MACHINE LEARNING

Grounded language acquisition through the eyes and ears of a single child

Wai Keen Vong[1]*, Wentao Wang[1], A. Emin Orhan[1], Brenden M. Lake[1,2]

Starting around 6 to 9 months of age, children begin acquiring their first words, linking spoken words to their visual counterparts. How much of this knowledge is learnable from sensory input with relatively generic learning mechanisms, and how much requires stronger inductive biases? Using longitudinal head-mounted camera recordings from one child aged 6 to 25 months, we trained a relatively generic neural network on 61 hours of correlated visual-linguistic data streams, learning feature-based representations and cross-modal associations. Our model acquires many word-referent mappings present in the child's everyday experience, enables zero-shot generalization to new visual referents, and aligns its visual and linguistic conceptual systems. These results show how critical aspects of grounded word meaning are learnable through joint representation and associative learning from one child's input.

Philosophers and cognitive scientists have argued that learning a new word requires sorting through a vast, potentially infinite set of candidate meanings (*1*–*3*). For instance, when a child hears the word "ball" in an utterance, how do they learn to associate this word with round, bouncy objects (i.e., the correct visual referents), rather than with other features, objects, or events? Young children are highly adept word learners: At 6 to 9 months, they begin connecting words to their visual counterparts (*4*). By 18 to 24 months, they can comprehend 300 words on average, mostly nouns (*5*, *6*). How do children get started on word learning? What ingredients (e.g., learning mechanisms and representational commitments) are needed to learn word-referent mappings from early experience?

The nature of these ingredients is the subject of intense interest and debate. One prominent theory is that word learning is driven by simple, domain-general, associative learning mechanisms (*7*–*11*), such as tracking the co-occurrences between stimuli in ways nonspecific to language. Alternative theories point to stronger constraints on word learning (e.g., innate knowledge or domain-specific inductive biases; for instance, that different words have different meanings) (*12*–*14*), or to other emerging cognitive abilities that actively support word learning (e.g., reasoning and social cognition) (*3*, *15*). Each account is well-supported by empirical studies in the lab (*3*, *9*, *13*, *14*, *16*, *17*), but acknowledging the evidence for multiple learning mechanisms, often measured across different developmental time points, does not reveal their relative importance. Nor does it provide sufficient guidance for building computational models that, like children, aim to learn outside the lab. If a model could perceive the world through a child's eyes and ears, would it need strong inductive biases or additional cognitive capacities to get word learning underway? Or would a simpler account driven by associative learning, in conjunction with feature-based representation learning (*18*), suffice?

In this article, we put the simplest theories of word learning to an unprecedented test: We examine the learnability of word-referent mappings from a single child's longitudinal head-mounted video recordings. To do so, we introduce the Child's View for Contrastive Learning model (CVCL, as shown in Fig. 1B) which instantiates a form of associative learning that is cross-situational, tracking the co-occurrences between words and possible visual referents to determine their mappings (*10*, *19*–*22*). CVCL interprets this idea through recent advances in multimodal (e.g., vision-and-language) machine learning that integrates representation learning and associative learning (*23*–*26*), using a contrastive objective that coordinates two neural networks, a vision encoder and a language encoder. Trained in a self-supervised manner (i.e., using only the recordings from the child's view and no outside labels), the contrastive objective brings together the embeddings (vectors) of video frames and linguistic utterances that temporally co-occur (treating the co-occurrences as positive evidence), while separating those that do not (treating the absence of co-occurrence as implicit negative evidence), as shown in Fig. 1B. Assuming that spoken utterances correlate with observable visual referents, CVCL converts these temporal associations into a smooth learning signal for learning and aligning its multimodal representations. Without strong constraints on word meaning, nor advance knowledge of possible visual referents, this combination of representation learning and associative learning enables the recovery of many, although not all, of the underlying word-referent mappings from a child's recorded input.

We train CVCL on the SAYCam-S dataset of longitudinal egocentric video recordings from an individual child (*27*), which consists of clips over a 1.5-year period of the child's life (6 to 25 months), with a total of 600,000 video frames paired with 37,500 transcribed utterances (extracted from 61 hours of video; data examples in Fig. 1A, with additional details in the Supplementary Materials or SM S.4). Thus, SAYCam-S provides an extended, first-person window into one child's experiences, but it only captures about 1% of the child's waking hours (*28*) and misses other aspects of their experience (e.g., action and embodiment). Despite these limitations, applying machine learning to the most realistic proxy experience to date can help illuminate the necessary ingredients for learning (*29*, *30*).

We find that CVCL can learn powerful multimodal representations from limited slices of one child's experience. In the following sections, we show that CVCL is capable of matching a range of everyday words to their corresponding visual referents in categorization tasks, generalizing to highly novel visual exemplars not seen during training, and achieving broad-scale alignment between visual and linguistic conceptual systems. Our results suggest that multimodal representation learning paired with domain-general, associative learning mechanisms provides a computational foundation for breaking into word learning.

Evaluating acquired word-referent mappings

After training was completed, we evaluated CVCL and various alternative models for the quality of their learned word-referent mappings. Adapting a common procedure for testing children (Fig. 1, C and D) (*31*), models were prompted with a target category label and selected the corresponding visual referent among four candidate images (based on their cosine similarity to the label). Figure 2A shows the results for Labeled-S: an evaluation dataset of frames annotated with 22 visual concepts that were jointly present in this child's visual and linguistic experience. This dataset was adapted from (*32*) [see supplementary materials (SM) S.5 for additional details]. Overall, CVCL's classification accuracy was 61.6%. Figure 2D shows the breakdown in performance across the different evaluation categories, where CVCL's performance for 11 out of the 22 concepts was found to be within 5% of the upper-bound estimate from CLIP (*25*), a similar image-text contrastive neural network, but trained on several orders of magnitude more data (400 million image-text pairs from the web). To address any potential issues related to category overlap in the evaluation frames, we also conducted a follow-up evaluation using a manually filtered subset with 15 mutually exclusive categories (see SM S5 and fig. S3).

[1]Center for Data Science, New York University, New York, NY, USA. [2]Department of Psychology, New York University, New York, NY, USA.
*Corresponding author. Email: waikeenvong@gmail.com

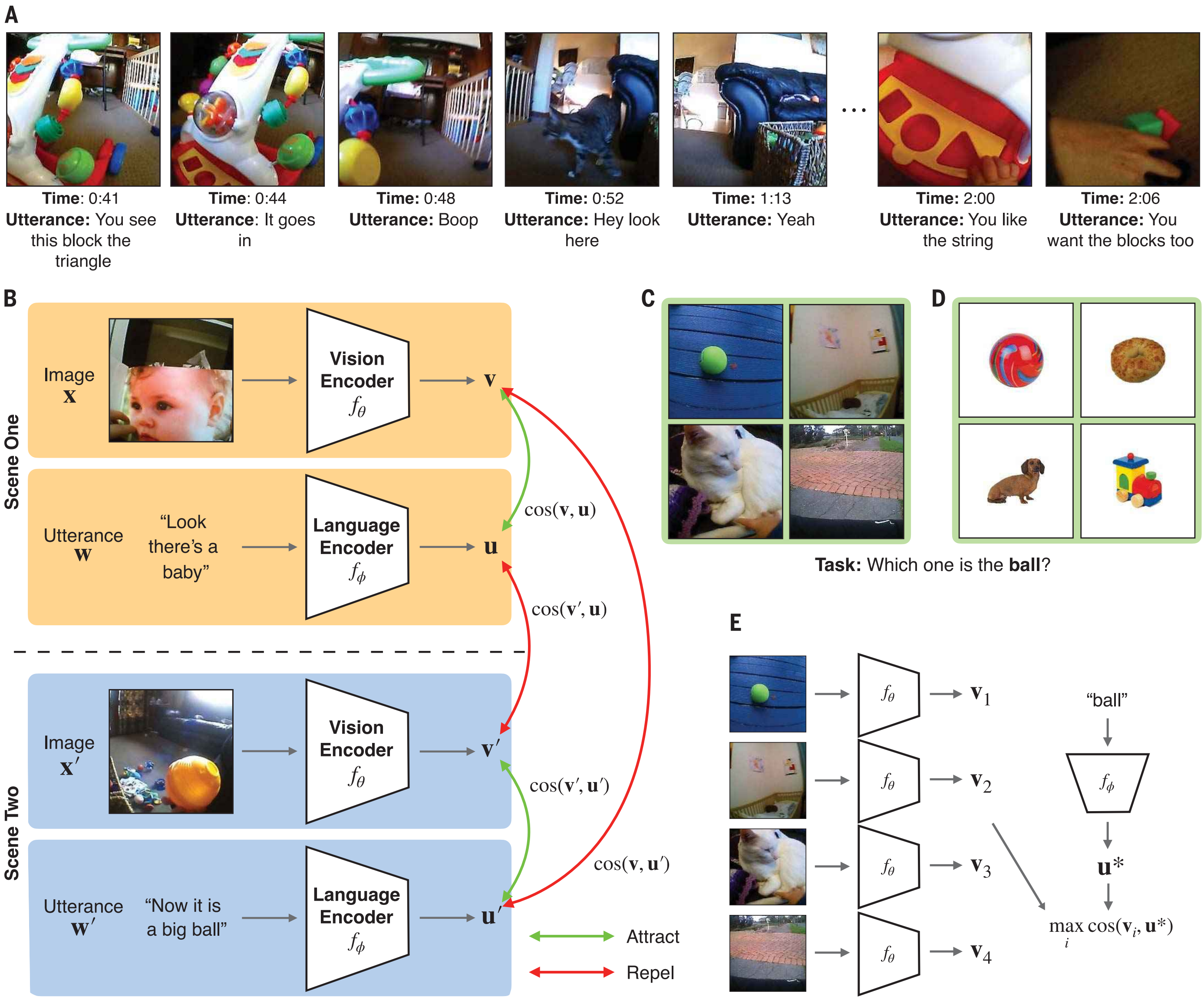

Fig. 1. CVCL model architecture and evaluation procedure. (**A**) Examples of paired frames and child-directed utterances (transcribed) from a single video in the SAYCam-S dataset, highlighting the noisy and sparse co-occurrences between visual and verbal information. (**B**) Images and utterances are embedded in a joint vector space through separate modality-specific neural networks. During training, matching pairs are brought closer (higher cosine similarity) whereas mismatching pairs are pushed apart. Example evaluation trials with (**C**) visually similar or (**D**) visually distinct images from those seen during training. The goal is to select the image matching the target concept's label. (**E**) During evaluation trials, the encoders produce embeddings for the target concept's label and each of the candidate images. The image with the highest cosine similarity with the label is selected.

CVCL was compared to alternatives (see SM S2 for details) that aimed to capture meaningful lower and upper bounds on performance (Fig. 2A). To lesion the visual-linguistic co-occurrences, we trained a variant using a training dataset in which the co-occurring frames and utterances were randomly shuffled and instead paired with other frames and utterances from the training set (CVCL-Shuffled), breaking the original co-occurrence links while preserving the information from each independent modality. This model performed at chance (mean, M = 26.7%), showing the critical role of consistent visual and verbal co-occurrence for learning. To lesion the use of strong visual embeddings (CVCL-Random Features), CVCL's vision encoder was randomly initialized and frozen during training. Again, performance dropped substantially (M = 38.0%), although a few concepts such as "sand" and "car" were partially acquired (Fig. 2D). We also estimated two upper bounds on performance based on models that use either outside or oracle training data, beyond what a child has access to. Evaluating CLIP (*25*) out-of-the-box achieved 66.7% accuracy, a 5.1% improvement over CVCL, owing to the relative improvement of a few concepts such as "kitchen," "toy," and "basket." Thus, CVCL's performance is comparable to a strong web-scale contrastive model when tested within-distribution. Finally, to examine the performance achievable with direct supervision with individual category labels (from the manually annotated Labeled-S evaluation set) rather than child-directed utterances, we trained a Linear Probe model. This Linear Probe was constructed by fitting a linear classifier on top of the frozen pretrained vision encoder (initialized from self-supervision) and achieved 81.6% accuracy based on thousands of within-distribution supervised examples.

As a follow-up, we aimed to quantify the value of a word occurring in a natural utterance versus in a directly labeled example. As shown in Fig. 2B, we trained additional Linear Probes with fewer labeled examples (using 10 and 1% of the available labeled data), with the number of natural language examples for CVCL and directly labeled examples for the Linear Probes displayed in table S2. Reducing the amount of directly labeled supervision resulted in an expected decrease in classification accuracy to 77.2 and 65.9%, respectively (with per category performance in fig. S2). Despite the limited number of labeled examples in the 1% Linear Probe, its performance was marginally better than and most comparable to that of CVCL (Fig. 2B). Furthermore, by comparing their relative frequencies, we can conservatively estimate that one directly labeled example is worth at least seven examples from natural language. Nevertheless, natural language supervision has the advantage of more accurately representing what children learn from, and enabling a flexible representational

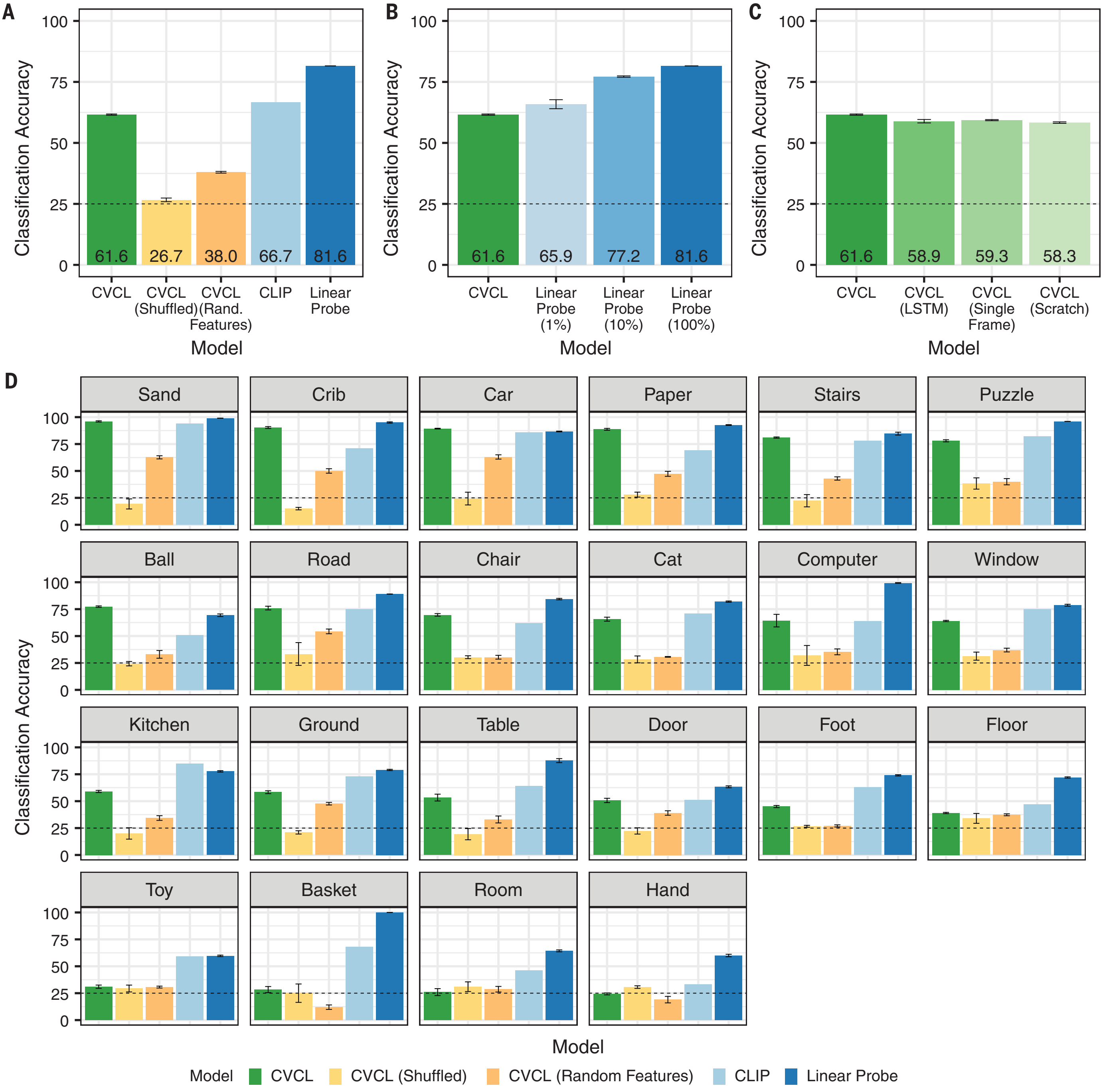

Fig. 2. Image classification accuracy from Labeled-S evaluation. (**A**) Performance of CVCL (green) compared to alternative models that represent upper and lower bounds. The performance of the upper bounds cannot be directly compared to CVCL because they are trained with much more (or cleaner) data. (**B**) Performance of CVCL compared to multiple Linear Probes trained with varying levels of direct supervision. (**C**) Performance of CVCL compared to CVCL model variants. (**D**) Performance broken down by target category. In each graph, error bars represent standard error across models trained with three different random seeds, and the dashed line represents chance accuracy.

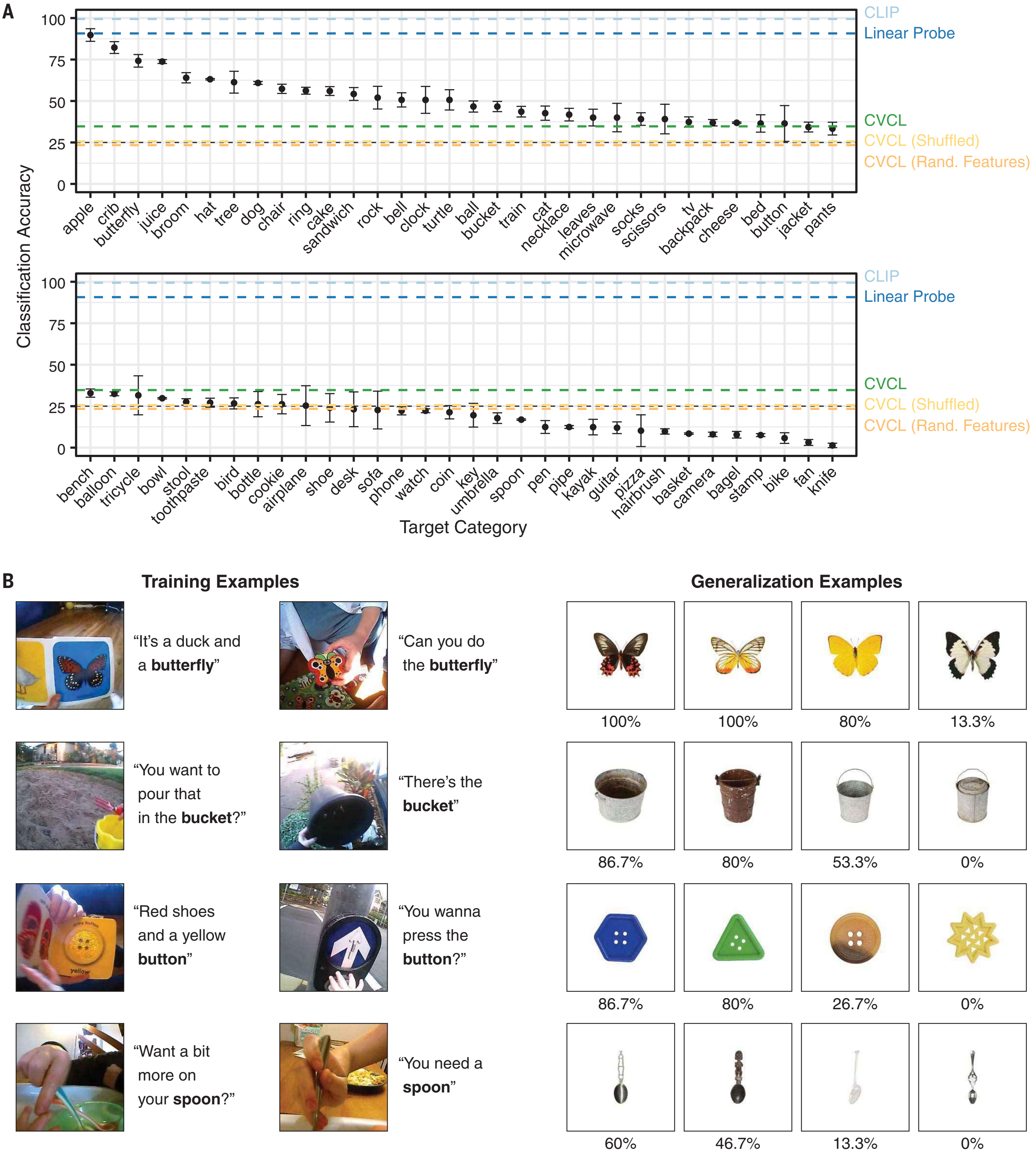

Fig. 3. Zero-shot classification accuracy from Konkle objects evaluation. (**A**) Performance of CVCL as evaluated using 64 visual categories. Error bars represent standard error across three models trained with different random seeds, and the black dashed line represents chance accuracy. The colored dash lines represent the overall accuracy across all trials for CVCL, as well as the other upper and lower bounds. CVCL performed significantly above chance, and better than either lower-bound estimate, but still struggled across many categories, whereas both upper bounds were close to ceiling (owing to training on these types of images). (**B**) In each row, two randomly selected training examples (image-utterance pairs) for four different visual concepts (in bold) are shown, alongside four test examples corresponding (left to right) to the two top, the median, and the worst exemplars. The percent correct below each generalization example refers to performance when this image is the target.

scheme that accommodates an unbounded number of visual concepts. Separately, to examine whether other factors influenced the learnability of word-referent mappings, we also trained and evaluated additional variants of the CVCL model, varying either aspects of the model architecture or the training procedure, although none performed better than CVCL itself (see Fig. 2C and SM S2 for details). Overall, our findings suggest that many of the earliest word-referent mappings can be acquired from as few as 10 to a hundred naturally occurring word-referent pairs.

Generalizing to novel visual exemplars

Using the same training runs, we also measured CVCL performance on the Konkle Objects evaluation dataset, containing naturalistic object categories derived from (*33*) [see Fig. 1D and fig. S1 (right panel) for example trials, and SM S5 for details]. This evaluation included 64 visual concepts whose corresponding words were all present in CVCL's vocabulary (*34*), with images containing a single object on a white background, inspired by the kinds of laboratory experiments used to study infant language development (*4*). This allowed us to examine whether the words learned by CVCL generalize to out-of-distribution visual stimuli—that is, novel category examples on an atypical background. As summarized in Fig. 3A, CVCL demonstrates modest knowledge of these additional visual concepts, with 16 of the 64 concepts scoring better than 50% correct and an additional 42 concepts scoring above chance (25%). The overall accuracy was 34.7%, and although this was lower than the Labeled-S evaluation, the task demands a larger set of concepts (whose word frequency in the training set was highly varied; see table S3), as well as the additional difficulty of out-of-distribution generalization. Additionally, both of the lower bounds were around chance accuracy (25.6 and 23.4% for the CVCL-Shuffled and CVCL-Random Features models respectively), whereas both upper-bound estimates were near ceiling (99.4 and 90.7% for CLIP and the Linear Probe models, respectively), as both models are trained on these types of images.

These results show how CVCL's multimodal representations can permit out-of-distribution generalization, consistent with other larger-scale demonstrations of this ability (*25*, *35*). To illustrate the degree of visual generalization required in this evaluation, Fig. 3B presents some naturalistic training instances (from the child's view) of a word embedded in an utterance, matched with novel test images used for evaluation (along with their classification accuracy). Furthermore, this evaluation closely resembles the kinds of stimuli presented in classic infant word learning experiments (*4*, *31*), demonstrating that representations acquired outside the lab can explain how infants generalize to novel visual stimuli inside the lab.

The organization of multimodal representations

In this section, we present three families of analyses exploring the structure of the learned multimodal representations in CVCL. First, we examined the extent to which CVCL's visual and linguistic conceptual systems align. For example, if both the vision and word embeddings for "car" are independently more similar to "road" than "ball," this would indicate good multimodal alignment (*36*, *37*).

Using the 22 concepts from Labeled-S, we computed a visual prototype for each concept by randomly sampling 100 annotated frames, extracting their image embeddings and averaging across frames. We also retrieved each concept's corresponding word embedding. Next, we computed all pairwise cosine similarities from these embeddings (both within and across modalities) and visualized their relationship using t-distributed stochastic neighbor embedding (t-SNE) as shown in Fig. 4, A and B. In Fig. 4A, the dashed lines represent the distance between each concept's corresponding visual centroid and word embedding. Because many of these cross-modal distances are small, we examined whether the within-modal similarities between concepts (via cosine) are related across vision and language, finding a significant degree of conceptual alignment (correlation coefficient $r = 0.37$, $p < 0.001$). These relationships did not hold for either of the two lower bounds for CVCL (fig. S4). Furthermore, alignment distance was also strongly negatively correlated to classification performance ($r = -0.65$, $p = 0.001$), with some of the least accurate categories exhibiting the largest distance between their respective visual prototype and word embeddings [e.g., "hand" (*38*); fig. S5]. Figure 4B illustrates a subset of labeled image embeddings from each concept, highlighting that different visual concepts vary in how tightly clustered their examples are. By considering visual variability as the average Euclidean distance of a concept's visual embeddings to its visual prototype (*37*), we also find a strong negative correlation to classification performance ($r = -0.48$, $p = 0.025$), suggesting that CVCL's difficulty with word-referent mappings such as "hand" and "toy" is linked to their visual variability, compared to tightly clustered concepts like "car" and "crib."

Next, we visualize how different word embeddings interact with image embeddings in CVCL (Fig. 4C). Examining three different concepts, we observe that the images that the model predicts to be most similar to a particular word embedding (shown in green) closely approximate the true set of labeled images from each class (shown in blue), with the full set of concepts shown in fig. S6. We find that CVCL learns to represent different sets of visually similar items from a concept as distinct subclusters, despite using a single vector per word. For example, the word embedding for "stairs" most strongly activates two separate clusters representing indoor versus outdoor stairs, whereas "puzzle" produces two other clusters that represent alphabet versus animal puzzles. Previous psychological theories of concept learning often required explicit, built-in mechanisms to capture substructure within concepts (*39*, *40*), but in CVCL, we find that multicluster representations emerge implicitly through contrastive learning.

We also qualitatively examined CVCL's ability to localize referents. For a given image, we obtained an attention map by applying Grad-CAM (*41*), highlighting image regions most relevant to the target category by computing a weighted sum of the final convolutional layer's feature maps (using weights based on a spatial average of the gradient of the image-text cosine similarity with respect to the feature maps). We can overlay this attention map over the image and check for any correspondence between the location of the referent and the attention map. Figure 5 presents multiple examples of attention maps from four concepts. For some classes, CVCL's attention maps provide evidence of object localization: The highest activating regions in the attention map closely track the location of the referent. Additional randomly selected attention maps are shown in fig. S7.

Discussion

In this article, we introduced the CVCL model, a deep neural network for grounded word learning from slices of one child's egocentric experience. Across a series of experiments, we found that CVCL can acquire word-referent mappings through naturalistic learning, generalize beyond the specific visual referents in the child's environment, and align its visual and linguistic representations. Our work builds on recent advances in multimodal machine learning (*23*, *25*, *26*, *42*), which also learn to associate words and visual referents, although through increasingly large and unrepresentative training datasets compared to how children learn. Owing to this data gap (*43*), the relevance of these machine learning advances for understanding early language acquisition has been, until now, unclear. Here, we showed how CVCL can effectively learn words from developmentally realistic data from an individual child. Establishing learnability from individual (rather than aggregate) data (*30*, *44*–*47*) is noteworthy because children must learn language from their own limited input. In this more rigorous and ecologically valid setting, CVCL suggests that paired representation and associative learning provides a genuine start to this problem.

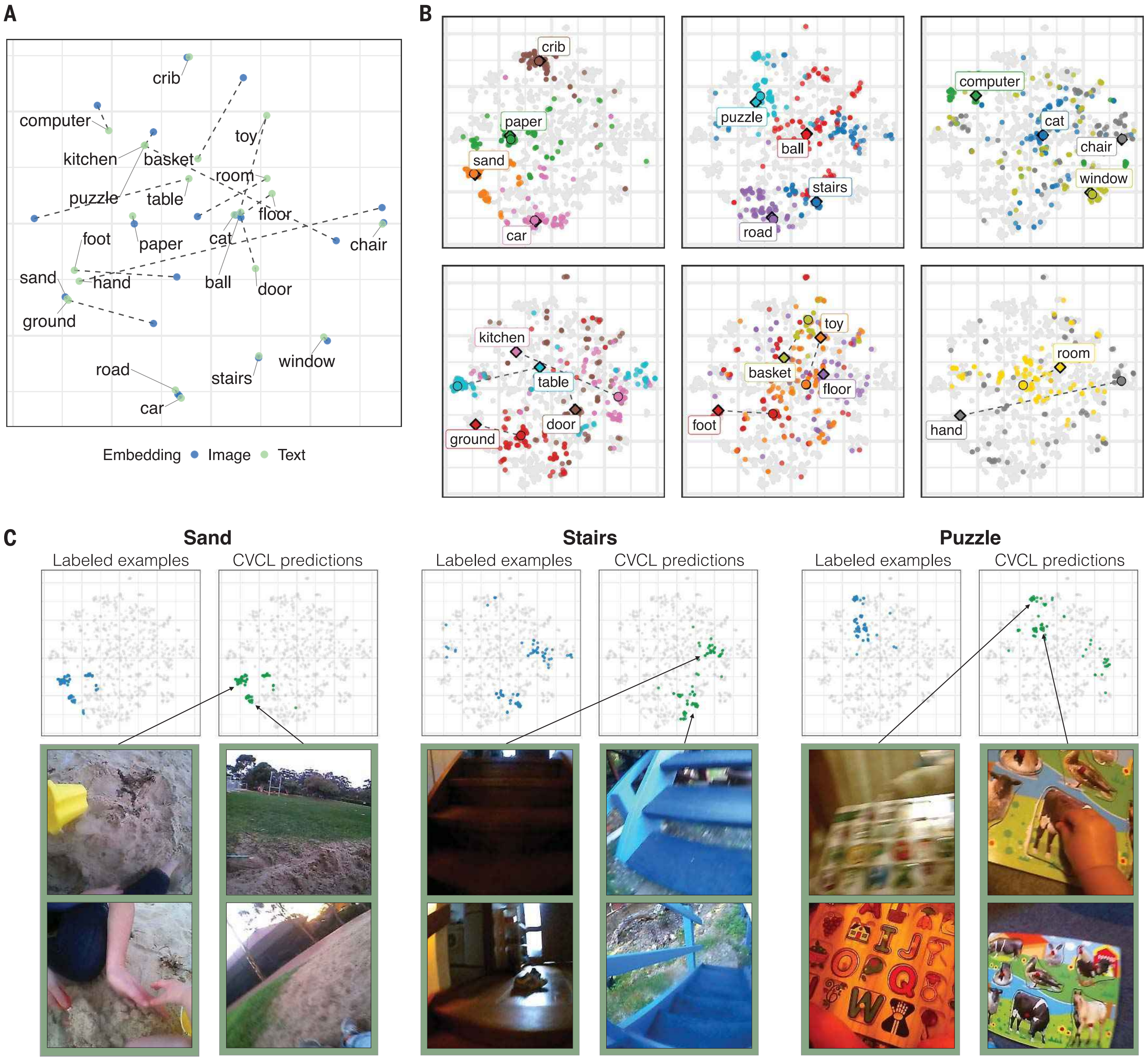

Fig. 4. Conceptual alignment of visual and language modalities. (**A**) A t-SNE plot derived from the cosine similarities between the mean image embeddings and text embeddings from concepts in Labeled-S, showing the high degree of alignment across the visual and linguistic conceptual systems. (**B**) t-SNE plots showing 100 labeled image embeddings for each concept (randomly chosen), highlighting how concepts vary both in how many distinct clusters are required to represent them and how tightly clustered or loosely scattered points from the same class can be. Additionally, for each concept, we also show its corresponding text embedding (diamond) and mean image embedding (circle). (**C**) In each plot, we visualize how CVCL predictions compare to the labeled examples using t-SNE, using a subset of the frame embeddings from the Labeled-S evaluation. The blue points on the left correspond to the 100 frames belonging to a particular category, and the green points on the right correspond to the 100 highest activating frames (based on the cosine similarity to each concept's word embedding from CVCL). Below each plot, we show multiple example frames belonging to either one or multiple subclusters for each concept, capturing how word embeddings interact with image embeddings in the joint embedding space. For example, for the word "stairs," we see that one cluster represents images of indoor wooden stairs, and the other main cluster represents images of the outdoor blue set of stairs. All of the t-SNE plots in these figures are derived from the same set of joint image and text embeddings.

CVCL's successes have broader implications for theories of word learning. Alternative theories posit reliance on strong inductive biases, specialized language machinery, or other cognitive abilities, in part, because word learning was assumed to be too hard otherwise (*2*, *3*, *13*, *14*) (with different perspectives focusing on evidence from different developmental ages). CVCL's focus on learnability with minimal ingredients shows how representation learning and associative, cross-situational mechanisms (*9*, *19*, *22*, *48*) are sufficient to acquire word-referent mappings from one child's first-person experiences. Rather than counting co-occurrences between discrete symbolic

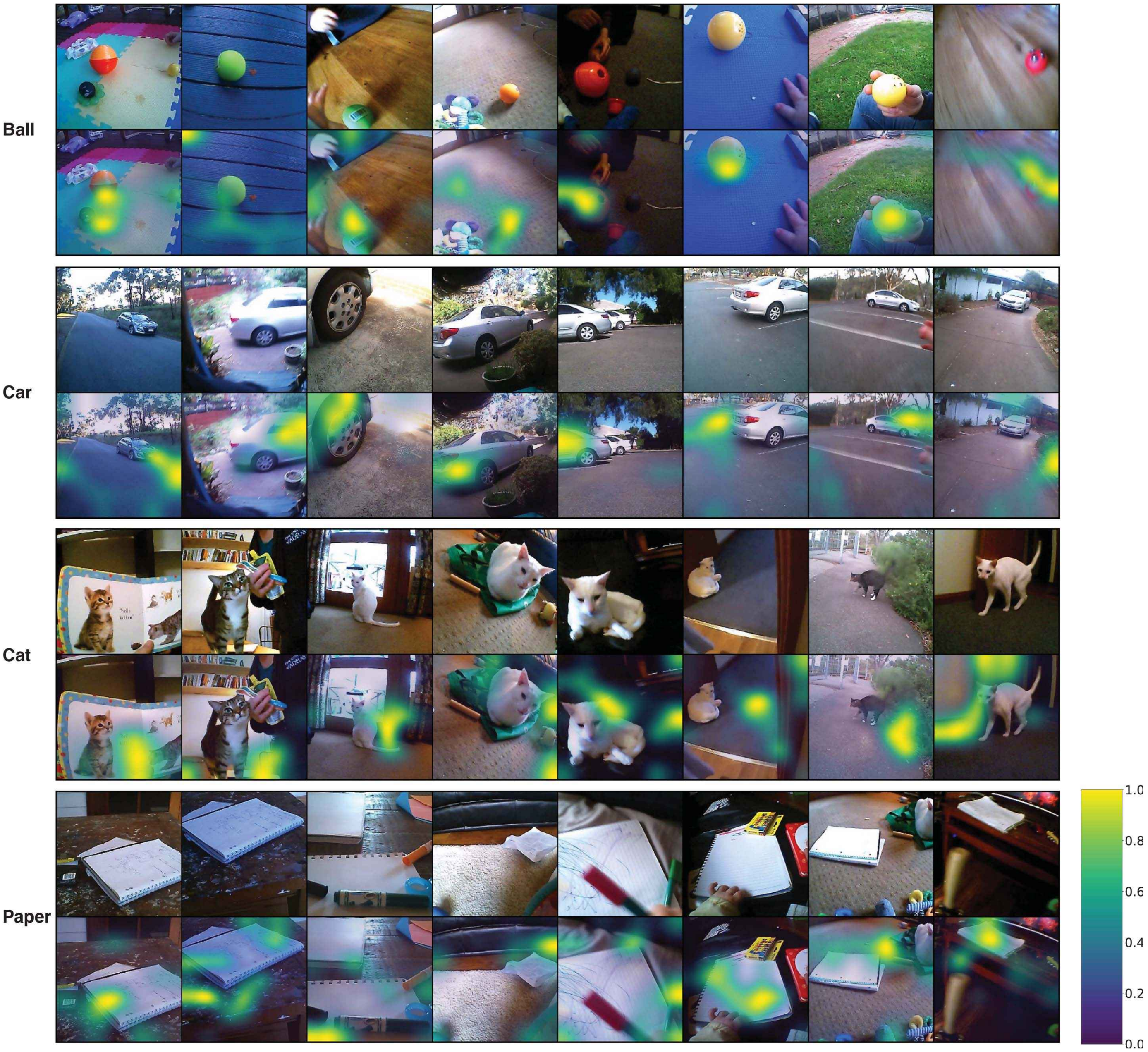

Fig. 5. Attention maps generated by Grad-CAM for four different categories showing some object localization capabilities in CVCL. Each plot contains eight different examples from a category, with the corresponding normalized attention map below, where yellow indicates regions with the highest attention. For the top two categories (ball and car), we see that the model can localize the intended referent across different views. However, in the bottom two categories (cat and paper), the attention maps are sometimes misaligned with the referent, showing that this ability to localize referents is not consistent across all categories. Images were manually selected from Labeled-S where the referent was visible and clear but did not take up the entire frame.

entities like traditional cross-situational models, CVCL encodes both words and images using distributed vectors (*49*–*52*). The contrastive objective leverages the temporal co-occurrence of words and images as an associative learning signal, enabling the incremental acquisition and alignment of multiple word-referent mappings jointly, and sidestepping previous conceptualizations of word learning as a discrete search over a vast number of candidate hypotheses (*1*, *53*). Contrastive learning is also a broadly applicable and domain-general learning strategy (*26*, *52*, *54*), allowing CVCL to learn representations informed by both within-domain [e.g., word-to-word (*49*–*51*)] and across-domain (e.g., word-to-image) correlations (*36*).

Although our primary aim was establishing the learnability of word-referent mappings with minimal ingredients, CVCL's successes do not rule out more sophisticated forms of representation and reasoning, especially ones that might emerge in later development (*55*). These include mutual exclusivity (*13*), the principle of contrast (*12*), the shape bias (*56*), syntactic cues (*57*), social or gestural cues (*15*), or hypothesis generation (*58*). Each of these additional factors has empirical support and their inclusion may further improve learning, to the extent that they do not already emerge from training. Subsequent research could systematically test for their contributions on top of CVCL, by incorporating these biases into the architecture or training procedure (*59*, *60*). Nevertheless, our findings suggest that they are not essential for making genuine progress on word learning from one child's experience.

Future work could aim to bring learning in children and models closer together by incorporating more cognitively plausible assumptions into the model. For example, children learn from temporally extended episodes (*61*), whereas CVCL learns from independent still frames, likely affecting the learnability of verbs and other abstract words. Second, children are fundamentally active, embodied learners, whereas CVCL must learn passively from recorded visual-linguistic experience. CVCL's successes are implicitly supported, in part, by the child's actions, attention, and social engagement, although other benefits of active learning are beyond the model's reach (*62*). Third, children learn continually from an ongoing stream of experience, whereas CVCL learns by revisiting its training data repeatedly over multiple epochs, although continual contrastive learning has been successful too (*63*). Finally, young children must learn from speech whereas CVCL learns from transcribed utterances, trading useful speech cues like intonation and emphasis for explicit word boundaries (*30*).

Even with these modeling and data limitations, CVCL demonstrates how grounded word learning is achievable from slices of a single child's experience. There are other aspects of word meaning, such as links to beliefs, intentions, and general semantic knowledge (*64*, *65*), that are beyond the scope considered here. Still, CVCL's promising performance on naturalistic word learning shows the power of combining representation learning and associative learning for addressing a long-standing challenge in early language acquisition.

REFERENCES AND NOTES

1. W. V. O. Quine, *Word and Object* (MIT Press, 1960).
2. S. Carey, *Linguistic Theory and Psychological Reality*, M. Halle, J. Bresnan, G. A. Miller, Eds. (MIT Press, 1978), pp. 264–293.
3. P. Bloom, *How Children Learn the Meanings of Words* (MIT Press, 2002).
4. E. Bergelson, D. Swingley, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3253–3258 (2012).
5. L. Fenson *et al*., *MacArthur-Bates Communicative Development Inventories* (Paul H. Brookes, 2007).
6. M. C. Frank, M. Braginsky, D. Yurovsky, V. A. Marchman, *J. Child Lang.* **44**, 677–694 (2017).
7. T. Regier, *Cogn. Sci.* **29**, 819–865 (2005).
8. E. Colunga, L. B. Smith, *Psychol. Rev.* **112**, 347–382 (2005).
9. C. Yu, L. B. Smith, *Psychol. Sci.* **18**, 414–420 (2007).
10. L. B. Smith, S. H. Suanda, C. Yu, *Trends Cogn. Sci.* **18**, 251–258 (2014).
11. J. Fiser, R. N. Aslin, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15822–15826 (2002).
12. E. V. Clark, B. MacWhinney, *Mechanisms of Language Acquisition* (Lawrence Erlbaum Associates, 1987), pp. 1–33.
13. E. M. Markman, *Categorization and Naming in Children: Problems of Induction* (MIT Press, 1989).
14. N. N. Soja, S. Carey, E. S. Spelke, *Cognition* **38**, 179–211 (1991).
15. M. Tomasello, M. J. Farrar, *Child Dev.* **57**, 1454–1463 (1986).
16. D. A. Baldwin, *Child Dev.* **62**, 875–890 (1991).
17. M. Bohn, M. H. Tessler, M. Merrick, M. C. Frank, *J. Exp. Psychol. Gen.* **151**, 2927–2942 (2022).
18. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**, 436–444 (2015).
19. D. K. Roy, A. P. Pentland, *Cogn. Sci.* **26**, 113–146 (2002).
20. L. Smith, C. Yu, *Cognition* **106**, 1558–1568 (2008).
21. M. C. Frank, N. D. Goodman, J. B. Tenenbaum, *Psychol. Sci.* **20**, 578–585 (2009).
22. A. Fazly, A. Alishahi, S. Stevenson, *Cogn. Sci.* **34**, 1017–1063 (2010).
23. A. Lazaridou, G. Chrupała, R. Fernández, M. Baroni, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016), pp. 387–392.
24. D. Harwath *et al*., in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 649–665.
25. A. Radford *et al*., Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] (2021).
26. M. Nikolaus, A. Alishahi, G. Chrupała, *Trans. Assoc. Comput. Linguist.* **10**, 922–936 (2022).
27. J. Sullivan, M. Mei, A. Perfors, E. Wojcik, M. C. Frank, *Open Mind (Camb.)* **5**, 20–29 (2021).
28. We obtain this estimate by approximating the total number of hours during this time span as 1.583 × 365 × 24 = 13,867 hours, and assuming that half of these are waking hours, then the proportion of time that SAYCam-S covers is 61/(0.5 × 13867) ≈ 0.01. Similarly, our training set consists of 225,000 linguistic tokens, and comparing this to estimates that suggest children hear between 2 million to 7 million words per year (*66*), suggests that the proportion of language is around 0.8% to 2.2% of the total language input received by this child.
29. E. Dupoux, *Cognition* **173**, 43–59 (2018).
30. A. Warstadt, S. R. Bowman, What Artificial Neural Networks Can Tell Us About Human Language Acquisition. arXiv:2208.07998 [cs.CL] (2022).
31. K. Hirsh-Pasek, R. M. Golinkoff, *Methods for assessing children's syntax, D. McDaniel, C. McKee*, H. S. Cairns, Ed. (The MIT Press, 1996), pp. 105–124.
32. E. Orhan, V. Gupta, B. M. Lake, "Self-supervised learning through the eyes of a child" in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (virtual).
33. T. Konkle, T. F. Brady, G. A. Alvarez, A. Oliva, *J. Exp. Psychol. Gen.* **139**, 558–578 (2010).
34. The word frequency of these concepts in the training dataset varied greatly, with a maximum of 481 examples for ball, to a minimum of 3 examples for desk, fan, kayak, crib, pizza, and tricycle. The frequency of each concept can be found in table S.3.
35. C. Jia *et al*., in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021), pp. 4904–4916.
36. B. D. Roads, B. C. Love, *Nat. Mach. Intell.* **2**, 76–82 (2020).
37. Y. Zhou, M. J. Tarr, D. Yurovsky, Quantifying the Roles of Visual, Linguistic, and Visual-Linguistic Complexity in Verb Acquisition. arXiv:2304.02492 [cs.CL] (2023).
38. Although the denominator in the contrastive objective aims to push away embeddings of frames and utterances that do not temporally co-occur, there are cases where word embeddings can still be very similar. One such case is the large discrepancy between the visual and word embeddings for "hand," because it is primarily spoken about only when the child is playing with sand, leading the model to incorrectly also associate the word "hand" with the referent sand. In this kind of situation, when two different words ("sand" and "hand") are both used to describe the same visual referent, the contrastive objective favors a solution where both the word embeddings for "sand" and "hand" are both associated with the referent for sand, and therefore end up similar to one another.
39. J. R. Anderson, *Psychol. Rev.* **98**, 409–429 (1991).
40. B. C. Love, D. L. Medin, T. M. Gureckis, *Psychol. Rev.* **111**, 309–332 (2004).
41. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 618–626.
42. J.-B. Alayrac *et al*., *Adv. Neural Inf. Process. Syst.* **35**, 23716 (2022).
43. M. C. Frank, Large language models as models of human cognition. PsyArXiv [Preprint] (2023); https://doi.org/10.31234/osf.io/wxt69.
44. A. Perfors, J. B. Tenenbaum, E. Wonnacott, *J. Child Lang.* **37**, 607–642 (2010).
45. B. C. Roy, M. C. Frank, P. DeCamp, M. Miller, D. Roy, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12663–12668 (2015).
46. W. Wang, W. K. Vong, N. Kim, B. M. Lake, *Cogn. Sci.* **47**, e13305 (2023).
47. A. E. Orhan, B. M. Lake, Learning high-level visual representations from a child's perspective without strong inductive biases. arXiv:2305.15372 [cs.CV] (2024).
48. B. McMurray, J. S. Horst, L. K. Samuelson, *Psychol. Rev.* **119**, 831–877 (2012).
49. J. L. Elman, *Cogn. Sci.* **14**, 179–211 (1990).
50. T. K. Landauer, S. T. Dumais, *Psychol. Rev.* **104**, 211–240 (1997).
51. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL] (2013).
52. S. Chopra, R. Hadsell, Y. LeCun, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 539–546 (2005).
53. S. Pinker, *Learnability and Cognition: The Acquisition of Argument Structure* (MIT Press, 1989).
54. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, in *International Conference on Machine Learning* (PMLR, 2020), pp. 1597–1607.
55. S. C. Meylan, E. Bergelson, *Annu. Rev. Linguist.* **8**, 77–99 (2022).
56. B. Landau, L. B. Smith, S. S. Jones, *Cogn. Dev.* **3**, 299–321 (1988).
57. L. Gleitman, *Lang. Acquis.* **1**, 3–55 (1990).
58. J. C. Trueswell, T. N. Medina, A. Hafri, L. R. Gleitman, *Cogn. Psychol.* **66**, 126–156 (2013).
59. K. Gulordava, T. Brochhagen, G. Boleda, in *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, pp. 2089–2095.
60. R. Geirhos *et al*., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness" in *International Conference on Learning Representations (ICLR)* (2019).
61. L. B. Smith, H. Karmazyn-Raz, *Trends Cogn. Sci.* **26**, 1064–1065 (2022).
62. 61. T. M. Gureckis, D. B. Markant, *Perspect. Psychol. Sci.* **7**, 464–481 (2012).
63. W. K. Vong, B. M. Lake, *Cogn. Sci.* **46**, e13122 (2022).
64. E. H. Wojcik, M. Zettersten, V. L. Benitez, *Wiley Interdiscip. Rev. Cogn. Sci.* **13**, e1596 (2022).
65. B. M. Lake, G. L. Murphy, *Psychol. Rev.* **130**, 401–431 (2023).
66. J. Gilkerson *et al*., *Am. J. Speech Lang. Pathol.* **26**, 248–265 (2017).
67. J. Sullivan, M. Mei, A. Perfors, E. Wojcik, M. C. Frank, Head cameras on children aged 6 months through 31 months (SAYCam), Databrary (2017); retrieved 16 October 2023. https://doi.org/10.17910/b7.564.

ACKNOWLEDGMENTS

We are grateful to the authors of the SAYCam article (*27*), and the volunteers who contributed to the data set, for making our article possible. We thank K. Adolph, G. Davidson, J. Guo, T. Kwan, G. Murphy, Y. Zhou, and four anonymous reviewers for valuable feedback on earlier versions of the article. Early versions of this work were presented as an abstract at the 34th CUNY Conference on Human Sentence Processing and the Philosophy of Deep Learning conference. **Funding:** This work was supported by the DARPA Machine Common Sense program and NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science. **Author contributions:** All of the authors contributed to conceptualizing the project and editing the manuscript. Conceptualization: W.K.V., W.W., A.E.O., B.M.L. Modeling: W.K.V., W.W., A.E.O., B.M.L. Investigation: W.K.V., B.M.L.; Visualization: W.K.V., W.W., B.M.L. Writing: W.K.V. (original draft), W.K.V., W.W., A.E.O., B.M.L. (editing). **Competing interests:** None declared. **Data and materials availability:** Pretrained models and code are available at https://github.com/wkvong/multimodal-baby. The SAYCam dataset is accessible to authorized users on Databrary (*67*). Access can be obtained by securing authorization from their institution through the formal Databrary Access Agreement (https://databrary.org/about/agreement.html), or through affiliation from a researcher who has authorization.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi1374
Materials and Methods
Figs. S1 to S5
Tables S1 to S3
References (*68*–*79*)
MDAR Reproducibility Checklist

Submitted 26 April 2023; resubmitted 17 July 2023
Accepted 31 December 2023
10.1126/science.adi1374