Quantifying infants' everyday experiences with objects in a large corpus of egocentric videos

Jane Yang (j7yang@ucsd.edu)
Department of Psychology, University of California, San Diego
La Jolla, CA 92093 USA

Tarun Sepuri (tsepuri@ucsd.edu)
Department of Psychology, University of California, San Diego
La Jolla, CA 92093 USA

Alvin Tan (tanawm@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305 USA

Michael C. Frank (mcfrank@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305 USA

Bria Long (brlong@ucsd.edu)
Department of Psychology, University of California, San Diego
La Jolla, CA 92093 USA

Abstract

While modern vision-language models are typically trained on millions of curated photographs, infants learn visual categories and the words that refer to them from very different training data. Here, we investigate which objects infants actually encounter in their everyday environments, and how often they encounter them. We use a large corpus of egocentric videos taken from the infant perspective (*N* = 868 hours, *N* = 31 participants), applying and validating a recent object detection model (YOLOE) to detect a set of categories that are frequently named in children's early vocabulary. We find that infants' visual experience is dominated by a small set of objects, with differences in individual children's home environments driving variability. We also find that young children tend to learn words earlier for more frequently encountered categories. These results suggest that visual experience scaffolds young children's early category and language learning and highlight that ecologically valid computational models of category learning must be able to accommodate skewed input distributions.

Keywords: early word learning; head-mounted cameras; object detection; naturalistic observations

Introduction

Children rapidly become remarkably adept at categorizing objects, producing words for many categories in their second year of life (Frank et al., 2017). However, modern machine learning models that can similarly categorize objects are trained on millions of curated images (Krizhevsky et al., 2012; Dosovitskiy et al., 2020; Radford et al., 2021; Frank, 2023). Here, we investigate children's "training data" for category learning by examining the distribution and diversity of objects that infants see during everyday experiences.

Head-mounted cameras have provided a unique window into the infant's point of view during early learning (Clerkin et al., 2017; Clerkin & Smith, 2022; Aslin, 2009; Sullivan et al., 2021; Yu & Smith, 2012; Franchak et al., 2011). Yet relatively little work has examined the statistics of the object categories in the infant view, due to the twin challenges of obtaining large amounts of head-mounted camera videos from young children and providing accurate object annotations for these videos.

One exception is Clerkin et al. (2017), who analyzed 8.5 hours of recordings from 8½ to 10½-month-old infants. Individual scenes were cluttered with many objects, but the frequency distribution of object categories across scenes was extremely right-skewed, with a small set of objects accounting for nearly a third of all object instances. While some object categories were pervasively present, they were named rarely (Clerkin & Smith, 2022). Yet these high-frequency objects (e.g., "cup") had names that tend to be learned early in development, suggesting that the frequency of visual categories themselves scaffolds early word learning. However, even this study focused on hand annotations for a small sample of infants (*N* = 8) and a single context (mealtimes). Further, the cameras in this and other prior studies have had relatively limited fields of view (Sullivan et al., 2021; Bergelson & Aslin, 2017)—especially in the vertical direction—limiting the ability to capture the objects infants were interacting with.

Recent advances in data quality and quantity (Long et al., 2023) as well as parallel improvements in object detection models (Wang et al., 2025) allow us to overcome these limitations, giving an unprecedented view on what objects children see and interact with during everyday learning. Here, we analyze the BabyView dataset, a growing set of recordings using a higher-resolution camera with a much larger vertical field of view (see Figure 1a). With these data, we (1) examine the frequency distribution of object categories, (2) patterns of variation across individual subjects and developmental age, and (3) their relationship to the age at which their corresponding object labels are learned.

Methods

Dataset

We analyze naturalistic, at-home egocentric video data collected with the BabyView camera (Long et al., 2023). We extracted frames at 1 frame per second for a total of 3,163,675 frames from 868 hours of data (release 2025.1), *N* = 31 subjects (age range 5 to 36 months). This sampling rate balanced computational efficiency with comprehensive coverage of infants' visual experiences.

Automatic object detection

The detected categories were restricted to align with nouns in the MacArthur-Bates Communicative Development Inventories (CDI) vocabulary items (Frank et al., 2017)(*N* = 288 words). We chose to focus on these categories in order to directly relate object frequency metrics to word learning outcomes. Age-of-acquisition (AoA) was derived from Wordbank (Frank et al., 2017), which provides a measure of when words typically enter children's productive vocabularies by estimating when 50% of children produce each word. Category detection was performed using the YOLOE-v8-L (Wang et al., 2025) (You Only Look Once Efficient model, see Figure 1a). Detection performance was validated by randomly sampling 100 extracted frames; two authors manually annotated all objects in each frame based on our CDI list. We then compared these ground-truth annotations to model's predictions. Overall, YOLOE detections achieved an F-score = 0.757 (*SD* = .43) across all 139 detected categories (average precision = 0.709, recall = 0.812). We included all detections above YOLOE's default confidence threshold of 0.25 as this yielded the highest F-scores. We excluded the frequent "person" and "picture" detections from subsequent analyses as they had heterogeneous referents. Future work will focus on fine-tuning detection accuracies for infrequently viewed categories. All preprocessed data and code are available at https://osf.io/qta5y/.

Results

We examined the overall distribution of object categories across all subjects. Similar to Clerkin et al. (2017), we found

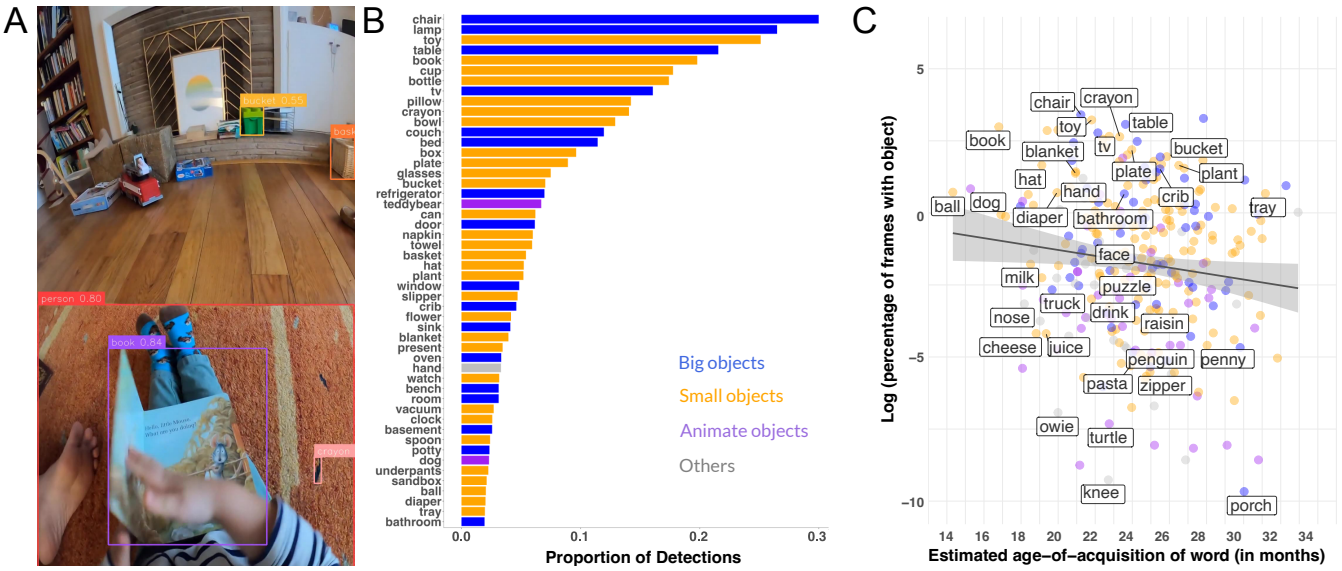

Figure 1: A: An example annotated frame from the infant view by YOLOE; each bounding box indicates a detected object. B: Top 50 object categories detected in the dataset, showing a right-skewed distribution; objects are colored according to their real-world size and animacy (including depictions). C: Log percentage of frames with each object by the estimated age in months at which the corresponding word is produced; line indicates a linear fit with a 95% confidence interval.

that the category distribution was extremely skewed, with a small number of object categories (e.g., *chair, toy*) appearing very frequently, and many others (e.g., *vitamin, penguin*) appearing less frequently in the infant view (see Figure 1b). We analyzed differences in detection frequencies by whether they are small, manipulable objects (*toy, bottle*), large, background objects (*chair, couch*), and objects that either are or refer to animates, as in Konkle & Caramazza (2013), observing skewed distributions for all three domains. These results thus extend and replicate prior work (Clerkin et al., 2017; Long et al., 2021) finding a skewed distribution of object categories in the infant view.

We next examined the consistency and variability in object frequencies across individual subjects and developmental age. We used a generalized mixed-effect model, where we examined whether frequencies were predicted by infants' age in months, including random slopes for the effect of age within each participant and object category. We then examined how variance in object detections was explained by the variation across individual categories and participants. We found that removing random slopes for the effect of age within participants or within categories led to worse model fits (full model vs. model without random slopes, $\chi^2(2) = 175{,}046$, $p < .001$), suggesting that the frequency of different categories in the infant view changes across early development.

Finally, we examined whether there was any relationship between the frequency of these categories in the infant view—collapsed across all participants and age ranges—and the age at which words for these categories tend to be learned in development. Extending Clerkin et al. (2017), words that were more frequent in the infant view tended to be learned earlier in development (see Figure 1C, $b = -0.427$, $SE = 0.194$, $t = -2.207$, $p < .05$). However, visual frequency did not explain unique variance in a linear mixed-effect model predicting age-of-acquisition alongside both word frequency and word concreteness ($b = 0.0128$, $SE = 0.139$, $t = 0.092$, $p = 0.926$). Word frequency and visual frequency are somewhat correlated ($r = 0.185$, $t = 3.180$, $p < 0.01$), suggesting that these metrics capture some overlapping aspects of a child's environment; however, neither measure was correlated with word concreteness (visual frequency vs. concreteness: $r = 0.048$, $t = 0.811$, $p = 0.418$).

Discussion

We leveraged an automated object segmentation model to analyze infants' everyday visual experiences, confirming that a small set of objects appears consistently while most are observed rarely. Despite some limitations in detection accuracy, our approach enabled the analysis of a substantially larger and more diverse dataset than manual coding methods, extending prior results to a much larger set of data than ever before. These detections enabled our replication of the finding with manual annotations that frequently viewed objects tend to have names learned earlier in development, but also found nuance: visual frequency did not explain unique variance once we accounted word frequency, highlighting the need for further analyses. More broadly, these results highlight that in order to understand children's robust visual categorization abilities—and to build ecologically-valid models of early learning—we must grapple with the immense variability in the frequency and diversity with which infants experience object categories and their corresponding labels.

Acknowledgments

This work was supported by an NIH R00HD108386 Pathways to Independence Award to B.L. and a Schmidt Futures gift to M.C.F. We gratefully acknowledge the families who contributed to the dataset, and to the members of the Language and Cognition Lab at Stanford and Visual Learning Lab at UC San Diego for their feedback.

References

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

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.

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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, *44*(3), 677–694.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

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).

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).

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. *arXiv preprint arXiv:2503.07465*.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.