Developmental changes in the precision of visual concept knowledge

Bria Long[1], Wanjing Anya Ma[2], Alvin W. M. Tan[2], Rebecca Silverman[2], Jason D. Yeatmean[2], & Michael C. Frank[2]

[1] University of California San Diego

[2] Stanford University

Author Note

.

Correspondence concerning this article should be addressed to Bria Long. E-mail: brlong@ucsd.edu

Abstract

How precise is children's visual concept knowledge, and how does this change across development? We created a gamified picture-matching task where children heard a word (e.g., "swordfish") and had to choose the picture "that goes with the word." We collected data from large sample of children on this task (N = 3467, 3-14 years of age) and adults (N = 211), and we modeled changes in the proportion of children who chose a given image for a certain word over this developmental age range. We found gradual changes across this age range in children's ability to identify the correct category, highlighting a protracted developmental trajectory. Error analysis revealed that children were more likely to choose higher-similarity distractors as they grew older; further, children's error patterns were increasingly correlated with target-distractor similarity in the linguistic and multimodal embedding spaces of a large multimodal language model. These analyses suggest a transition from coarse to finer-grained visual representations over early and middle childhood, while emphasizing that even young children have partial knowledge for many difficult visual concepts. More broadly, these findings demonstrate the utility of combining gamified experiments and similarity estimates from computational models to probe the content of children's evolving visual representations.

Keywords: visual concepts, receptive vocabulary, large language models, object recognition

Word count: X

Developmental changes in the precision of visual concept knowledge

Introduction

When a child hears a word — like a "whale" — this activates a mental representation of its referent in the visual world. Depending on how old a child is – and how much they know about whales – a child might imagine a canonical exemplar of a blue whale, a specific whale from a picture book, or perhaps just vaguely an animal that lives in the ocean. How precise are the visual representations that underlie children's understandings of words across early and middle childhood?

Early in development, children experience an astonishing rate of vocabulary growth as they begin to communicate with their caregivers about the objects, people, and places around them (Bloom, 2000; Braginsky, Yurovsky, Marchman, & Frank, 2019). Infants as young as 6-months of age associate some shape information with common words (Bergelson et al., 2009), and 14-18 month-olds extend newly learned words to atypical exemplars of these categories in looking-while-listening tasks (Weaver et al., 2024). By around their second birthday, children also extend nouns to stylized, 3D exemplars (Smith, 2003) as they learn that shape is a valuable cue to basic-level category membership (Rosch et al., 1976). Thus, at least for within-category exemplars, very young children exhibit relatively sophisticated generalization abilities for common visual concepts, in line with a broad-to-narrow view of category development (Waxman & Gelman, 2009), where infants construe words as initially referring to many items and subsequently refine their representations across development.

From this perspective, children's visual representations may change relatively little beyond these first early years; instead, children may continue to gradually acquire new visual concepts and then change in how they represent the relationships between categories. For example, children may learn that whales are mammals, and then appropriately group them with other land mammals vs. with fish when asked to make taxonomic classifications

(Vales, Stevens, & Fisher, 2020). Accordingly, empirical work on children's developing ability to recognize objects (Ayzenberg & Behrmann, 2024) has also focused on the first few years of childhood as the most critical period in which object recognition abilities develop.

To overcome these methodological barriers, we created a gamified picture-matching task where children heard a word (e.g., "swordfish") and had to choose the picture "that goes with the word". Critically, we chose distractor items with high, medium, and low concept similarity to each target word; distractors were paired via cosine similarity of the target and distractor words in a large multimodal language model (CLIP, Radford et al., 2021). This task was then deployed in online, preschool, and school contexts to 3599 children aged 3-15 years and 211 adults years of age. Using this large dataset, we find gradual changes in how children represent visual concepts across childhood, with older children becoming both more accurate at identifying the correct referents throughout this extended age range; however, we also found that even young children were more likely to choose the related vs. unrelated distractors, highlighting a gradual change from coarse, representations that encompass both the target and related distractors to fine-grained, specific representations that the visual information that words refer to. We then use both unimodal and multimodal embeddings from this same model to examine how visual, linguistic, and multimodal similarity explain children's error patterns across development.

Contrary to this simplified account, here we provide evidence that children's visual concepts continue to change throughout childhood, with an extended developmental trajectory that continues in parallel with later vocabulary learning and formal schooling. Children's vocabulary knowledge – often assessed via paper-and-pencil, closed, expensive traditional assessments – grows and expands across childhood, but there has been relatively little consideration of the visual representations that support children's performance on picture vocabulary tasks. Some work on children's production and recognition of line drawings of common objects hints at this kind of protracted developmental timeline (Long et al., 2024): in a large observational study, children became increasingly able to both depict and recognize line drawings of common object categories. However, no work has directly tested children's visual recognition behaviors for a wide variety of visual concepts, in part because of the difficulty of obtaining data from large samples of children on a consistent set of items with variability over a large developmental age range.

To overcome these methodological barriers, we created a gamified picture-matching task where children heard a word (e.g., "swordfish") and had to choose the picture "that goes with the word". Critically, we chose distractor items with high, medium, and low concept similarity to each target word; distractors were paired via cosine similarity of the target and distractor words in the language encoder of a large multimodal language model (Contrastive Language-Image Pre-training model, or CLIP, Radford et al., 2021) (see overview in Figure 1a). This task was then deployed in online, preschool, and school contexts to 3467 participants aged 3-14 years and 211 adults. Using this large dataset, we found gradual changes in how children represent visual concepts across childhood, with older children becoming both more accurate at identifying the correct referents throughout this extended age range. We also found that even young children were more likely to choose the related vs. unrelated distractors, highlighting a gradual change from coarse, representations that encompass both the target and related distractors to fine-grained, specific representations that the visual information that words refer to. We then use both unimodal and multimodal embeddings from this same model to examine how visual, linguistic, and multimodal similarity explain changes in children's error patterns across development.

Methods

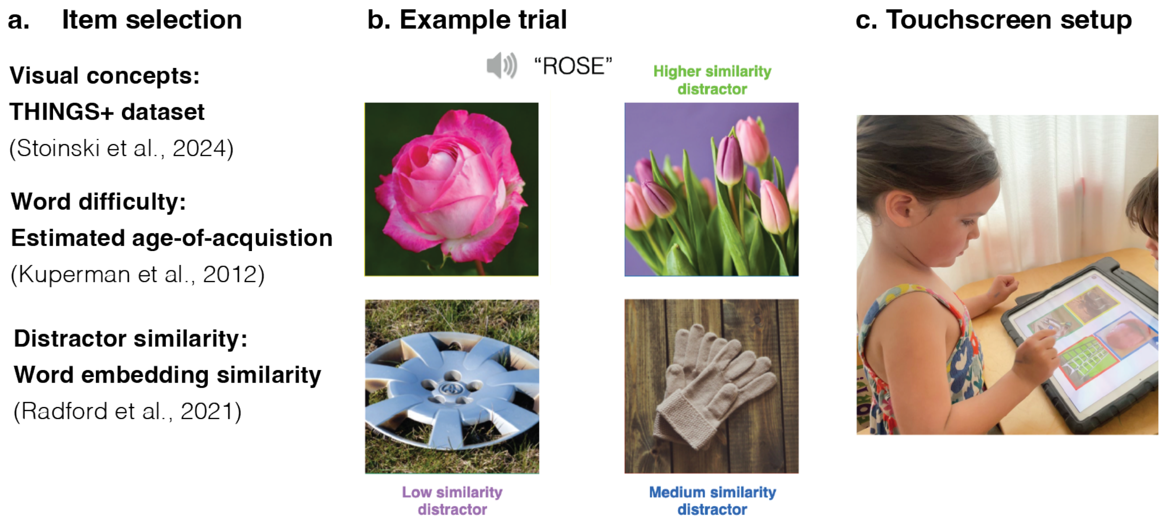

Figure 1. Overview of the (a) databases and models involved in item selection, (b) an example trial, and (c) a touchscreen setup for younger participants.

Procedure

Children were invited to participate in a picture matching game where children were asked to help "teach aliens on another planet about some of the words on our planet" and children picked a particular alien to "accompany them on their journey." Before the stimuli appeared, children heard a target word (e.g., "apple") and then were asked to "choose the picture that goes with the word". The four images appeared in randomized locations on the screen, and one of the images always corresponded to the target word (see example trial in Figure 1b). On practice trials, the distractor images were all very dissimilar to the target concept, and the target word was relatively easy. The tablet played a chime sound if children chose the correct image, and a slightly unpleasant sound if they responded incorrectly. Each child viewed a random subset of the item bank, and the items they viewed were displayed in a random order. Children were allowed to stop the game whenever they wanted to. Different versions of the game included varying amounts of trials or items;

these games were developed as part of a project to develop an open-sourced measure of children's vocabulary knowledge. Here, we analyze children's responses to items that were generated using the THINGS+ dataset with distractors of varying difficulty (see Stimuli).

Participants

To obtain a large sample, we collected data from children in several different testing contexts. We collected data from children in an in-person preschools (N = 65, 3-5 year-olds), from the Children Helping Science Platform, (N=243, 3-7 year-olds), elementary schools across multiple states (N=3332, 5-14 year-olds) and adults online (N=211 adults, recruited via Prolific; half of the adults spoke English as a second language). Most participants responded directly via a keyboard, except those recruited online: however, children's parents responded via clicking on the image on Children Helping Science, and adults responded via clicking on the images.

We included data for a total of 3786 participants from preschools, schools, and online testing contexts around the United States (range 84 to 654), who completed, on average, 25.02 4AFC trials that were sampled randomly from the stimuli set (max = 86; different maximum numbers of trials were included in different testing contexts). All participants who contributed data and scored above 30% accuracy were included, even if they did not complete the assessment (minimum trials = `min(trials_by_participant$num_trials)`, maximum trials = 86, average number of trials = 25.02, We tested an additional 84 participants who scored near chance on 4AFC trials (chance = 25%, threshold = 30%) and were school-aged (>6 years of age) and who we excluded from analyses; these participants completed an average of 17.72 trials.

Stimuli selection

We capitalized on publicly available existing image and audio databases to generate stimuli. Visual concepts were taken from the THINGS+ dataset (Stoinski et all., 2023), after filtering out non-child safe images (e.g., weapons, cigarettes) and images with low nameability (<.3), as per the released norming data. We used the copy-right free, high-quality image released for each visual concept. We then subset to visual concepts that had available audio recordings in the MALD database as well as age-of-acquisition (AoA) ratings from a previous existing dataset (Kuperman, 2012).

Using this subset, we sampled distractors with high, medium, and low similarity to the target word as operationalized via embedding similarity of the words in the language encode of a multimodal large language model (Radford et al., 2021).We determined high-, medium, and low similarity values relative to the distribution of all possible target-distractor pairing values for each word in the THINGS+ dataset. Stimuli were selected to optimize for having a maximum number of trials with unique target and distractors, in addition, we constrained the sampling such that target-distractor pairs had estimated age of acquisition within 3 years of each other. All stimuli and their meta-data are available on the public repository for this project. For each target word, we first selected a high-similarity distractor that had the highest cosine similarity to the target (and was itself not one of the target words). For medium-similarity distractors, we randomly sampled a distractor word was the same animacy as the target word, and unique to the dataset. For low-similarity words, we sampled a unique distractor words that had the lowest cosine similarity among the remaining distractors. In our final set, we had 108 items with a range of different estimated age-of-acquisitions (e.g., hedgehog, mandolin, mulch, swordfish, waterwheel, bobsled) with all unique targets and distractors. See Appendix, Figure XX for a visualization of the cosine similarity values for each distractor type for each word.

Model features

We obtained all model features features using the Open AI available implementation of CLIP available at https://github.com/openai/CLIP. For language similarity, we computed the cosine similarity of the embeddings of the target word to each distractor word on each trial (e.g,. rose – tulip, rose – glove, rose – hubcap). For visual similarity, we repeated this procedure but by obtaining image similarity vectors in the vision transformer for each target image and distractor image on each trial. For multimodal similarity, we computed the cosine similarity of the embedding of the target word in the language model to the embeddings for each of the distractor images; this is possible because the embedding spaces for the vision and language transformers in the CLIP model are aligned and have the same number of dimensions.

Results

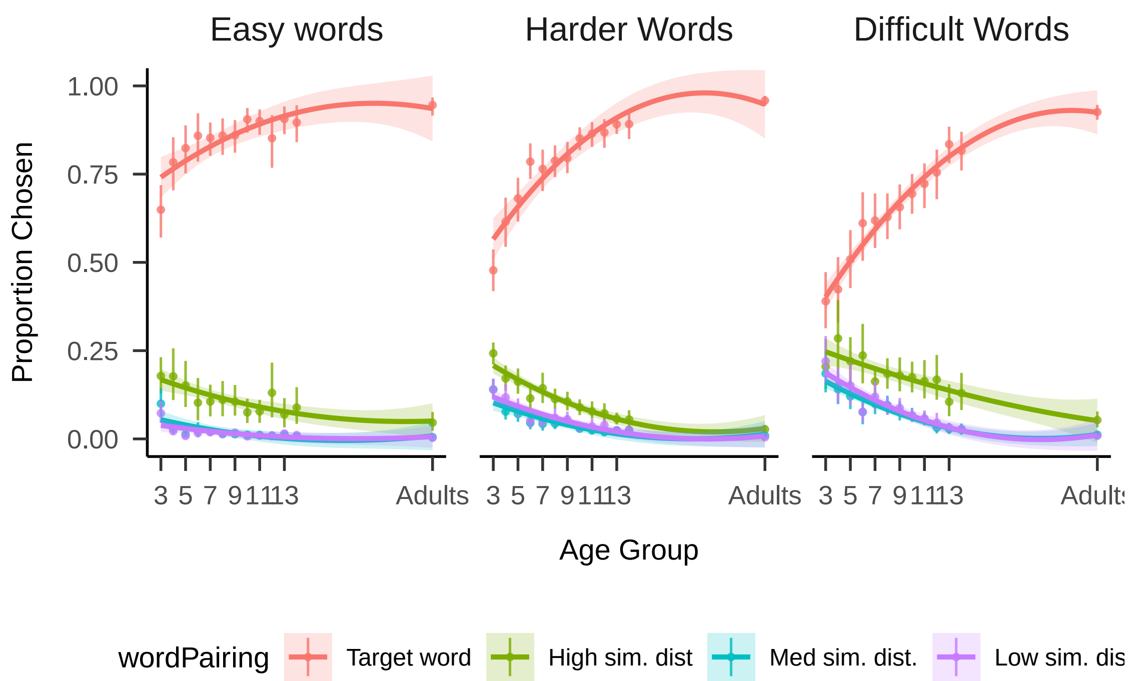

Figure 2. Visual vocabulary task performance as a function of the age of the child completing the task, plotted separately for relatively easy, harder, or difficult words; words are binned into terciles based on the estimated AoA from Kuperman et al., 2012. Lines refer to the proportion of words that children chose the target (red), high-similarity (green), medium similarity (turqouise), or low similarity (purple) distractor at each age; error bars represent boostrapped confidence intervals.

A protracted developmental trajectory

We found a gradual increase in children's ability to correctly identify the target word across our entire age range, extending into early adolescence; Figure 2 shows the proportion of time that children identified the target word, highlighting a protracted developmental trajectory. We found this developmental trend for both relatively "easy" words, with an average estimated age-of-acquisition (AoA) of 4.81 years (SD = 0.87), more difficult words (average AoA =6.95 years, SD = 0.65 years), and challenging words (average AoA =9.60 years, $SD = 1.21$)).

At an item level, the words that showed the greatest change across age included some animals (e.g., "swordfish") as well as inanimate objects ("prism","antenna","sandbag", "turbine") but also parts of larger buildings ("scaffolding","gutter"). However, some of our developmental trends likely also stem from differences in executive control: for some words that had very semantically similar distractors but were relatively easy (e.g., "cheese" vs "butter"), we still saw steep developmental changes, highlighting that this "simple" picture vocabulary matching tasks still assess many different cognitive abilities beyond the fidelity of children's visual representations.

Changes in the precision of visual concepts

Next, we thus aimed to understand whether we were indeed observing changes in the precision of children's visual concepts. Indeed, one possibility is that children are mostly becoming more accurate at ignoring relevant distractors (due to developmental changes in executive function capacity), but often have knowledge of the target concepts. If this is the case, then we should only observe changes how accurate children are at identifying the target word, with no changes in the types of distractors that children choose when they choose incorrectly. However, if children's visual concepts are proceeding from a coarse representations that is overly broad, or starting from no representation at all, then we would expect younger children to be more likely to choose distractors of all types, whereas we would expect older children to almost exclusively choose related distractors.

We thus examined whether we observed systematic changes in how children made errors across age, shown in Figure 3. Consistent with the latter hypothesis, we found that children increasingly choose related distractors throughout development, with adults being still more likely to choose the related distractors relative to the oldest children (14-year-olds) in our sample.

Table 1

Fixed effect coefficients a linear mixed effects model assessing changes in the proportion of related distractors chosen over development. The model included random intercepts for participants. Age and number of trials were standardized prior to analysis.

effect	Predictor	b	SE	t	df	p
fixed	Intercept	0.62	0.01	122.41	3407.593	$< .001$
fixed	Age (scaled)	0.06	0.01	11.26	3419.562	$< .001$
fixed	Number of trials (scaled)	0.00	0.00	-0.42	3404.545	0.671

We confirmed this result via a linear mixed effect models, modeling the proportion of errors that each children chose related distractors as our dependent variable as a function of children's age (in years); we also included a fixed effect of the number of errors each child made as this varied widely by participant and age group. We found a main effect of age (see Table 1): older children were more likely to choose related distractors relative to unrelated distractors. We also modeled these effects at the item level in a second linear mixed-effect model, with random intercepts for each item, finding a fixed effect of age (see SI), and thus the same pattern of effects. Children become more likely to choose related distractors across development, suggesting a progression where children gradually build detailed knowledge about the visual referents of many challenging words.

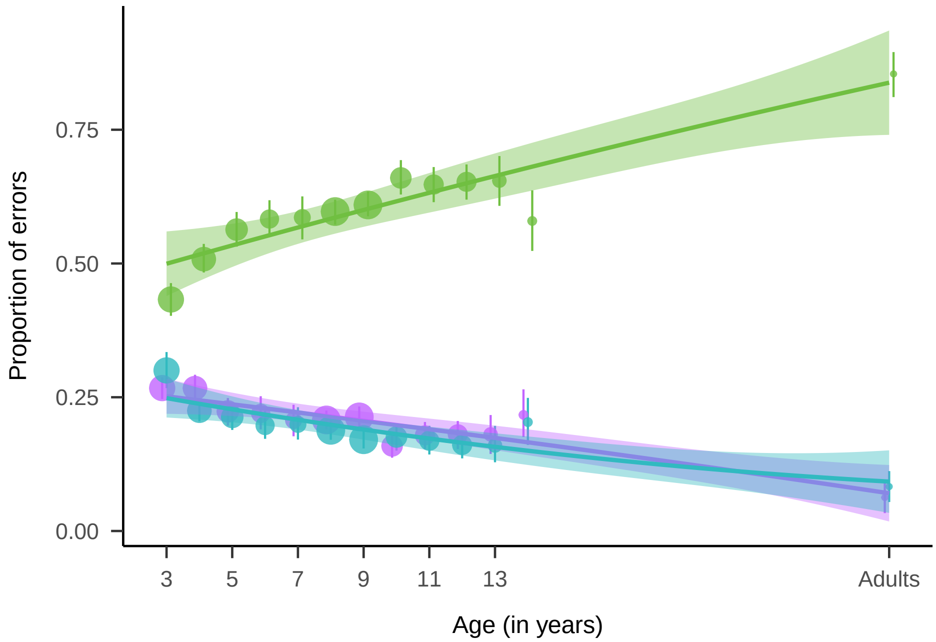

Figure 3. Changes in the proportion of errors chosen as a function of childrens age, where green lines reflected higher similarity distractors. Dot size represents the number of errors made by children in each age group. Error bars represent 95 percent bootrstrapped confidence intervals.

Modeling changes in children's error patterns

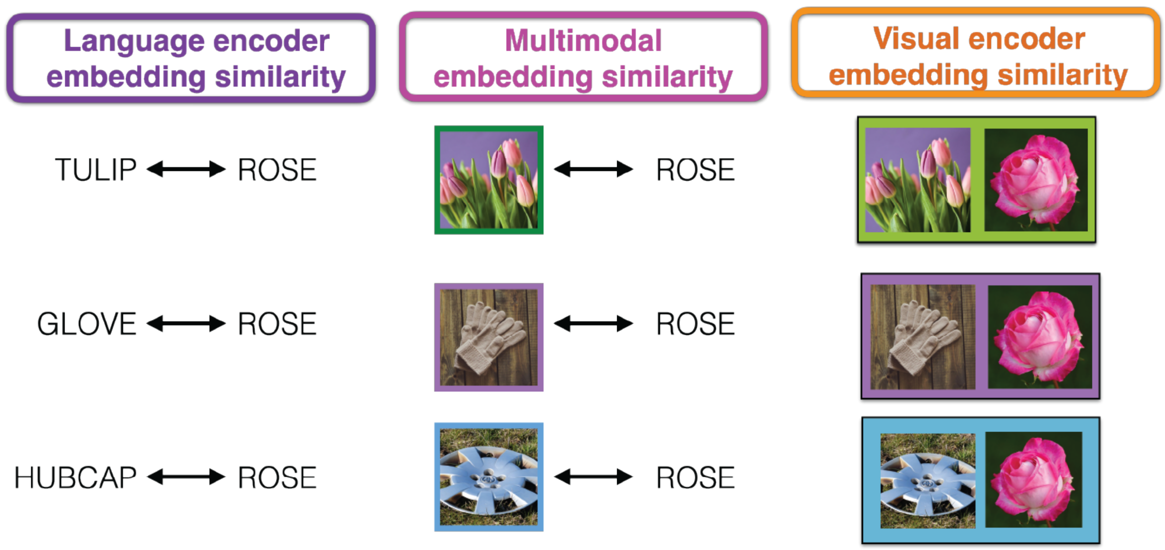

Figure 4. Schematic of the three different ways that embedding simialrity was calculated in CLIP (Radford et al., 2021)

In a set of final analyses, we aimed to understand the source of these changes in children's error patterns by leveraging the high-dimensional embeddings of our linguistic and visual stimuli in the same large, multi-modal language model (Radford et al., 2021), acknowledging that our stimuli were not necessarily designed to pull apart the contributions of changes in semantic versus visual similarity. Nonetheless, our stimuli were generated by using similarity in a linguistic embedding space, and so some stimuli on certain trials were nonetheless related to the target concept semantically but not necessarily visually (e.g, gardening "gloves" were a distractor for the target word "rose"). We thus sought to understand the degree to which children's error patterns in this task reflected changes in how they processed the visual similarity of the targets and distractors, their semantic similarity, or–perhaps most likely–some combination.

To do so, we used a series of cross-validated linear mixed effect models, where we examined the degree to which visual, linguistic, and multimodal similarity metrics (and their combination) derived by large language model embeddings could explain children's error patterns. Specifically, we modeled the proportion of time that children chose each distractor for a target word as a function of the difficulty of the target word (as estimated by the estimated AoA metric), the age (in years) of the children participating, and (1) the similarity of the target word to each distractor word (linguistic embeddings), (2) the similarity of the target image to each distractor image (visual embeddings), and (3) the similarity of the target word to each distractor image (multi-modal embeddings), and (4) a combined model with all predictors combined. We iteratively sampled 80/% of the dataset 50 times, and then evaluated the conditional R-squared for each model for each split; these values are plotted in Figure 5. These exploratory results revealed that combining both the visual and linguistic embeddings – either in one, large mixed-effect model, or via multi-modal embeddings – led to increase explained variance in children's error patters. These results thus suggest that changes in children's error patterns across age are not solely due to changes in children's ability to reject the distractor images that are visually similar to the target concept.

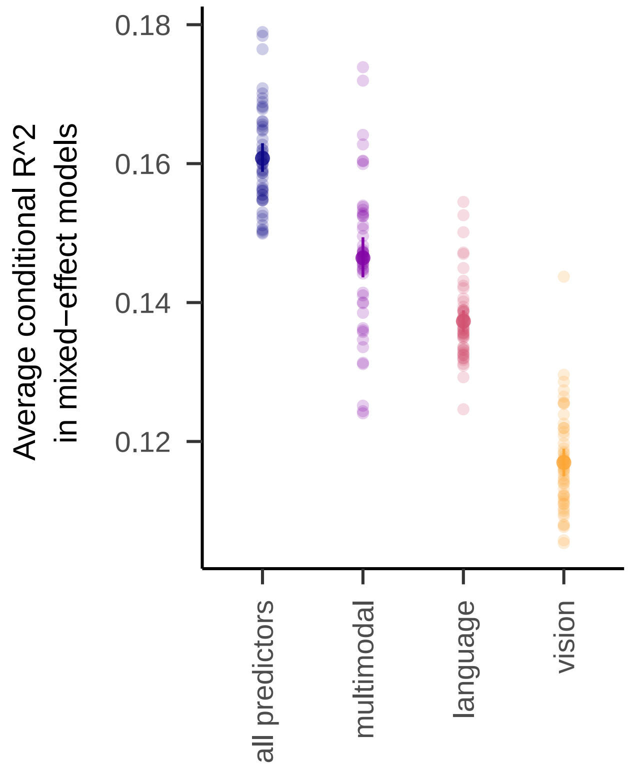

Figure 5. Average explained variance in children's error patterns (conditional R-squared in linear mixed effect models) by linguistic, visual, multimodal, or combined predictors in cross-validated mixed efffect models. Error bars represent bootstrapped 95 percent confidence intervals across 50 iterations.

In a final analyses, we aimed to understand the source of these changes in children's error patterns by leveraging high-dimensional embeddings of our linguistic and visual stimuli in a large, multimodal language model (CLIP, Radford et al., 2021). We chose a set of stimuli where visual similarity was colinear with semantic similarity to a large degree, as it often is in the real-world for most visaul concepts. Thus, our stimuli were not necessarily designed to pull apart the contributions of changes in semantic vs. visual similarity. Nonetheless, our stimuli were generated by using similarity in a linguistic embedding space, and so some stimuli on certain trials were nonetheless related to the target concept semantically but not necessarily visually (e.g, gardening gloves were a distractor for "rose"). We thus sought to understand the degree to which children's error patterns in this task reflected changes in how they processed the visual similarity of the targets and distractors, their semantic similarity, or–perhaps most likely–some combination.

To do so, we used a series of cross-validated linear mixed effect models, where we examined the degree to which visual, linguistic, and multimodal similarity metrics (and their combination) derived by large language model embeddings could explain children's error patterns. Specifically, we modeled children's proportion of time that children chose each distractor for a target word as a function of the difficulty of the target word (as estimated by the estimated AoA metric), the age (in years) of the children participating, and (1) the similarity of the target word to each distractor word (linguistic embeddings), (2) the similarity of the target image to each distractor image (visual embeddings), and (3) the similarity of the target word to each distractor image (multi-modal embeddings), and (4) a combined model with all predictors combined. We iteratively sampled 80/% of the dataset, and then evaluated the conditional r-squared for each model for each split; these values are plotted in Figure 5. These exploratory results revealed that combining both the visual and linguistic embeddings – either in one, large mixed-effect model, or via multi-modal embeddings – led to increase explained variance in children's error patters. These results thus suggest that changes in children's error patterns across age are not solely due to changes in children's ability to reject the distractor images that are visually similar to the target concept.

Discussion

How precise is children's visual concept knowledge, and how does this change across development? Here, we collect and analyze a large dataset of picture matching performance across development, finding evidence for a transition from coarse to finer-grained visual representations over early and middle childhood. Children became more accurate at identifying the referents of words over this entire age range, and their error patterns progressed from relatively random towards related distractors.

Broadly, these data support a theoretical view where these is substantial enrichment and change in existing representations for everyday visual concepts throughout childhood. For example, certain visual features may become more or less salient in children's visual concepts as children understand their functional roles (e.g., camels have humps to store water) or the degree to which they help delineate a category boundary (e.g., between whales and whale sharks). On this account, even school-aged children's visual representations may undergo substantial change as they learn more about the world around them.

This protraction of the timeline for visual concept learning into middle childhood substantially broadens the scope of potential learning mechanisms beyond associative label-object matching. For example, children's learning environments extend beyond the home and into structured educational contexts; children's learning partners include their peers, teachers, and siblings (who may be more or less reliable), and children's individualized experiences, interests, and hobbies may influence which words they have detailed visual representations for. As children begin to learn why animals and objects are classified the way they are, this semantic learning likely influences the visual features that are prioritized in children's representations. To make matters even more complicated, children also likely learn about visual features from generic utterances (e.g., "Tigers have stripes") in verbal conversations where visual referents are nowhere to be found. Thus in order for our models and theories of visual concept learning to account for this full developmental trajectory, we need to think beyond labelled (or even captioned) photos or videos of referents.

Indeed, we suspect that visual concept learning extends into adulthood, and that many adults have coarse visual representations for many different words (and indeed, we culled some items during pilot testing because adults could not discriminate them!). Consider that while many adults in Western contexts experience the referents of some visual concepts relatively frequently e.g., trees, computers, cups, cars – other words refer to visual concepts that different individuals may have varying amounts of interest in and frequency in interacting with – like telescopes, or antelopes. Visual concept learning is likely influenced by both individuals pre-occupations and very intense interests, be they professional or not. And indeed decades of work has established that birding experts, car aficionados, and graphic artists have both qualitatively and quantitatively different kinds of visual representations for the visual concepts that they engage with (CITE, CITE).

There are several limitations to the current work that future work could address. While we include data from a diverse group of children over a wide developmental age range, at present our conclusions are drawn primarily from around one hundred experimental items and distractors; further work that expands the range and diversity of the visual concepts – and that expands to populations outside of the continental U.S. (Henrich et al., 2010) will be necessary to understand the generalizability of these findings. In addition, the present data are large but cross-sectional, and thus cannot provide evidence for changes in the precision of representations within individual minds. Dense data collected from children over longer ranges of developmental time could confirm the hypotheses and theories raised by these analyses. Nonetheless, the present work highlights the promise of large-scale, online games for collecting large datasets that can be used to examine the consistency and variability in visual concept representations across childhood.

Overall, these findings suggest that children's visual concepts gradually become more precise across childhood, and broaden our view on the timeline and mechanisms for visual concept learning. We hope that future work will build on the tools and ideas developed here to understand the visual mind in both children and adults.

References

Ayzenberg, V., & Behrmann, M. (2024). Development of visual object recognition. *Nature Reviews Psychology*, *3*(2), 73–90. https://doi.org/10.1038/s44159-023-00266-w

Bloom, P. (2000). *How children learn the meanings of words*. Boston, MA: MIT Press.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind*, *3*, 52–67. https://doi.org/https://doi.org/10.1162/opmi_a_00026

Vales, C., Stevens, P., & Fisher, A. V. (2020). Lumping and splitting: Developmental changes in the structure of children's semantic networks. *Journal of Experimental Child Psychology*, *199*, 104914.