# Examining the precision of infants' visual concepts by leveraging vision-language models and automated gaze coding

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

**Martin Zettersten (mzettersten@ucsd.edu)**
Department of Cognitive Science, University of California, San Diego
La Jolla, CA 92093 USA

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

## Abstract

**Infants rapidly develop knowledge about the meanings of words in the first few years of life. Previous work has examined this word knowledge by measuring how much infants look at a named target image over a distractor. Here, we examine the specificity of that knowledge by manipulating the similarity of the target and distractor. We measured looking behavior in 91 14- to 24-month-old infants, enabled by automatic gaze annotation and online data collection. Using a vision-language model to quantify target-distractor image and text similarity, we find that infants' looking behavior is shaped by the high-level visual similarity of competitors: infants' looking to the target image was inversely correlated with image similarity but not with visual saliency. Our findings demonstrate how multimodal models can be used to systematically examine the content of infants' early visual representations.**

**Keywords:** early word learning; vision-language models; partial knowledge; visual concepts; early representations

## Introduction

How precise are the visual concepts that support the robust advances in early word learning (Bergelson, 2020)? Parent-reported data suggest that the average child can understand around 300 words by 24 months (Frank et al., 2021). Yet, to fully understand the visual meaning of words, children need to know how to correctly generalize them: for example, to learn what the word "bulldozer" means, children need to learn what bulldozers look like—and don't look like—arguably a computationally challenging feat (Vong et al., 2024). Accordingly, some theoretical accounts have characterized visual concept learning as a slow and incremental process (Swingley, 2010; Wagner et al., 2013), positing that children's representations begin rather coarse (Rescorla, 1980; Long et al., 2024). In other words, infants may have partial visual knowledge of many of the words they ostensibly learn rapidly.

The looking-while-listening paradigm has provided initial evidence for early partial word knowledge but methodological barriers have obviated any strong conclusions. In this paradigm, infants are shown two images on a screen and are asked to "find the [bulldozer]" (Fernald et al., 2008). Then, the nature of visual concept knowledge can be characterized by how the similarity of a distractor interferes with word recognition (Arias-Trejo & Plunkett, 2010). For example, when asked to look at a stroller, 6-month-old infants' attention is more likely to be drawn to a car, a distractor with high semantic similarity, than a hand, a less similar distractor (Bergelson & Aslin, 2017). However, to date, all distractor interference studies operationalize similarity as a dichotomous variable based on subjective experimenter judgments. These studies may differ in their emphasis on perceptual and conceptual similarities, rendering synthesis across studies difficult. Additionally, collecting infant data often requires large amounts of hand-annotated gaze data. Thus, looking-while-listening studies typically consist of small sample sizes (averaging *N*=25; Bergmann et al., 2018) and small stimuli sets with relatively easy words. Together with idiosyncratic similarity measures, this leads to a low-data regime which makes it challenging to assess theoretical predictions about the nature of children's early word knowledge.

Here, we expand on prior work by systematically examining the precision of visual concepts in a large sample of infants and across a broad range of items. To do so, we examine infants' understanding of how words refer to naturalistic images taken from a dataset of visual concepts (Stoinski et al., 2024). We then use a multimodal transformer model (Radford et al., 2021)—with both vision and language encoders—to examine how the image and text similarity metrics of these stimuli influence infants' looking behaviors (Tan et al., 2024). Based on previous findings (Bergelson & Aslin, 2017), we hypothesize that the more similar a target and distractor are, the more infants will be drawn away from the target. We introduce a novel, automated pipeline for implementing this looking-while-listening task by utilizing the Children Helping Science online platform (Scott & Schulz, 2017) and by automating gaze coding (Erel et al., 2022; Raz et al., 2024).

## Methods

**Participants** We tested 91 children between the ages of 14-24 months (*M*=19.75 months) on the online research platform Children Helping Science (Scott & Schulz, 2017). Data from 23 additional children were excluded due to insufficient looking (less than 50% looking to the screen) on more than 50% of trials. Our pre-registered sample size of 90 participants provided 90% power to detect an estimated effect size of *d*=0.35.

**Procedure** Each caregiver led their infant through our study. The study was asynchronous, meaning that no experimenter was present. Infants were shown 32 trials, each lasting 7200ms, interspersed with 4 attention-getters, consisting of a pair of images on the left and right of the screen and an audio stimulus labeling one of the two images.

**Stimuli** We operationalized similarity for stimuli selection by taking the cosine similarity between text embeddings from CLIP (Radford et al., 2021). Eight target words, with one similar and one dissimilar distractor each, were chosen from the THINGS+ dataset (Stoinski et al., 2024; Figure 1A). These stimuli formed the basis of our 32 trial design: 8 trials each where the target (e.g., *bulldozer*) had a dissimilar (e.g., *orange*) and a similar distractor (e.g., *tractor*), and 16 trials where the distractors were themselves the target (e.g., *orange, tractor*). A female native speaker of American English recorded all audio in infant-directed speech. One of four carrier frames was chosen for each trial (e.g., "Look at the [target]"). Each auditory stimulus was normalized in amplitude and duration.

**Analysis Plan** Looking time data was coded using iCatcher+, which predicts whether a child is looking left, right, or away in a frame by using a face detector and a gaze

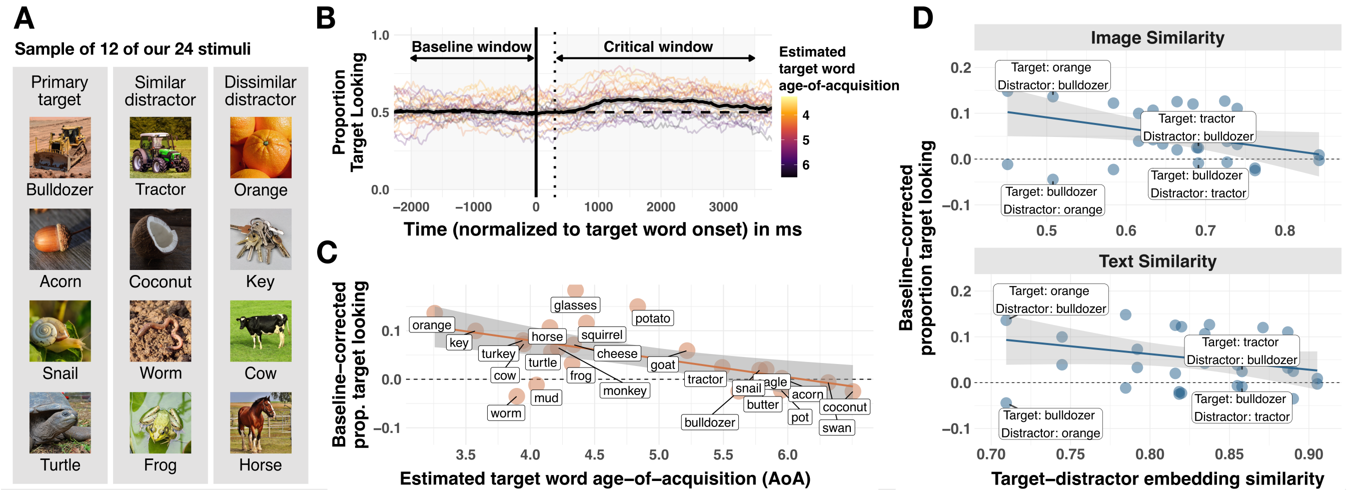

Figure 1: (A) Breakdown of items employed in the experiment design, (B) Average timecourse of proportion target looking for each item. Black line indicates the average across all items with a 95% CI error band, (C) Looking behavior by target word difficulty, regression line indicates a linear fit with a 95% CI error band (D) Looking behavior by item-pair similarity, regression lines indicate a linear fit with 95% CI error bands.

direction classifier (Erel et al., 2022). The tool's performance has been validated as comparable to that of human coders (Erel et al., 2023; Luchkina et al., 2024). Word recognition accuracy was then assessed at the trial level using the proportion of looking to the target vs distractor; specifically, we analyzed the difference between looking in the critical (300-3500 ms post label onset) vs the baseline window (-2000-0 ms relative to label onset; Weaver et al., 2024) .

We then predicted infants' proportion target looking, in two separate linear mixed-effects models, using the cosine similarity of target-distractor image embeddings and text embeddings from CLIP (using a ViT-B/32 vision encoder; Radford et al., 2021). The models included an interaction between infant age and similarity, a random slope for similarity by-subject, and a random intercept for the target image. In a secondary analysis, we incorporated two predictors, expecting them to influence looking behavior: (1) word difficulty, estimated using the age-of-acquisition (AoA) ratings from Kuperman et al. (2012) and (2) visual saliency differences, which we measured using the GBVS toolbox (Harel et al., 2006; measured as the difference between the mean visual saliency of the target and distractor). All experiment code, stimuli, and analyses are available at the OSF repository `https://osf.io/925t6/`.

## Results

We examined whether infants showed evidence of graded visual concept knowledge. To do so, we analyzed how well infants could identify the referents of visual concepts in naturalistic images, using distractors that varied in similarity to target images. Infants looked more at the target image when the distractor was more dissimilar in image similarity space (see Figure 1D; estimated with a linear mixed-effects model: $b$=−0.06, $SE$=0.02, $p$<.05). The effect of our text similarity measure—which was largely colinear with image similarity ($r$=0.77, $p$<.001)—trended in the same direction but was not statistically significant ($b$=−0.05, $SE$=0.03, $p$=.08). As expected, older infants looked more at target images in general ($b$=0.06, $SE$=0.02, $p$<.05); we did not find any interaction between infant age and our similarity measures ($b$=−0.03, $SE$=0.02, $p$=.12).

In addition, we anticipated that more difficult words would be more challenging for infants to recognize. Consistent with this prediction, the AoA of the target word correlated inversely with the proportion of time infants looked at the target image (Figure 1C; $b$=−0.09, $SE$=0.03, $p$<.01). Yet, while target word AoA and target-distractor image similarity were not colinear ($r$=0.26, $p$=.16), they did not explain unique variance in this sample. Finally, we verified that these effects were not driven by differences in how well the stimuli captured infants' attention: target-distractor visual saliency differences from a GBVS model did not predict variance in infants' looking behaviors ($b$=0.01, $SE$=0.03, $p$=.89).

## Discussion

These results suggest that in their second year, infants have partial visual knowledge for many difficult words. When the distractor and target images had more similar high-level visual features, infants had more trouble identifying the correct visual referent and were often unable to identify it at all. More broadly, this work combines advances in vision-language models and gaze annotation techniques to examine the precision of infants' visual concept knowledge. Building an automated pipeline enabled the use of a large sample and a diverse item set, which in turn helped us examine effects for individual items, though our conclusions remain limited by our item set and age range. Overall, this framework paves the way to further investigate the progression of visual concept knowledge in early development.

## Acknowledgments

We gratefully acknowledge the families who participated in our study. This work was supported by an NIH Pathways to Independence Award (R00HD108386) to B.L.. We also thank the members of the Visual Learning Lab and the Language & Infant Learning Lab at UC San Diego for their valuable feedback on this work.

## References

Arias-Trejo, N., & Plunkett, K. (2010). The effects of perceptual similarity and category membership on early word-referent identification. *Journal of Experimental Child Psychology*, *105*(1), 63–80. doi: 10.1016/j.jecp.2009.10.002

Bergelson, E. (2020). The Comprehension Boost in Early Word Learning: Older Infants Are Better Learners. *Child Development Perspectives*, *14*(3), 142–149. doi: 10.1111/cdep.12373

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. doi: 10.1073/pnas.1712966114

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting Replicability in Developmental Research Through Meta-analyses: Insights From Language Acquisition Research. *Child Development*, *89*(6), 1996–2009. doi: 10.1111/cdev.13079

Erel, Y., Potter, C. E., Jaffe-Dax, S., Lew-Williams, C., & Bermano, A. H. (2022). iCatcher: A neural network approach for automated coding of young children's eye movements. , *27*(4), 765–779. doi: 10.1111/infa.12468

Erel, Y., Shannon, K. A., Chu, J., Scott, K., Kline Struhl, M., Cao, P., ... Liu, S. (2023). iCatcher+: Robust and Automated Annotation of Infants' and Young Children's Gaze Behavior From Videos Collected in Laboratory, Field, and Online Studies. *Advances in Methods and Practices in Psychological Science*, *6*(2), 25152459221147250. doi: 10.1177/25152459221147250

Fernald, A. E., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. M. Fernández, & H. Clahsen (Eds.), *Developmental Psycholinguistics: Online methods in children's language processing* (pp. 97–135). John Benjamins Publishing Company. doi: 10.1075/lald.44.06fer

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project.* MIT Press.

Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. In *Advances in Neural Information Processing Systems* (Vol. 19). MIT Press.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. doi: 10.3758/s13428-012-0210-4

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191. doi: 10.1038/s41467-023-44529-9

Luchkina, E., Simon, L. R., & Waxman, S. (2024). *Catching up with iCatcher: Comparing analyses of infant eyetracking based on trained human coders and automated gaze coding software.* doi: 10.31219/osf.io/2aq9z

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763).

Raz, G., Cao, A., Saxe, R., & Frank, M. C. (2024). *A stimulus-computable rational model of habituation in infants and adults.* doi: 10.1101/2024.08.21.609039

Rescorla, L. A. (1980). Overextension in early language development. *Journal of Child Language*, *7*(2), 321–335. doi: 10.1017/S0305000900002658

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A New Online Platform for Developmental Research. *Open Mind*, *1*(1), 4–14. doi: 10.1162/OPMI_a_00002

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. doi: 10.3758/s13428-023-02110-8

Swingley, D. (2010). Fast Mapping and Slow Mapping in Children's Word Learning. *Language Learning and Development*, *6*(3), 179–183. doi: 10.1080/15475441.2010.484412

Tan, A. W. M., Yu, S., Long, B., Ma, W. A., Murray, T., Silverman, R. D., ... Frank, M. C. (2024). *DevBench: A multimodal developmental benchmark for language learning* (No. arXiv:2406.10215). doi: 10.48550/arXiv.2406.10215

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. , *383*(6682), 504–511. doi: 10.1126/science.adi1374

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. , *127*(3), 307–317. doi: 10.1016/j.cognition.2013.01.010

Weaver, H., Zettersten, M., & Saffran, J. R. (2024). Becoming word meaning experts: Infants' processing of familiar words in the context of typical and atypical exemplars. *Child Development*, *95*(5), e352-e372. doi: 10.1111/cdev.14120