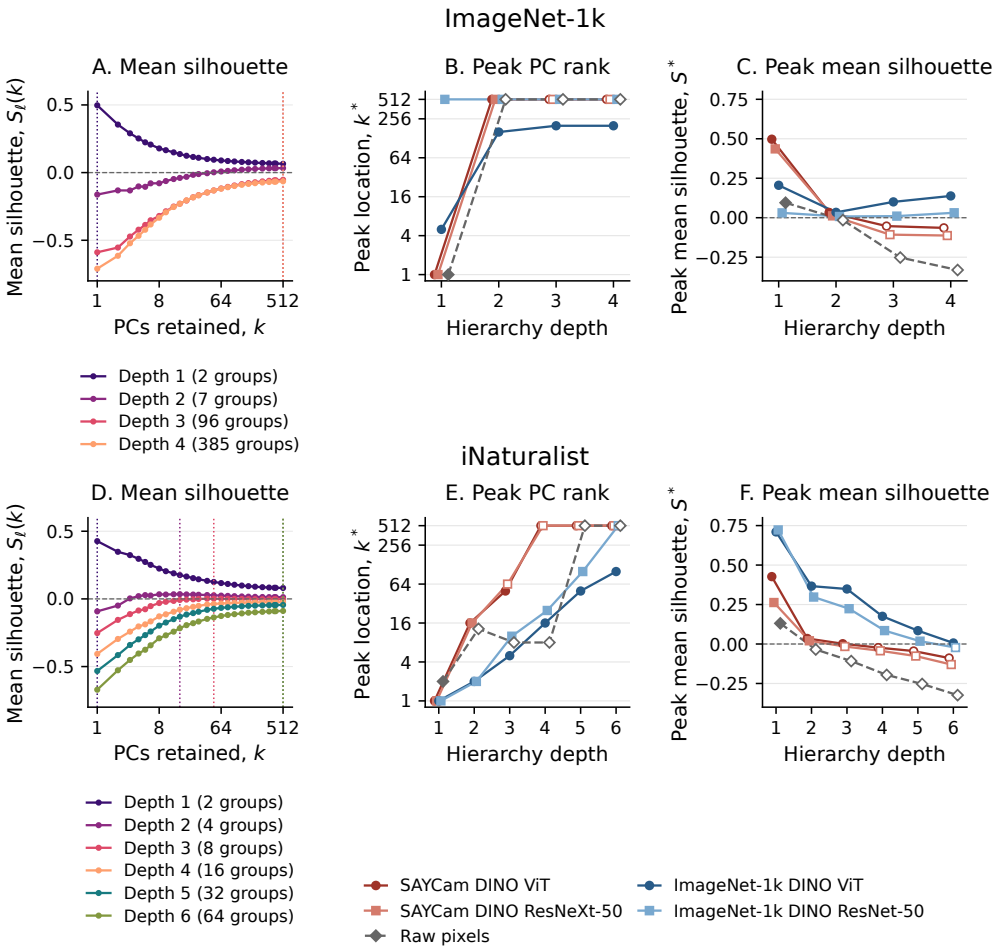


Figure 12: Category organization on ImageNet and iNaturalist. **A, D:** Silhouette curves for SAYCam DINO ViT-B/14; dotted lines mark peak ranks. **B, E:** Peak rank across models. **C, F:** Peak silhouette. Hollow markers indicate non-positive peaks. All models use a search ceiling of 512; maxima there are boundary peaks.

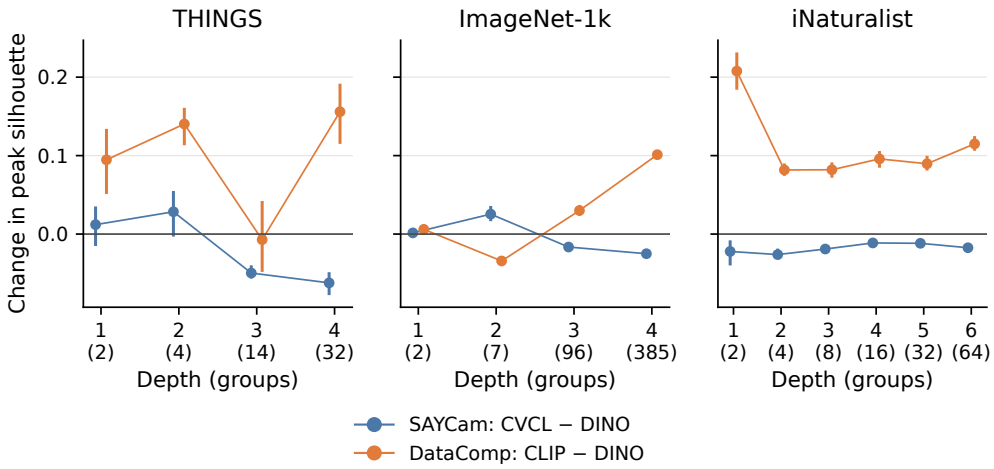


Figure 13: Changes in peak silhouette for CVCL versus SAYCam DINO and DataComp CLIP versus DINO. Positive values indicate stronger separation in the language-grounded model. Counts in parentheses indicate the number of categories at each depth.

C.2 Language grounding and taxonomic separation

Section 4.3 finds modest gains in conceptual and contextual alignment after language grounding. Here we examine whether these gains accompany stronger taxonomic separation.

Setup. We evaluate the same model pairs on THINGS, ImageNet, and iNaturalist: SAYCam DINO and CVCL (ViT-B/14, child S), and DataComp DINO and CLIP (ViT-B/16). The SAYCam comparison uses CVCL's frozen DINO backbone, while the DataComp models were trained on the same web images. CVCL and CLIP use their final projected image representations. At each depth, we subtract DINO's peak silhouette from that of the corresponding language-grounded model. We select peaks separately on the same rank grid, with search ceilings of 383 on THINGS and 512 on ImageNet and iNaturalist. These ranges are shared by the 768-dimensional DINO and 512-dimensional projected representations.

Results. CVCL slightly increases peak silhouette at the two broadest THINGS depths ($+.012$ and $+.028$), but decreases it at the finer depths ($-.050$ and $-.062$; Figure 13). The ImageNet evaluation shows the same pattern, while iNaturalist shows decreases at all six depths. DataComp CLIP, by comparison, improves over DINO at the finest depth on all three datasets, by $.156$, $.101$, and $.115$, respectively, although its ImageNet and iNaturalist peaks remain non-positive. Thus, the similar gains in human alignment reported in Section 4.3 accompany different changes in taxonomic separation. One possible explanation is that language grounding can strengthen fine distinctions only when the visual backbone already provides sufficiently differentiated features, which may be limited in the child's-view model.