Characterizing the visual representation of objects from the child's view

Jane Yang[1], Tarun Sepuri[1], Alvin W.M. Tan[2],

Khai Loong Aw[3], Michael C. Frank[2], Bria Long[1]

[1]Department of Psychology, University of California San Diego

[2]Department of Psychology, Stanford University

[3]Department of Computer Science, Stanford University

Abstract

Children acquire object category representations from their everyday experiences in the first few years of life. What do the inputs to this learning process look like? We analyzed first-person videos of young children's visual experience at home from the BabyView dataset ($N = 31$ participants, 868 hours, ages 5–36 months), using a supervised object detection model to extract common object categories from more than 3 million frames. We found that children's object category exposure was highly skewed: a few categories (e.g., cups, chairs) dominated children's visual experiences while most categories appeared rarely, replicating previous findings from a more restricted set of contexts. Category exemplars were highly variable: children encountered objects from unusual angles, in highly cluttered scenes, and partially occluded views; many categories (especially animals) were most frequently viewed as depictions. Surprisingly, despite this variability, detected categories (e.g., giraffes, apples) showed stronger groupings within superordinate categories (e.g., animals, food) relative to groupings derived from canonical photographs of these categories. We found this same pattern when using high-dimensional embeddings from both self-supervised visual and multimodal models; this effect was also recapitulated in densely sampled data from individual children. Understanding the robustness and efficiency of visual category learning will require the development of models that can exploit strong superordinate structure and learn from non-canonical, sparse, and variable exemplars.

Characterizing the visual representation of objects from the child's view

Introduction

To form a visual category – for example, an APPLE – young children need to extract the relevant visual features from diverse exemplars that may vary in shape, size, and even representational format (e.g., apple sauce, a line drawing of an apple, or a green apple on a table). Yet young children show evidence of achieving this computationally challenging feat quickly. Even four-month-olds can distinguish dogs from cats by silhouette alone (Quinn, Eimas, & Tarr, 2001). More extensive representations emerge gradually throughout the first and second years of life (Bergelson & Aslin, 2017; Mandler & McDonough, 1993; Mareschal & Quinn, 2001) as infants robustly recognize objects across different representational formats (e.g., line drawings), map them to labels, and even distinguish them from conceptually similar distractors (Bergelson & Aslin, 2017; DeLoache, Pierroutsakos, Uttal, Rosengren, & Gottlieb, 1998; Zhu, Kilonzo, Zhu, Fan, & Frank, 2025).

How do young children come to understand their visual world so quickly? One key to answering this puzzle is to examine the actual input data for visual categorization, as any developmentally-plausible theory or model of visual learning must operate over these inputs. Videos taken from the infant perspective using head-mounted cameras (Aslin, 2009; Yoshida & Smith, 2008) suggest the child's view of the world is indeed dramatically different from that of adults (Yoshida & Smith, 2008), and varies considerably as children learn to locomote on their own and interact actively with the objects, places, and people around them (Aslin, 2009; Franchak, Kretch, Soska, & Adolph, 2011; Kretch, Franchak, & Adolph, 2014; Long, Sanchez, Kraus, Agrawal, & Frank, 2022; Smith, Yu, Yoshida, & Fausey, 2015; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021; Yoshida & Smith, 2008). Initial evidence analyzing the objects in the infant view suggests that children see some categories dramatically more than others, and experience highly variable, non-canonical views and exemplars of these categories (Clerkin, Hart, Rehg, Yu, & Smith, 2017; Long, Kachergis, Bhatt, & Frank, 2021).

Children's naturalistic visual inputs may even be beneficial for learning. For

example, children's ability to manipulate and rotate objects gives rise to diverse object views, supporting object perception in both young children and models (Bambach, Crandall, Smith, & Yu, 2018; Soska, Adolph, & Johnson, 2010). Additionally, both children and adults can learn visual category distinctions better when categories are sampled from skewed distributions versus uniform distributions (Lavi-Rotbain & Arnon, 2021). In one study, a small number of object categories were both pervasively present during mealtime for 8–10-month-olds (e.g., spoons, cups) and among infants' first-learned words (Clerkin et al., 2017; Clerkin & Smith, 2022).

Despite this initial work, we still have an impoverished understanding of what the actual input data is for visual category learning. For example, how variable are the exemplars that children experience within and across categories? How often do young children tend to see certain categories (e.g., zoo animals) in real life versus as symbolic referents in storybooks or as toys (DeLoache, 2004)? And how consistent are children's visual experiences with object categories across individual households, even within relatively affluent and Western environments (Casey et al., 2022)?

The answers to these questions have implications for both our theories as well as for models of visual category learning by constraining or expanding the set of relevant learning mechanisms. At present, modern neural network models for visual recognition – including convolutional neural networks (CNNs) and vision transformers – show immense promise as instantiations of statistical learning mechanisms that could operate over children's category learning inputs. Intriguingly, activations in these models to images of object categories successfully predict variation in both human object perception and neural responses to the same object categories (Conwell, Prince, Kay, Alvarez, & Konkle, 2024; Yamins et al., 2014). In particular, responses from object-selective cortex, deep neural networks, and human behavior exhibit a consistent geometry, with clustering within broad categories (e.g., animals versus inanimate objects) (Conwell et al., 2024; Muttenthaler & Hebart, 2021).

Yet despite their promise, these models are clearly receiving extraordinarily different kinds and amounts of visual data. Many models still require immense amounts

of curated images to acquire useful representations relative to children (Ayzenberg, Sener, Novick, & Lourenco, 2025; Frank, 2023a; Huber, Geirhos, & Wichmann, 2023) and learn best when given unrealistic amounts of explicitly labeled or captioned photographs (Liu, Li, Wu, & Lee, 2023; Radford et al., 2021; Schuhmann et al., 2022). Thus, despite their correspondence with behavioral and neural data in adults, these models are likely quite far from being mechanistic models of children's visual category learning.

Here, we characterize the input data for category learning from the child's view by analyzing a large, longitudinal dataset of everyday experiences, with the goal of informing the learning mechanism that can explain how children learn visual categories so efficiently. To do so, we capitalize on new data and innovations in computer vision to allow us to annotate these videos. Until recently, existing open egocentric child video datasets have been relatively low-resolution videos with a narrow field of view from a handful of participants (Long et al., 2021; Sullivan et al., 2021) and any analyses required laborious hand annotations. To overcome these challenges, we use the BabyView dataset (Long et al., 2024), an open dataset of egocentric video, whose first release contains 868 hours of data from $N = 31$ children (release 2025.1, ages 5–36 months) from children living mostly in the United States. We also leverage new object detection and classification models (Radford et al., 2021; Siméoni et al., 2025; Wang et al., 2025) to identify common object categories across the entire dataset (Frank, 2023a; Huber et al., 2023).

These new tools and data allow us to analyze the frequency, diversity, and similarity structure of the visual categories in the child's view, and to quantify how often children see real-life exemplars vs depictions across various formats – as toys, drawings, or in media. We compare exemplars of the categories in the child's view to images from THINGS (Stoinski, Perkuhn, & Hebart, 2024), a curated dataset of images used widely in the vision science community. To do so, we use visual representations for categories derived from the embedding spaces from both a self-supervised vision model (Siméoni et al., 2025) and a vision–language model (Radford et al., 2021). We then use

activations from these models to quantify the visual similarity of each category between BabyView and THINGS and the representational geometry that emerges in each dataset (Kriegeskorte, Mur, & Bandettini, 2008).

Using these data, we quantify how different the objects in the child's view are from the objects in curated datasets: We find that the distribution of objects in infants' visual experience is indeed long-tailed, with some categories appearing dramatically more often than others, and that young children's experience contains substantial variation in viewpoint and across representational formats, with a high prevalence of depictions for many categories. Despite the variation, we observed that detections within superordinate categories were *more* similar to each other in the child's view versus in a curated dataset. These results were recapitulated in individual data from densely sampled children, despite their idiosyncratic experiences. Overall, our findings both highlight the dramatic divergence between children's everyday experiences and curated datasets and open new avenues towards understanding how children's everyday category experiences scaffold their learning about the visual world.

Results

Identifying the object categories in the child's view

We first extracted frames at 1 frame per second for a total of 3.68M frames from the 868 hours of the 2025.1 release of the BabyView dataset (Long et al., 2024). We then selected concrete nouns from the MacArthur-Bates Communicative Development Inventories (a commonly used questionnaire for measuring children's vocabulary; CDI) (Marchman, Dale, & Fenson, 2023) and performed category detection for $N = 129$ categories using the YOLOE-v8-L model (Wang et al., 2025); our final list of detected categories was based on iterative rounds of model validation (see Methods). We then filtered all object detections using CLIP, a vision–language model (Radford et al., 2021), to decrease the probability of false alarms (see Supplemental Information (SI) 1.1). This pipeline resulted in a set of 2,994,667 detections for 129 categories across the entire dataset.

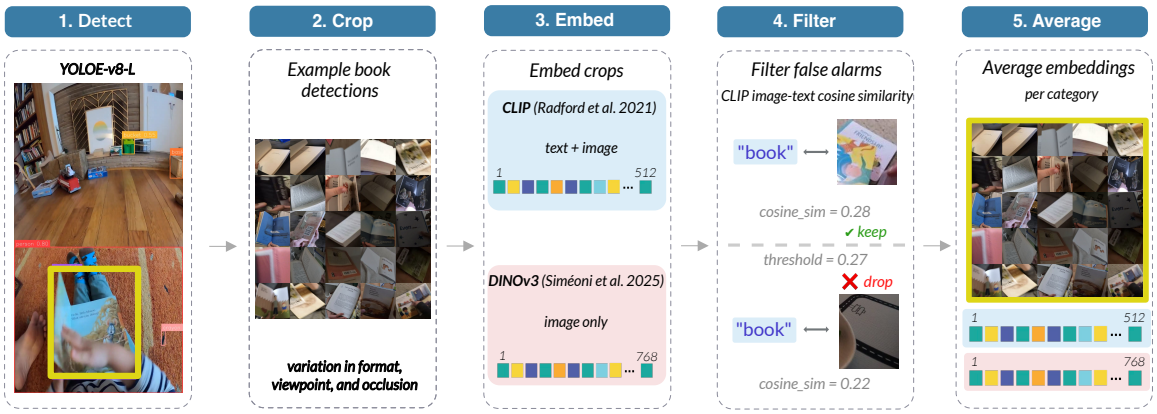

Figure 1. Overview of the automated detection pipeline. (1) YOLOE detects exemplars of CDI vocabulary categories in each frame; (2) bounding boxes are cropped, with example BOOK detections shown. (3) Each crop is embedded with both a vision–language model (CLIP) and a self-supervised vision model (DINOv3). (4) Detections are filtered by the cosine similarity between the CLIP image embedding of the crop and the CLIP text embedding of its predicted label, retaining only crops above a fixed threshold (0.27). (5) For category-level analyses, embeddings of retained crops are averaged within each category and dataset.

For our full set of detections, we validated the precision of the pipeline detections by crowd-sourcing categorizations from human participants for a set of detections (100 per category) that were stratified across participants and videos. Using these data, we observed an average precision of .67 for our 129 categories (SD = .22, see SI 1.2). For a strict set of detections, we constructed a set of human-validated, crowd-sourced detections for the 85 categories that had a precision greater than .60 to yield a "gold" set of 7,018 validated detections; in this subset, the average precision was .80 (SD = .11, range = .61–.99). In both sets of detections, we observed that precision and detection frequency in the dataset were modestly correlated (full set of categories, r = .34; strict set of categories, r = .22; see Supplemental Figure A3), suggesting that variation in observed frequencies is not a direct product of detection accuracy.

As a final check on the robustness of our detection pipeline, we prompted a video question-answering (VideoQA) model, VideoLLaMA3 (Zhang et al., 2025), to detect *all*

objects in 10-second, contiguous chunks of the dataset (Sepuri et al., 2025). We compared these detected categories and their frequencies with those extracted from our pipeline. The observed frequencies were relatively similar across these two distinct model pipelines ($r = .72$, $p < .01$, $N = 99$ overlapping categories, see Supplemental Figure A1). While object detection at scale in naturalistic egocentric video from the child's viewpoint is still a difficult challenge, this level of precision and our filtering pipeline (combined with robustness checks using a smaller set of human annotations) allows us to access a vastly larger amount of data about children's early visual experience than in prior work.

The distribution of objects in infants' visual experience is long-tailed

We observed a skewed distribution of object categories in the child's view: a small number of object categories (e.g., CHAIR, TOY) appeared very frequently while many others (e.g., PENGUIN) appeared less frequently, consistent with prior work (Clerkin et al., 2017). Figure 2 shows the distribution of the top 50 most frequent categories that were detected in the dataset (excluding the most frequent PERSON and PICTURE detections, which only further amplify the skewed distribution); these detections are further broken down by their semantic category using categories from the CDI.

To quantify the shape of these distributions, we fit a power-law function to these frequencies, finding a power-law exponent of $\alpha = 1.93$ across all 129 categories, comparable to the $\alpha = 2.44$ reported by Clerkin et al. (2017) for hand-annotated object frequencies from egocentric videos of infants' mealtimes. Distributions were long-tailed even within each CDI category, with $\alpha = 1.23$ (clothing), $\alpha = 1.98$ (furniture), $\alpha = 1.93$ (household objects), $\alpha = 1.68$ (toys), $\alpha = 2.36$ (body parts), $\alpha = 1.68$ (food and drinks), $\alpha = 1.98$ (outside), $\alpha = 3.13$ (vehicles), and $\alpha = 1.69$ (animals).

As a robustness check, we repeated all key analyses after restricting to human-annotated detections from 85 categories with human-validated precision and in the detections resulting from the VideoQA model pipeline (see Supplemental Figure A1). The core pattern of results was unchanged in this high-precision subset or in the

VideoQA model detections (see Supplemental Figure A2). Overall, these results suggest that skewed distributions of object categories are a highly robust, general property of naturalistic visual environments as experienced by young children.

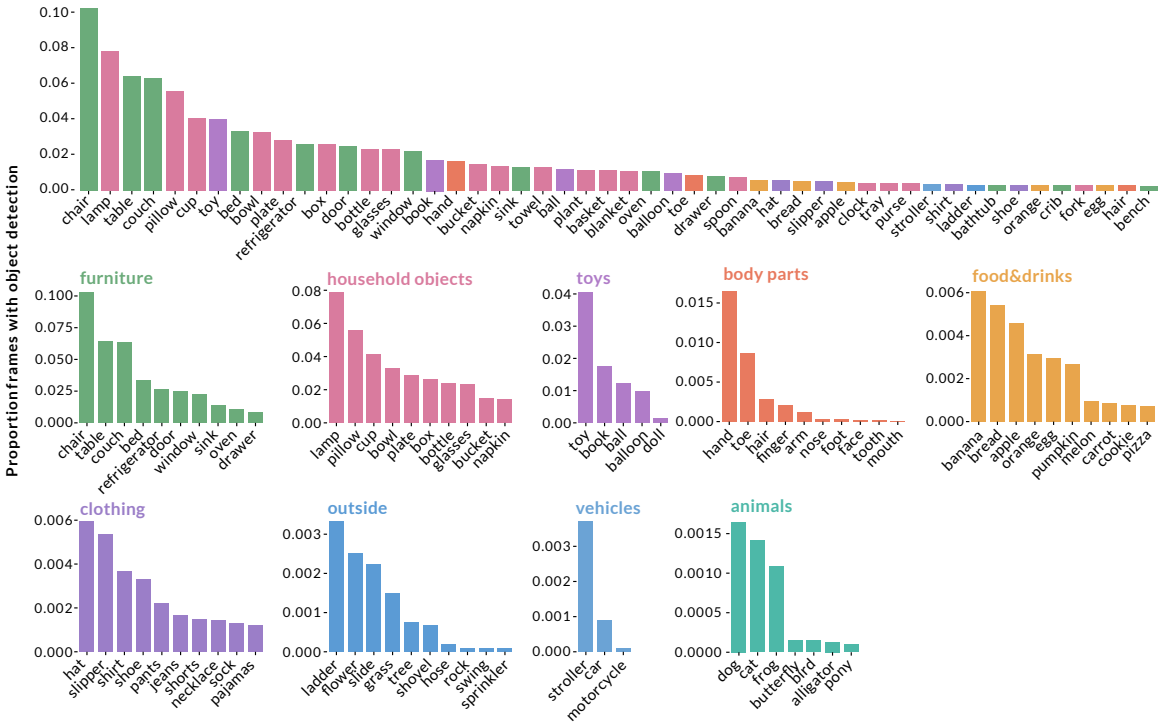

Figure 2. Long-tailed distribution of object categories in young children's everyday visual experience. Each bar shows the proportion of the 3.68M sampled frames containing at least one filtered detection of a given category. Top: the 50 most frequent categories overall, excluding PERSON and PICTURE (which dominate detections and further amplify the skew). Bottom: the most frequent categories within each CDI superordinate domain. Bars are colored by CDI domain.

Object categories in the child's view vary in their similarity to canonical views in curated datasets

Next, we examined how children experience individual exemplars and views of these categories. We compared the visual similarity between the average CUP that infants experienced to the average photograph of a CUP in a curated dataset – THINGS (Stoinski et al., 2024), which is widely used to examine visual representations in both

adults and children. To do so, we first took all filtered object crops in BabyView and all images for each category in the THINGS dataset (see Methods), and extracted embeddings from both a vision–language model (CLIP; Radford et al., 2021) and a self-supervised vision model (DINOv3; Siméoni et al., 2025) for all individual cropped images (see Methods) and averaged across exemplars. In both cases, we computed the cosine similarity between the average category embeddings across the two datasets for 129 categories.[1]

We found that similarity values varied substantially across categories (average cosine similarity between BabyView vs THINGS, CLIP: range = .14–.80, mean $\cos(\theta) = 0.61$; DINOv3: range = .04–.85, mean $\cos(\theta) = 0.48$; see Figure 3). There was relative agreement across the two feature spaces on the question of which categories were more or less similar (Pearson's correlation between DINOv3 vs. CLIP category-wise similarity values: $r = 0.70$, $p < .01$). There were some differences, however, particularly for categories with diverse exemplars, where the DINOv3 embeddings appeared to capture more of this variability. For example, DISH in BabyView tended to be close up views of plates vs. collections of dishes in THINGS (see Figure 3)—here, the DINOv3 embeddings were more sensitive than CLIP to this viewpoint and exemplar variability.

[1] We anticipated that vision–language model (VLM) embeddings might better capture the higher-level similarity between the infant view and curated datasets, for two reasons. First, VLMs are trained on inputs that vary in representational format (e.g., line drawings) and with semantic supervision (i.e., natural language captions for images) – and, as noted, the child's view contains views of many depicted categories. However, we acknowledge that in all following analyses, that there is some circularity that could inflate any observed similarity between the child's view and curated dataset: Since all object detections were filtered for false alarms using embeddings from a VLM, our resulting set of filtered detections more already skewed towards the representational space of the VLM. Thus, the following results are likely to be an upper-bound on the true value, which we suspect might be somewhat lower if we were able to detect all possible *valid* exemplars. Thus, we also utilized embeddings from a self-supervised vision model (DINOv3) to provide an independent estimate of the visual similarity between the detections in these datasets.

Figure 3. Example montages of detected exemplars from the child's view (BabyView) versus a curated dataset (THINGS), for categories with relatively high (left) and low (right) cross-dataset embedding similarity. Cosine similarity values between mean category embeddings are reported separately for CLIP and DINOv3; category labels are colored by their CDI superordinate domain.

However, there are many ways in which the objects in the child's view differ from curated datasets; this variation does not fall neatly along domain boundaries. In our analysis, no broad category domain (animate objects, furniture, toys) was consistently more or less similar to its curated counterpart. The differences we observed not only appears to reflect differences in visual format but also in the range of viewpoints and contexts through which children encountered particular categories; objects in children's everyday experiences are often seen from non-canonical angles, under variable lighting conditions, and often partially occluded.

Variation across visual formats. Qualitatively, we observed that the objects in the child's view spanned a wide range of visual formats: birds were experienced as photographs, alligators were experienced as illustrations and line drawings, and many

children had toy versions of cars, trucks, and trains. However, curated datasets rarely capture this variation in visual format: for example, all of the exemplars in the THINGS dataset are photographs of realistic exemplars. Accordingly, we observed that categories with high variation in visual format across their exemplars (e.g., real objects, toys, and depictions) tended to show lower correspondence between BabyView and THINGS in DINOv3.

We additionally conducted a manual annotation of the detected exemplars for animals to analyze how frequently children experienced real-life exemplars, given the observed range of visual formats they were observed in. Trained annotators simply classified exemplars as belonging to a real-life vs. a depicted viewpoint (including photographs, media, stuffed animals, and line drawings). These analyses revealed that depictions constituted a substantial proportion of the animal exemplars, with some variation across categories. For example, 100% of the exemplars in PONY were depictions (98% in BUTTERFLY, 94% in BIRD, see examples in Figure 4). Conversely, animals that children might experience as household pets – dogs and cats – had overall lower proportions of depictions (23% in DOG, and 5% in CAT) These results suggest that for many categories, the majority of children's learning experiences in the home – and perhaps even outside of it, e.g., in daycare – will consist of depicted exemplars.

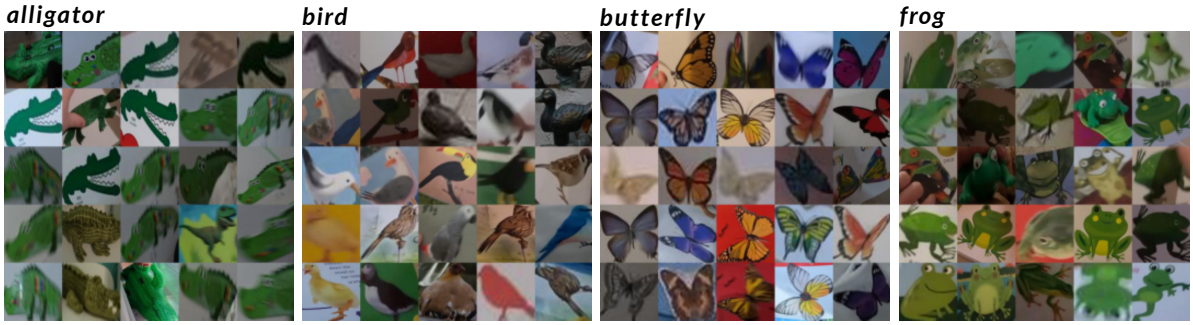

Figure 4. Human-validated exemplars from selected animal categories, illustrating the prevalence of depictions (toys, drawings, photographs, and screen media) in young children's everyday experience.

Between-category representational geometry in young children's object experiences

We used Representational Similarity Analysis (RSA) to compare between-category geometry in BabyView versus THINGS (see Methods). For each dataset, we computed a 129×129 representational dissimilarity matrix (RDM) from pairwise cosine distances between category-mean embeddings, then correlated the strict lower triangles between datasets. Cross-dataset agreement was moderate in both feature spaces (CLIP: $\rho = 0.55$, $p < .01$; DINOv3: $\rho = 0.40$, $p < .01$; Figure 5). Although moderate in absolute size – as expected given differences in viewpoints, context, and noise – these effects indicate reliable agreement in *which* category pairs are relatively similar vs. dissimilar across datasets.

When we examined the RDMs from both BabyView and THINGS, we observed coherent groupings for the CDI superordinate semantic categories: animals, body parts, clothing, large household objects, and small household objects clustered together. For example, the average embedding of CAT was more similar to other animals than to clothing or household objects. We found this to be the case when we constructed RDMs using embeddings from either CLIP or DINOv3, with some variability (see Figure 5).

Remarkably, these broad category clusters appeared overall *stronger* in data from the child's view vs. in THINGS. To quantify this CDI-domain cluster strength, we constructed a between-minus-within distance statistic Δ_d for each superordinate domain (Δ_d^{BV} and Δ_d^{TH}; see Methods), where larger values indicate stronger within-domain clustering. As a random baseline, we then shuffled CDI labels across categories while preserving domain counts (using the same shuffled labeling in BabyView and THINGS on each permutation draw (see Methods; SI 1.7). We first examined activations from CLIP: here, we found that clustering in BabyView was strongest for body parts ($\Delta_d^{\mathrm{BV}} = 0.440$), vehicles (0.403), and furniture (0.390), and weakest for household objects (0.108), a broad and heterogeneous domain. THINGS showed the same rank ordering but weaker separation overall (body parts $\Delta_d^{\mathrm{TH}} = 0.325$, furniture 0.318, vehicles 0.298; household objects 0.087). After within-model false discovery rate

correction, the BabyView > THINGS domain contrast was significant in 6 of 9 domains for CLIP and in all 9 of 9 domains for DINOv3 (See Figure 5C & D). Collectively, these findings indicate that CDI superordinate structure is present in both datasets and is generally more pronounced in the child's naturalistic visual input.

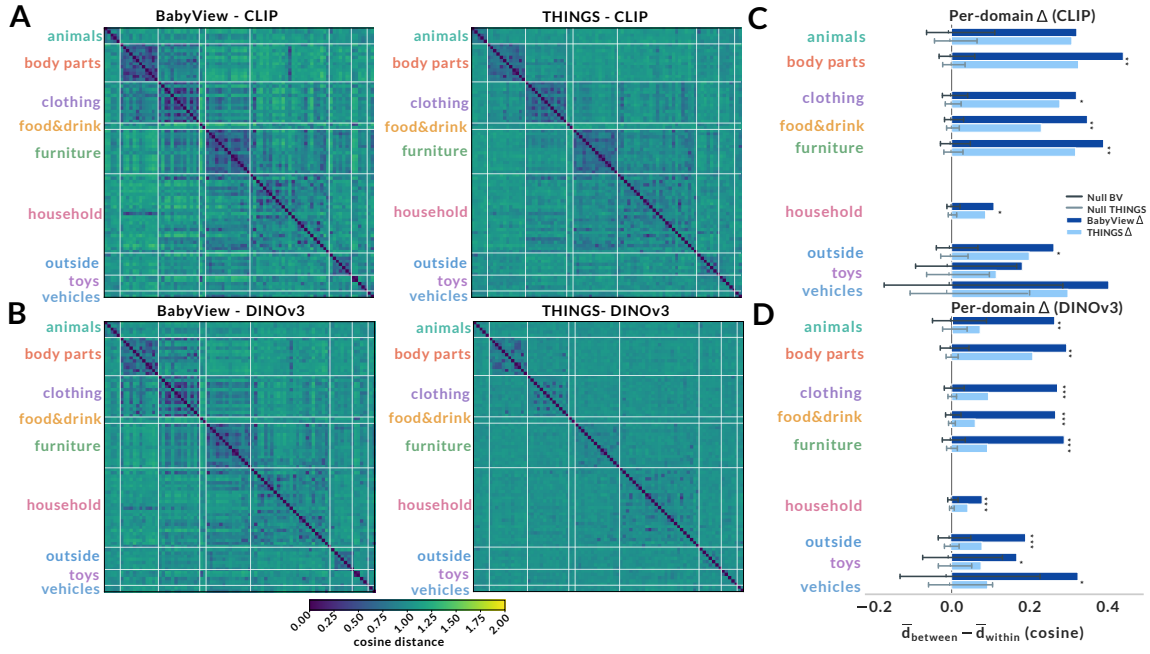

Figure 5. Between-category representational geometry and CDI-domain clustering in BabyView versus THINGS (129 categories). (A) BabyView–CLIP versus THINGS–CLIP RDM comparison. (B) BabyView–DINOv3 versus THINGS–DINOv3 RDM comparison. In both RDM panels, matrices are 129×129 pairwise cosine distances between category centroids, ordered by CDI superordinate domain; darker diagonal blocks indicate stronger within-domain similarity. (C) Per-domain cluster strength in CLIP. (D) Per-domain cluster strength in DINOv3. For panels C–D, bars show $\Delta_d = \bar{d}_{\text{between}} - \bar{d}_{\text{within}}$ for BabyView and THINGS, with permuted null intervals shown for each dataset/domain. Across most domains, clustering is stronger in BabyView than in THINGS.

Idiosyncratic experiences across children support similar category structures

Every child has their own unique home environment. Does the category structure we observed in an average across families reflect consistent properties of each individual

child's visual experience? To examine this question, we constructed separate RDMs for the eight children with the densest recording data (> 32.5 hours of data per family, range = 32.5–91.5 hours, ages 7–29 months). Despite the unique visual environments of each family, the broad organizational structure observed in the aggregate was remarkably consistent. Figure 6 shows the individual RDMs in DINOv3 embedding space, highlighting again that coherent clusters for animals, body parts, and large household objects were identifiable in each family's RDM (categories are in the same order as Figure 5). Statistically, we found that these individual RDMs were relatively similar to each other in embeddings from both models (average between-subject RDM correlation for CLIP: $r = 0.776$, $SD = 0.039$; average between-subject RDM correlation for DINOv3: $r = 0.787$, $SD = 0.032$; see Figure 6). Thus, this category structure appears across the families in our sample in both embeddings from a vision–language model and a purely visual self-supervised model. Overall, these findings indicate that the category structure documented in our aggregate analyses is not merely an artifact of averaging across diverse inputs but is a property that arises from the visual experiences of individual children.

Discussion

How do young children experience visual categories? We quantified the statistical properties of young children's everyday visual experiences with object categories in a large dataset of egocentric videos using innovations in computational models and human annotations. Within our sample, children saw some objects dramatically more than others: a heavily skewed frequency distribution dominated by common household objects, with substantial variability at the exemplar level, consistent with the distributions observed in prior work (Clerkin et al., 2017; Clerkin & Smith, 2022; Long et al., 2021).

Further, we found stronger clusters of individual categories within superordinate groupings in the child's view vs curated datasets – highlighting that children experience *more* redundancy within superordinate groupings than is present in third-person

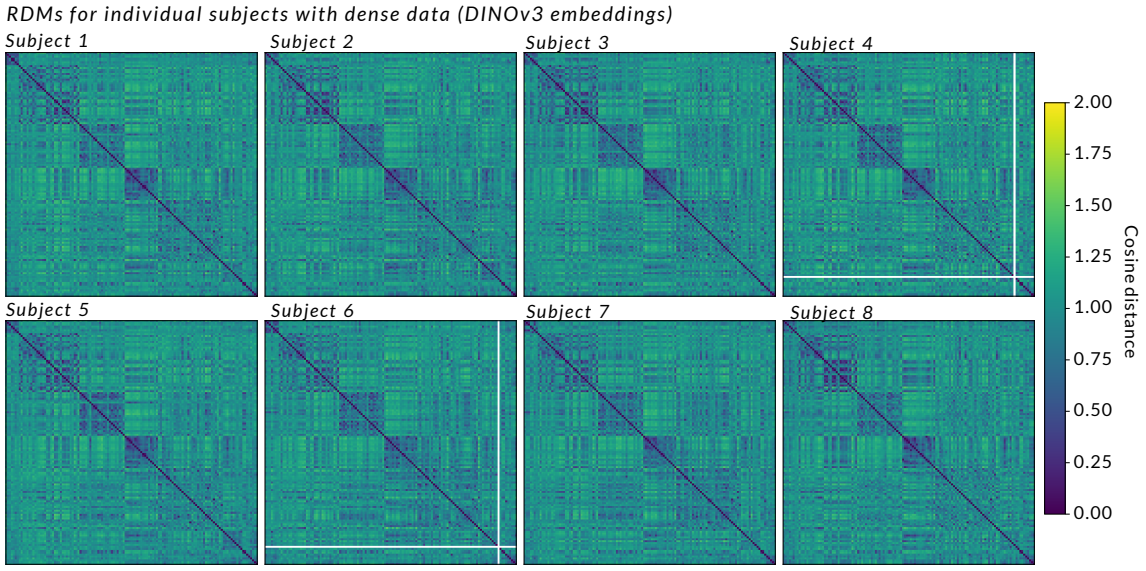

Figure 6. Individual-family RDMs (DINOv3 embeddings, all 129 categories) for the eight participants with the densest recordings. Categories are ordered by CDI superordinate domain, matching Figure 5. Thick white lines indicate missing categories. Darker values indicate visually similar category pairs (low cosine distance); lighter values indicate dissimilar pairs. The same broad superordinate structure visible in the aggregate RDM is recoverable in each individual family, despite idiosyncratic home environments.

curated photographs. Furthermore, we found that this between-category geometry was highly stable across the everyday experiences of the individual children in our dataset.

First, these results constrain theories of visual category learning by extending findings from prior work. Our results reinforce the idea that early category learning operates over repeated encounters with a small set of objects – children's cups, toys, household objects, a handful of real-life animals, and many depictions (Clerkin et al., 2017; Clerkin & Smith, 2022; Long et al., 2021). These skewed distributions may be both ubiquitous in natural environments (Newman, 2005) and in fact beneficial for visual statistical learning in adults (Lavi-Rotbain & Arnon, 2021).

However, our results also highlight how objects may be experienced in dramatically differ ways across broad category domains. For small, handheld objects,

repeated, self-generated views are likely beneficial for building robust 3D object representations. Children who have more experience sitting and rotating objects with their hands tend to succeed on a 3D object completion task (Soska & Adolph, 2014), and variance in these self-generated views tend to predict vocabulary growth (Slone, Smith, & Yu, 2019). However, many object categories in the infant view do not have this property and cannot be rotated or manipulated by children. Indeed, children must navigate around or on top of large, immovable objects (chairs, slides, tables) (Konkle & Caramazza, 2013). And other categories (e.g., many animals) may be mostly experienced as depictions: as "child-directed" visuals that emphasize their diagnostic features. Quantitative theories of children's category learning will need to expand to incorporate these very different kinds and quantities of experience across broad category domains.

Indeed, a key finding of this work is that children's experiences of objects cluster *within* superordinate categories more strongly than in curated datasets: that is, children experience more variable exemplars than is captured in curated datasets, but this additional variability is consistent *within* superordinate clusters. The redundancy within broad, superordinate categories may provide a wedge for children to parse the "blooming buzzing confusion" into meaningful categories that can be mapped to words in their native language (Bergelson & Aslin, 2017). Indeed, this intersects with a foundational idea in cognitive development that children acquire global representations (e.g., animacy) before basic-level representations (Mandler, 2000; Mandler & Bauer, 1988; Quinn & Eimas, 2000). This is consistent with an emerging computational perspective that training computational models with these broad category distinctions – as opposed to basic-level or subordinate level labels – is sufficient for the emergence of human-like object representations (Mehta & Bonner, 2026).

Our results also have implications for research that seeks to emulate child-like learning in machine learning models. While the "data gap" between humans and computational models of visual learning has so far been quantified in terms of hours of data or number of images (Ayzenberg et al., 2025; Frank, 2023b; Huber et al., 2023;

Orhan, 2021) our work suggests that part of the reason for this gap may be a fundamental difference not only in the quantity but in the *types* of objects that children experience. Indeed, while an emerging literature has trained unsupervised models on images and videos from head-mounted camera data (Long et al., 2024; Orhan, 2021; Orhan & Lake, 2024; Zhuang et al., 2021), they have only found mixed success compared with models trained on canonical training datasets such as ImageNet (Deng et al., 2009). Our findings thus point towards a fundamental algorithmic gap between humans and models: the fact that models require such radically different learning inputs suggests they are not simply running a noisier or slower version of the same algorithm; modeling research may need to seek new algorithms fundamentally different in kind. For example, model architectures that can leverage the relational structure between categories might better succeed at learning from these non-canonical, sparse exemplars. Our work does not prescribe a particular algorithmic solution, but contributes a "cognitive target" and suggests properties that future benchmarks should capture.

There are several limitations to these findings. First, the YOLOE detections were far from perfect, highlighting a need for annotation tools that are suited to the child's perspective; while we found convergent results on the frequencies of different categories using detections from a different model class (a VQA model; see Sepuri et al., 2025), we are likely missing certain categories or exemplars in our analysis. Our detection and filtering approach is conservative – that is, by passing all detections through a vision–language model, we likely removed from exemplars or viewpoints of object categories that happen to be even more unusual – and these viewpoints may be relatively challenging for YOLOE to start with. However, we still observed the same pattern of results when we included all possible detections and categories vs. filtered detections. Thus, our results may represent an upper-bound on the similarity between children's everyday experience and curated datasets, and we suspect that a complete characterization of the child's view would reveal even more divergence.

Second, our sample size is relatively small for some of our densest longitudinal analyses, limiting generalizability. Other aspects of the dataset also limit

generalizability: for example, the dataset captures primarily indoor environments due to privacy constraints, potentially missing important aspects of visual experience that occur outdoors and in daycare or caregiving settings. Finally, this dataset captures the experiences of infants from a specific geographic and cultural context, and the generalizability of these statistical patterns to other populations remains an open question (Henrich, Heine, & Norenzayan, 2010).

Overall, this work provides empirical data about the visual input available to developing minds by quantifying the object frequencies, exemplar variability, and category relationships in children's everyday, naturalistic settings. The combination of skewed frequencies, variable exemplars, and stable relational structure characterizes a learning environment that differs markedly from typical laboratory or computational settings. Children's ability to build categorical knowledge from this input highlights the need for developmental theories that can account for how everyday learning occurs across messy, embodied, naturalistic contexts.

Acknowledgments

We gratefully acknowledge the families who participated in the BabyView Dataset. We are grateful to Mira Mateo, Dora Deng, and Jason Yang for assistance with video annotation and ground-truth coding. This work was funded by an NIH R00HD108386 grant to B.L., by a grant from Schmidt Futures, by a gift from Meta, by the Stanford Center for the Study of Language and Information John Crosby Olney Fund, and by the Stanford Human-Centered AI Initiative (HAI) Hoffman-Yee grant program.

Methods & Materials

Dataset

Data for this study came from the 2025.1 release of the BabyView dataset (Long et al., 2023, 2024), consisting of 868 hours of egocentric video recorded using head-mounted cameras from 31 infants (aged 5–36 months) during their everyday activities. Data are available at `https://www.databrary.org/volume/1882`. We

sampled frames at 1 frame per second, resulting in 3.68 million frames for analysis.

Automatic Object Detections using YOLOE

We detected objects in each frame using the YOLOE-v8-L model (Wang et al., 2025). Frames from egocentric videos are out of domain for YOLOE. To validate YOLOE's performance on our dataset, we employed two complementary manual annotation strategies: ground-truth annotations to assess detection completeness and recall, and model corrections to assess precision and label accuracy. First, the detected categories were restricted to align with concrete nouns in the MacArthur-Bates Communicative Development Inventories (CDI) vocabulary items (Marchman et al., 2023) ($N = 295$ words). We included all detections above YOLOE's default confidence threshold of .25. We excluded the frequent PERSON and PICTURE detections from subsequent analyses as they had heterogeneous referents. Detections of PERSON commonly included body parts, such as toes and ears. Detections of PICTURE commonly included detections of photo frames and drawings. Additionally, we excluded any categories that had less than 100 total exemplars, yielding a set of 129 categories.

Ground-truth annotations. Next, two trained coders were each independently given 58 randomly sampled frames (116 frames in total). Annotators were instructed to annotate all possible categories within the CDI vocabulary list ($N = 295$), as well as add new labels if a label were the most suitable to an object but not within the existing CDI vocabulary list. We obtained a limited number of ground-truth frames only because it was labor-intensive. For each object, annotators were also asked to add an overlapping bounding box of PICTURE or TOY to the object if it were a depiction of an object (e.g., "a picture of a dog"). Overall, this initial round of annotations revealed that YOLOE detections were reasonably accurate for $N = 163$ categories; however, many categories were very infrequently present, with sometimes only a few exemplars present per category in the frames themselves.

Detection filtering. We aimed to exclude false alarms for each category that could increase noise in our similarity analyses. To create a cleaner subset of

high-confidence true positive detections for the THINGS comparison, we implemented a CLIP-based filtering procedure. For each YOLOE detection, we extracted the cropped image region within the bounding box and passed it through CLIP's image encoder (ViT-B/32; Radford et al., 2021), yielding a 512-dimensional image embedding. We also encoded the predicted category label (e.g., CUP, DOG, CHAIR) using CLIP's language encoder, yielding a corresponding text embedding. We then computed the cosine similarity between the image embedding and its associated label embedding. This similarity score reflects how well the visual content of the cropped region matches the semantic meaning of the predicted category label according to CLIP's learned visual–linguistic alignment.

Detections were retained only if their image-text cosine similarity exceeded a threshold of .27, which was determined through qualitative examination of detections at various thresholds. At this threshold, we observed that retained detections were predominantly true positives with clear, recognizable instances of the labeled category, while filtered detections often contained either incorrect objects, extremely partial views, heavy occlusions, or empty/ambiguous regions where YOLOE had hallucinated an object. For example, a detection labeled DOG showing a clear, frontal view of a dog would typically achieve similarity scores above .30, while a false alarm showing a piece of furniture mislabeled as DOG would score well below .26. This filtering procedure removed a substantial majority of the original YOLOE detections, with filtering rates varying by category. Categories with clearer visual appearance and less ambiguity (e.g., BALL, BOOK) had lower filtering rates, while categories with more variable appearance or greater susceptibility to YOLOE errors (e.g., TOY, FOOD) had higher filtering rates.

At this threshold, CLIP filtering removed 83.06% of raw YOLOE detections overall (retaining 2,994,667 of 17,674,191 detections). Filtering rates varied substantially across categories. Because this filter is conservative, it likely removes some true positives with atypical appearance, partial visibility, or unusual viewpoints, potentially biasing retained detections toward more prototypical exemplars.

Model detection verification. We assessed the precision of model detections by asking human annotators to verify filtered detections via a crowd-sourced annotation task. We obtained three independent human ratings verifying our model detections for each of 100 exemplars from 129 different categories (12,900 crops in total). To construct this validation set, we sampled from CLIP-filtered detections with additional diversity constraints: categories were required to have at least 100 eligible exemplars across a large spread of subjects and videos (see SI 1.2).

This model verification strategy thus allows us to compute the accuracy of the filtered detections for each category. However, this annotation task was considerably more challenging than annotating in the objects in the full-view frames taken from the video, due to the loss of the surrounding contextual information. In our main analyses, we thus report the main results for all $N = 129$ categories, as well as for a stringent filtered subset of $N = 85$ categories that have a precision $> .60$. This dual reporting strategy is intended to preserve comparability while quantifying sensitivity to false-alarm-prone categories.

Object representations

We investigated the category structure of the BabyView dataset using embedding representations and representational similarity analysis (Kriegeskorte et al., 2008).

Object embeddings. We embedded all object crops from the filtered dataset using the CLIP image encoder (ViT-B/32; Radford et al., 2021 as well as the DINOv3 image encoder (Vit-B/16; Siméoni et al., 2025). CLIP embeddings represent more semantically aligned representations (due to the joint language–image pretraining), whereas DINOv3 embeddings represent more vision-focused representations (due to image-only self-supervised pretraining).

Category-level summaries. For each included category we summarized detection frequencies from the filtered frame-level detections (1 frame per second sampling) and grouped categories by CDI superordinate domain labels used throughout the manuscript. For BabyView versus THINGS alignment at the *exemplar* level, we

compared mean category embeddings using cosine similarity (CLIP and model-matched DINOv3), as reported in the Results.

CLIP-threshold sensitivity analysis. To test whether representational results depended on the CLIP image–text operating threshold, we reran the $N = 129$ aggregate pipeline at thresholds $t \in \{.26, .27, .28\}$ while holding category inclusion and category order fixed; see Supplemental Figure A6. For each threshold, BabyView category centroids were recomputed from the retained detections; THINGS centroids were unchanged. We then recomputed the RDM-based diagnostics used in the main analyses (BabyView–THINGS Spearman correlation in CLIP and DINOv3 spaces, and within-dataset CLIP–DINO RDM agreement), and also tracked total retained detections within the included category scope. This analysis quantifies the tradeoff between stricter filtering (fewer retained detections) and stability of the recovered representational geometry.

Representational dissimilarity matrices (RDMs) and cross-dataset alignment

Construction and Spearman correlation. We asked whether the *pattern* of between-category distances in BabyView matches THINGS when both datasets use the same $N = 129$ categories in the same order. For each embedding model (CLIP and DINOv3), we averaged exemplar embeddings within category (separately for BabyView and THINGS), computed all pairwise cosine distances between category centroids, and formed one symmetric RDM per dataset. We then vectorized each RDM's strict lower triangle (one entry per unique category pair) and computed Spearman's rank correlation ρ between BabyView and THINGS vectors. Two-sided p-values came from the standard Spearman test on those paired vectors.

Label-shuffle null for cross-dataset ρ. To test whether cross-dataset ρ depends on true category alignment, we implemented a *single-sided label-shuffle null* (separately for CLIP and DINOv3). BabyView was fixed. On each of $n_{\text{perm}} = 5{,}000$ draws, we sampled one random permutation of the 129 category labels and applied it to THINGS *rows only* (columns unchanged). We then vectorized the strict lower triangle

of this row-permuted THINGS matrix and recomputed Spearman's ρ against the fixed BabyView lower-triangle vector. Empirical two-sided p-values were computed as $(k{+}1)/(n_{\text{perm}}{+}1)$, where k is the number of null draws with $|\rho_{\text{null}}| \geq |\rho_{\text{obs}}|$.

CDI superordinate alignment in embedding space

We quantified CDI superordinate structure using a *per-domain* statistic in each fixed RDM, $\Delta_d = \overline{d}_{\text{between}(d)} - \overline{d}_{\text{within}(d)}$, computed separately for BabyView and THINGS. Inference/benchmarking for this analysis used label shuffling across categories with domain counts preserved (5,000 draws), with the same shuffled domain assignment applied to both datasets on each draw.

Domain-level cluster strength. Let $D(i)$ denote the CDI domain of category i and d_{ij} the pairwise cosine distance between categories i and j in a fixed RDM ($i < j$ throughout). For each domain d, we compute

$$\Delta_d \;=\; \overline{d}_{\,\text{between}(d)} - \overline{d}_{\,\text{within}(d)},$$

where $\overline{d}_{\,\text{within}(d)}$ is the mean pairwise distance among categories assigned to domain d, and $\overline{d}_{\,\text{between}(d)}$ is the mean distance from categories in d to categories outside d. We denote this quantity as Δ_d^{BV} for BabyView and Δ_d^{TH} for THINGS. Larger Δ_d indicates stronger superordinate clustering for that domain (between-domain distances exceed within-domain distances by a larger margin). Because domains reuse the same underlying pairwise distances, these per-domain summaries are not independent; we therefore interpret them primarily as a profile of *where* clustering is stronger vs. weaker across the CDI taxonomy.

Marginal label-shuffle intervals for Δ_d. To provide a visual null benchmark for each Δ_d, we estimated *marginal* label-shuffle intervals. On each draw, we held both RDMs fixed and shuffled CDI domain labels across categories (category identities and distances unchanged), while preserving exact domain counts. The same shuffled domain assignment was applied to BabyView and THINGS in parallel on that draw. We recomputed Δ_d^{BV} and Δ_d^{TH} for all domains, repeated this over 5,000 draws, and took the 2.5th and 97.5th percentiles as the central 95% interval. Under this null, a label such as

"animals" refers to a random subset of categories of matching size, not the true animal set. These intervals are used as a descriptive visual benchmark, not as a multiplicity-corrected decision rule across domains. For clarity of presentation, we visualize this same 2.5–97.5% marginal null interval in equivalent styles (e.g., capped whiskers/error bars in the supplement, shaded bands where used elsewhere). All styles are constructed from identical per-domain permutation draws and differ only in rendering.

Agreement across domains between datasets. We quantified agreement between BabyView and THINGS in the *pattern* of domain strengths by Spearman rank correlation between the paired vectors $\{\Delta_d^{\mathrm{BV}}\}_d$ and $\{\Delta_d^{\mathrm{TH}}\}_d$ within each embedding model (two-sided p-values from SciPy's `spearmanr`).

Per-domain BabyView-versus-THINGS inference and FDR control. For each CDI domain d and embedding model, we quantified the directional contrast $\Delta_d^{\mathrm{BV}} - \Delta_d^{\mathrm{TH}}$ and computed a one-sided permutation p-value for the hypothesis BabyView > THINGS using the shared count-preserving label-shuffle draws described above. Because this yields a family of nine per-domain tests per model, we controlled false discoveries within each model using Benjamini–Hochberg false discovery rate correction across domains and report corresponding q-values in supplemental outputs.

Analyses of individual-family RDMs (dense-recording subsample)

To examine whether aggregate geometry reflects stable structure within individual children's homes, we identified the eight families with the densest usable recording time under the same filtering and category inclusion rules used for the aggregate RDMs. Operationally, we ranked families by total hours of analyzed video contributing to the included category set and selected the top eight families that each exceeded 32.5 hours (range 32.5–91.5 hours; child ages at recording 7–29 months). For each family we computed centroid RDMs in CLIP and DINOv3 embedding spaces using the same pipeline as for the aggregate BabyView RDM. We then quantified pairwise similarity of these family-specific RDMs by Spearman correlation over the strict lower triangle

(restricted to category pairs present for both families in a given comparison) and summarized the distribution of between-family correlations (mean and SD across pairs), providing empirical bounds on how consistent superordinate structure is across idiosyncratic visual environments.

References

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and Vision Science*, *86*(6), 561–565.

Ayzenberg, V., Sener, S. B., Novick, K., & Lourenco, S. F. (2025). Fast and robust visual object recognition in young children. *Science Advances*, *11*(27), eads6821.

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in neural information processing systems*, *31*.

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Casey, K., Elliott, M., Mickiewicz, E., Silva Mandujano, A., Shorter, K., Duquette, M., ... Casillas, M. (2022). Sticks, leaves, buckets, and bowls: Distributional patterns of children's at-home object handling in two subsistence societies. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

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

Clerkin, E. M., & Smith, L. B. (2022). Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences*, *119*(18), e2123239119.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, *15*(1), 9383.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in cognitive sciences*, *8*(2), 66–70.

DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengren, K. S., & Gottlieb, A. (1998). Grasping the nature of pictures. *Psychological Science*, *9*(3), 205–210.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child development*, *82*(6), 1738–1750.

Frank, M. C. (2023a). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992.

Frank, M. C. (2023b). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61–83.

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2023). The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *Journal of vision*, *23*(7), 4–4.

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

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child development*, *85*(4), 1503–1518.

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

Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in zipfian distributions. *Cognition*, *206*, 104492.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.

Long, B., Goodin, S., Kachergis, G., Marchman, V. A., Radwan, S. F., Sparks, R. Z., ... others (2023). The babyview camera: Designing a new head-mounted camera to capture children's early social and visual environments. *Behavior Research Methods*, 1–12.

Long, B., Kachergis, G., Bhatt, N. S., & Frank, M. C. (2021). Characterizing the object categories two children see and interact with in a dense dataset of naturalistic

visual experience. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Long, B., Sanchez, A., Kraus, A. M., Agrawal, K., & Frank, M. C. (2022). Automated detections reveal the social information in the changing infant view. *Child Development*, *93*(1), 101–116.

Long, B., Sparks, R. Z., Xiang, V., Stojanov, S., Yin, Z., Keene, G. E., ... others (2024). The babyview dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *arXiv preprint arXiv:2406.10447*.

Mandler, J. M. (2000). What global-before-basic trend? commentary on perceptually based approaches to early categorization. *Infancy*, *1*(1), 99–110.

Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? *Cognitive development*, *3*(3), 247–264.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive development*, *8*(3), 291–318.

Marchman, V. A., Dale, P. S., & Fenson, L. (2023). *The macarthur-bates communicative development inventories: User's guide and technical manual, 3rd edition*. Brookes Publishing Company.

Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. *Trends in cognitive sciences*, *5*(10), 443–450.

Mehta, Y., & Bonner, M. F. (2026). An extremely coarse feedback signal is sufficient for learning human-aligned visual representations. *arXiv preprint arXiv:2605.05556*.

Muttenthaler, L., & Hebart, M. N. (2021). Thingsvision: a python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics*, *15*, 679838.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, *46*(5), 323–351.

Orhan, A. E. (2021). How much human-like visual experience do current self-supervised learning algorithms need in order to achieve human-level object recognition? *arXiv preprint arXiv:2109.11523*.

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

Quinn, P. C., & Eimas, P. D. (2000). The emergence of category representations during infancy: Are separate perceptual and conceptual processes required? *Journal of Cognition and development*, *1*(1), 55–61.

Quinn, P. C., Eimas, P. D., & Tarr, M. J. (2001). Perceptual categorization of cat and dog silhouettes by 3-to 4-month-old infants. *Journal of experimental child psychology*, *79*(1), 78–94.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... others (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, *35*, 25278–25294.

Sepuri, T., Aw, K., Tan, A., Sparks, R., Marchman, V., Frank, M., & Long, B. (2025). Characterizing young children's everyday activities using video question-answering models. *NeurIPS Dataset and Benchmarks Workshop*.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., ... others (2025). Dinov3. *arXiv preprint arXiv:2508.10104*.

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental science*, *22*(6), e12816.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407–419.

Soska, K. C., & Adolph, K. E. (2014). Postural position constrains multimodal object exploration in infants. *Infancy*, *19*(2), 138–161.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: Motor

skill acquisition facilitates three-dimensional object completion. *Developmental Psychology*, *46*(1), 129–138. Retrieved from http://dx.doi.org/10.1037/a0014618 doi: 10.1037/a0014618

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2024). Thingsplus: New norms and metadata for the things database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, *56*(3), 1583–1603.

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

Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., & Ding, G. (2025). Yoloe: Real-time seeing anything. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 24591–24602).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? using a head camera to study visual experience. *Infancy*, *13*(3), 229–248.

Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., ... others (2025). Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhu, R., Kilonzo, T. N., Zhu, L. Z., Fan, J. E., & Frank, M. C. (2025). Cross-contextual variability in children's early understanding of visual media. *Topics in Cognitive Science*.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.

Appendix

Supplemental Information

SI 1.1. CLIP filtering pipeline and false-alarm control

To reduce category-specific false alarms from open-vocabulary YOLOE detections, we applied a CLIP-based image-text consistency filter to every detection crop. For each predicted object, we computed cosine similarity between the CLIP image embedding of the crop and the CLIP text embedding of the predicted category label, and retained detections only when similarity exceeded .27.

At this threshold, CLIP filtering removed 83.06% of raw YOLOE detections overall, retaining 2,994,667 of 17,674,191 detections. Filtering rates varied by category, with ambiguous categories showing higher rejection rates than visually specific categories. This conservative filter improves precision for downstream representational analyses, while likely excluding some true positives with atypical viewpoints, heavy occlusion, or low-resolution appearance.

SI 1.2. Crowdsourced annotation protocol and derivation of the human-validated subset

We constructed a crowdsourced validation set by sampling 100 detection crops per category from the CLIP-filtered YOLOE detections (129 categories; 12,900 regular crops in total in the annotation manifest). Annotators ($N = 12$) were shown 25 images at a time and asked to "Please identify all invalid detections of [cat]"; each trial included a pre-specified invalid image as an attention check. Annotators who failed more than 20 attention checks were replaced. Sampling was constrained to increase diversity and reduce over-representation from repeated clips: categories were required to have at least 100 exemplars after CLIP filtering, at least 8 unique subjects, and sufficient spread across videos in the dataset to include at most 2 exemplars from a single video when forming the annotation set (average video length = 510.77s).

Each sampled crop received 3 independent human ratings (per-file table: $N_{\text{raters}} = 3$). For each crop, we computed an error rate from the crowd responses and

converted this to crop-level precision (precision = 1 − error rate); class-level precision was then computed by averaging crop-level precision within category. Across the 129 categories, mean class precision was .669 (reported as 0.67 in the main text; $SD = 0.22$).

For high-precision robustness analyses, we defined a stringent subset of 85 categories using a fixed threshold of precision > .60 at both levels: (i) category-level precision and (ii) exemplar-level precision, intersected with the pre-specified category list (N = 129) and the sampled manifest rows. This procedure yielded 85 categories and 7,018 unique human-validated exemplar crops (the ∼7k set used in supplemental robustness analyses).

SI 1.3. Convergence with video question-answering model detections

To validate our automated pipeline, we prompted a video question-answering (VideoQA) model, VideoLLaMA3 (Zhang et al., 2025), to detect all objects in 10-second, contiguous chunks of the dataset. We utilized the same chunks of the dataset as presented in Sepuri et al. (2025), sampling every third chunk for a total of $N =$ 289.33 hours. We prompted the model with the phrase "This a video from the point-of-view of a camera mounted on a child's head. Strictly return a list detailing each object, animal and person present in this video, comma separated like so: 'ball,tennis racquet,person,sofa'." We compared detections both for the full set of $N =$ 129 categories and the stringent high-precision subset of $N = 85$ categories. For the full set of categories, we found 99 categories that overlapped between the VideoQA model and our automated detections. The 30 missing categories were all small objects (for example, TOE and ORANGE). One reason these detections were missing in longer videos could be that they were not salient across longer 10-second chunks of the dataset and occurred for shorter durations.

We found strong correlations in the observed frequencies of overlapping categories across both the full set of categories ($r = .72$, $p < .01$, $n = 99$) and the stringent high-precision subset of categories ($r = .69$, $p < .01$, $n = 70$). Additionally, we observed a skewed distribution of objects. We fit a power-law function to the frequencies in the

high precision set of categories, finding a power-law exponent of $\alpha = 2.57$ across the overlapping 70 categories, comparable to the $\alpha = 2.44$ reported by Clerkin et al. (2017) and the $\alpha = 1.93$ found with our YOLOE model pipeline.

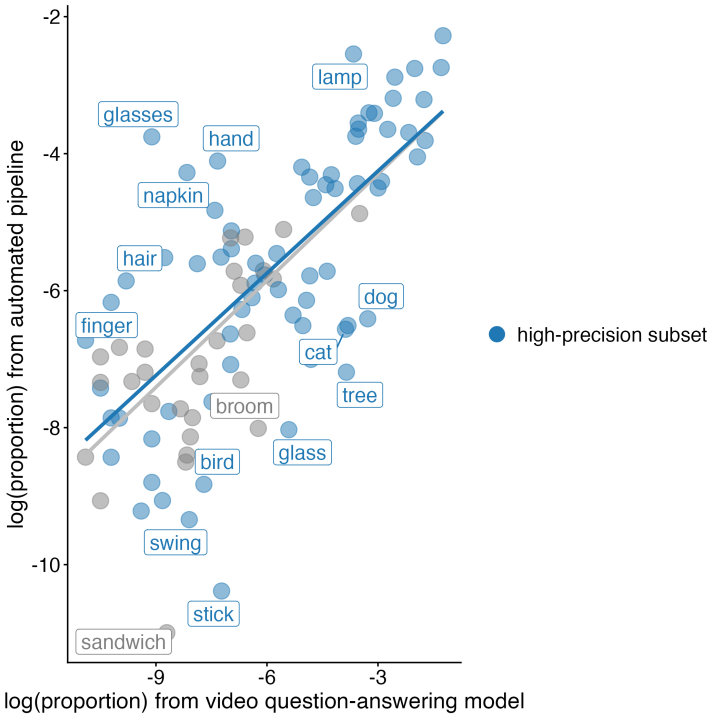

Figure A1. Convergence between the automated YOLOE+CLIP pipeline and an independent video question-answering pipeline (VideoLLaMA3). Each point is a category, plotted by its log proportion of detections in each pipeline. Gray points and regression line: full set of overlapping categories ($N = 99$); blue points and line: high-precision human-validated subset ($N = 70$). Thirty categories detected by YOLOE but not returned by the VQA model are excluded.

SI 1.4. Long-tailed category distribution in the high-precision subset ($N = 85$)

We observed the same long-tailed structure when restricting analyses to the high-precision human-validated subset ($N = 85$). Fitting a power-law function to category frequencies yielded an exponent of $\alpha = 1.92$ across the 85 included categories. Semantic-level fits (when sufficient categories were available) also showed long-tailed structure, with estimated exponents of $\alpha = 1.97$ (animals), $\alpha = 2.35$ (body parts),

$\alpha = 1.23$ (clothing), $\alpha = 2.00$ (furniture), $\alpha = 1.61$ (household objects), $\alpha = 2.30$ (outside), and $\alpha = 1.68$ (toys). Food & drinks and vehicles each contained only two categories in this subset and were therefore too sparse for stable semantic-level power-law fitting.

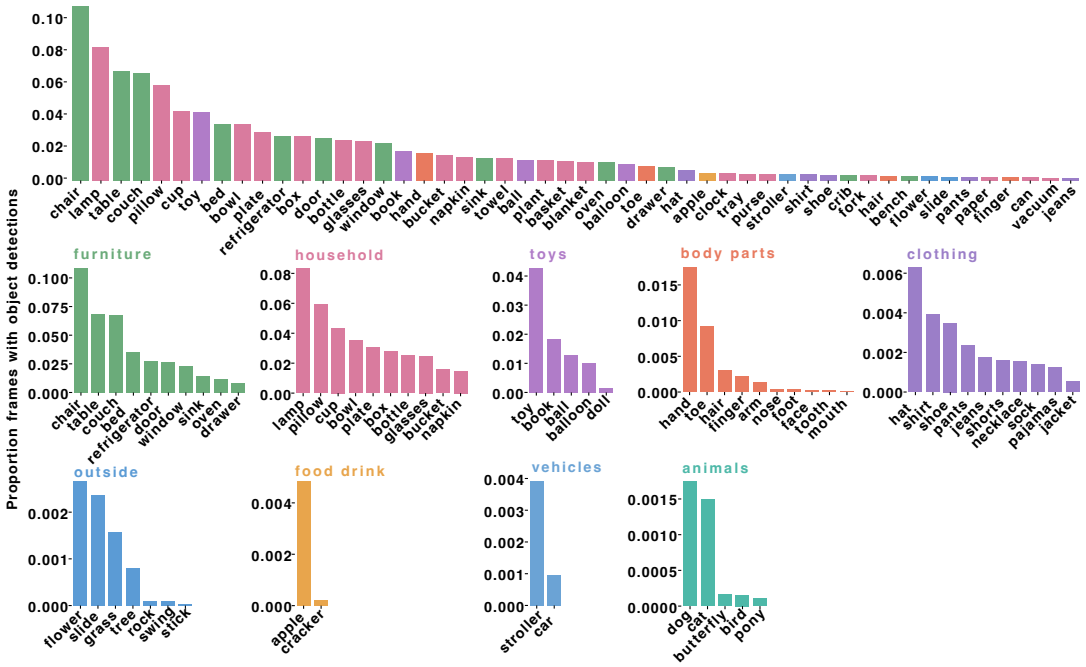

Figure A2. Long-tailed distribution of object categories in the high-precision human-validated subset ($N = 85$). As in Figure 2, each bar shows the proportion of the 3.68M sampled frames containing at least one filtered detection. Top: all 85 categories, ranked by frequency. Bottom: category frequencies within each CDI superordinate domain. The skewed shape observed in the full set of 129 categories is preserved in this stricter subset (overall power-law exponent $\alpha = 1.92$).

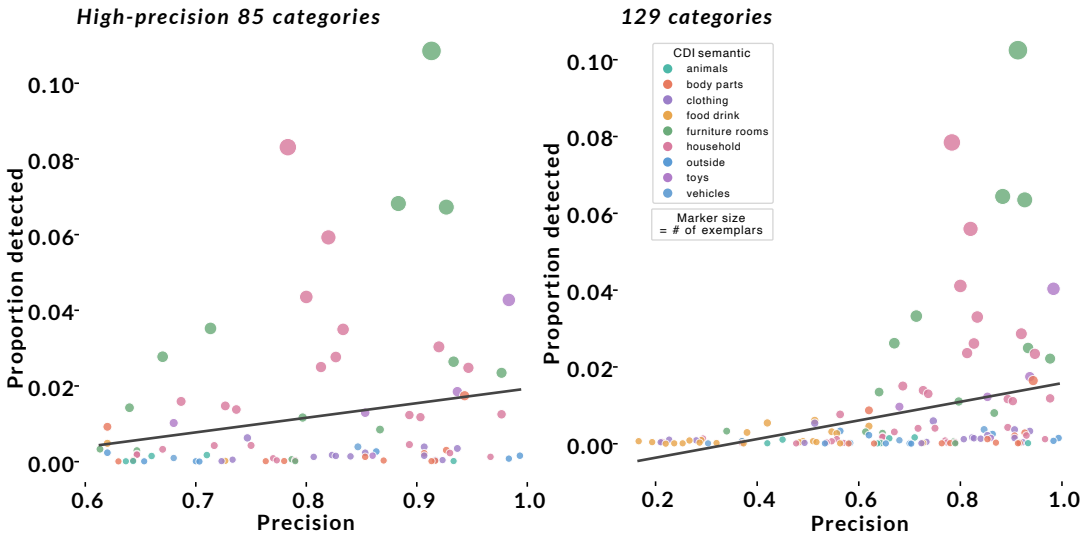

Figure A3. Relationship between detection frequency and human-validated precision. Each point is a category, plotted by its crowdsourced precision (x) against its proportion of CLIP-filtered detections within the analysis set (y). Left: high-precision subset ($N = 85$); right: full set ($N = 129$). Points are colored by CDI superordinate domain and sized by total detection count; lines show ordinary least-squares fits. Frequency and precision are only modestly correlated, indicating that the most frequent categories are not disproportionately driving precision estimates.

SI 1.5. Category-wise BabyView–THINGS similarity distribution ($N = 85$)

For category-wise similarity (BabyView vs. THINGS), values showed substantial heterogeneity across the 85 categories in both embedding spaces. In CLIP space, cosine similarity ranged from .32 to .82 (mean = .62, median = .63, $SD = .11$). In DINOv3 space, cosine similarity ranged from .07 to .84 (mean = .50, median = .49, $SD = .18$).

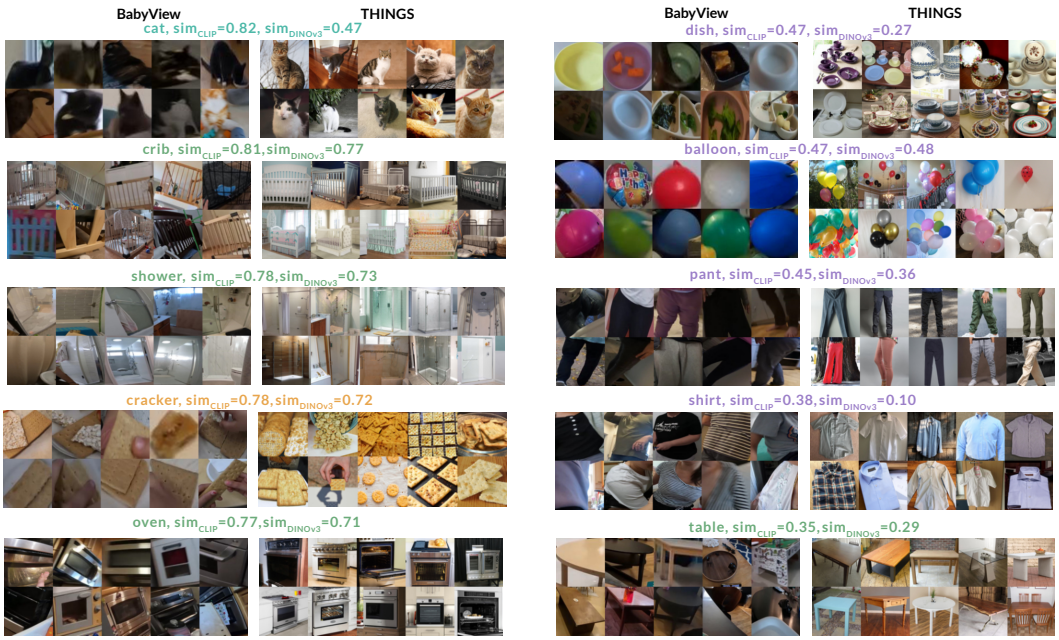

Figure A4. Example exemplar montages for the high-precision subset ($N = 85$), shown as in Figure 3. Left: categories with relatively high cross-dataset similarity between BabyView and THINGS; right: categories with relatively low similarity. Cosine similarity values are reported separately for CLIP and DINOv3.

SI 1.6. Representational geometry convergence with THINGS ($N = 85$)

Using the high-precision human-validated subset ($N = 85$), we observed moderate convergence between representational structure in BabyView and THINGS. In CLIP space, the BabyView-vs-THINGS RDM correlation was Spearman's $\rho = .57$ ($p < .01$); in DINOv3 space, the corresponding correlation was Spearman's $\rho = .46$ ($p < .01$).

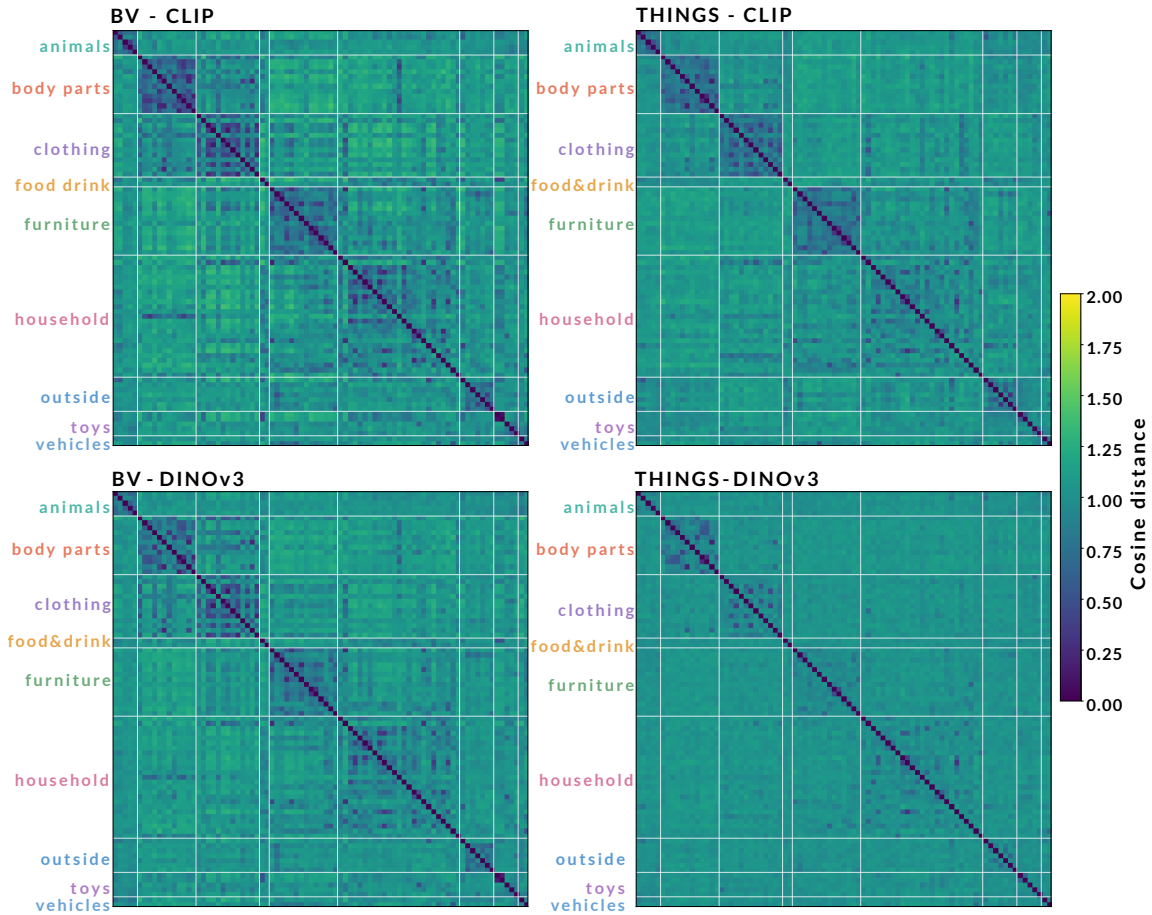

Figure A5. Between-category representational geometry for the high-precision subset ($N = 85$ categories). Each panel shows an 85×85 RDM of pairwise cosine distances between mean category embeddings, with categories ordered by CDI superordinate domain. Top row: CLIP; bottom row: DINOv3; left column: BabyView; right column: THINGS. The aggregate pattern from Figure 4 is preserved: superordinate clusters are visible in both datasets and are more pronounced in BabyView (CLIP: Spearman's $\rho = .57$; DINOv3: Spearman's $\rho = .46$; both $p < .01$).

SI 1.7. Within- versus between-cluster separation for CDI semantic groups

Procedures for domain-level Δ_d and its label-permutation null ($n_{\text{perm}} = 5{,}000$) are specified in *Methods* (main text). Here we restate the implemented shuffling procedure for clarity and show the null interval as whiskers for readability.

For each CDI domain d and model, we compute $\Delta_d = \overline{d}_{\text{between}} - \overline{d}_{\text{within}}$ separately for BabyView and THINGS. On each permutation draw, we keep both RDMs fixed and

shuffle CDI superordinate labels across categories (equivalently: random category-to-domain reassignment), while preserving the exact count of categories per domain. The same shuffled label vector is then applied in parallel to BabyView and THINGS for that draw. We recompute Δ_d under this shuffled labeling and repeat over 5,000 draws. Whiskers denote the 2.5th and 97.5th percentiles of these per-domain null draws (central 95% interval); bars denote observed Δ_d. Thus, whiskers are a visual benchmark for "how large Δ_d is under label-randomized domains" and are not a multiplicity-corrected decision rule across domains.

Domain-level Δ_d was largest for body parts, vehicles, and furniture/rooms and smallest for household in both embeddings; Spearman rank correlation of $\{\Delta_d^{\mathrm{BV}}\}$ vs. $\{\Delta_d^{\mathrm{TH}}\}$ across domains was CLIP $\rho = 0.88$, $p = 0.0016$, and DINOv3 $\rho = 0.73$, $p = 0.025$ ($k = 9$ domains), indicating similar domain ordering with larger separation in BabyView for most domains.

For per-domain directional contrasts (BabyView > THINGS), we computed one-sided permutation p-values and applied Benjamini–Hochberg FDR correction across the 9 domains within each model. In CLIP, 6/9 domains survived FDR ($q < .05$: body parts, clothing, food/drink, furniture/rooms, household, outside), while animals, toys, and vehicles did not. In DINOv3, all 9/9 domains survived FDR ($q < .05$), indicating a more pervasive BabyView > THINGS separation profile in that embedding space.

SI 1.8. Agreement between CLIP and DINOv3 pairwise geometry

For each dataset (BabyView and THINGS) and each category set, we formed the symmetric RDM of pairwise cosine distances between mean category embeddings, then correlated the off-diagonal entries of the CLIP RDM with those of the DINOv3 RDM (same category order; CLIP detection filter threshold .27 as elsewhere). Table A1 summarizes Pearson and Spearman correlations ($n(n-1)/2$ pairs). Agreement was much stronger for BabyView than for THINGS, indicating that CLIP and DINOv3 largely recover the same between-category structure in infant-view centroids but diverge more on THINGS centroids under the same category inventory.

Table A1

Correlation between CLIP and DINOv3 cosine-distance RDMs (vectorized lower triangle, excluding the diagonal).

Category set	Dataset	Pearson r	Spearman ρ	Pairwise pairs
$N = 129$	BabyView	0.907	0.874	8,256
$N = 129$	THINGS	0.679	0.487	8,256
$N = 85$	BabyView	0.889	0.839	3,570
$N = 85$	THINGS	0.702	0.514	3,570

All p-values were effectively zero at machine precision ($p < 10^{-100}$ for the reported coefficients); we omit exact floating-point p here.

SI 1.9. CLIP-threshold sensitivity for representational geometry and retained detections (N=129)

To evaluate sensitivity to the CLIP image–text detection filter, we repeated the $N = 129$ analysis over thresholds .26, .27, and .28 while keeping category scope and ordering fixed. For each threshold, BabyView category centroids were recomputed from retained exemplar crops, and RDM-based geometry metrics were recalculated against fixed THINGS centroids. We also tracked the total number of CLIP-filtered detections retained within the included category scope.

Figure A6 reports the two diagnostics used in the main robustness check: (A) representational geometry correlations across thresholds, and (B) total retained detections by category scope. Across this range, the geometry pattern is stable (small variation in RDM correlations) while retained-detection totals change monotonically with threshold, indicating that the main geometric conclusions are not driven by a single idiosyncratic cutoff at .27.

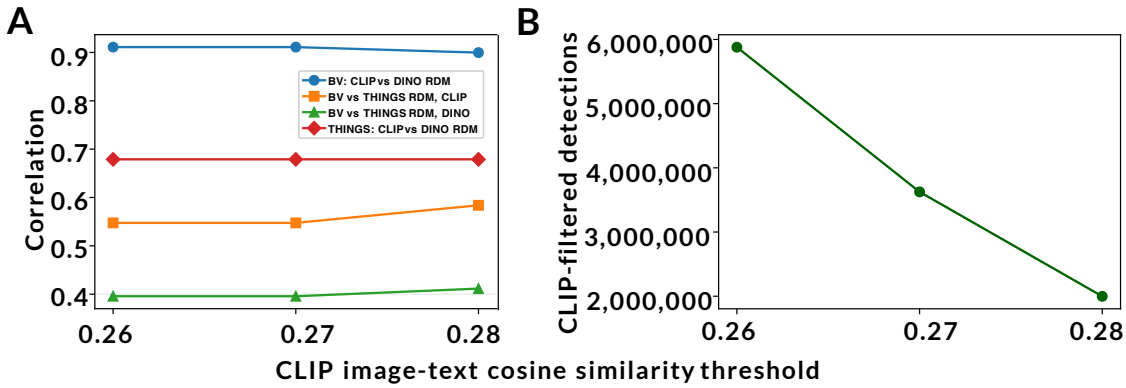

Figure A6. CLIP-threshold sensitivity analysis for the 129-category set. (A) Representational geometry correlations across thresholds, including BabyView–THINGS RDM correlations (CLIP and DINOv3) and within-dataset CLIP–DINO RDM agreement. (B) Total CLIP-filtered detections retained as threshold varies. The operating point used in main analyses ($t = 0.27$) is marked for reference. Geometry metrics remain comparatively stable across .26–.28, while retained detections decrease monotonically as threshold increases.