Science
AAAS

Supplementary Materials for

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

Wai Keen Vong *et al*.

Corresponding author: Wai Keen Vong, waikeenvong@gmail.com

Science **383**, 504 (2024)
DOI: 10.1126/science.adi1374

The PDF file includes:

Materials and Methods
Figs. S1 to S5
Tables S1 to S3
References

Other Supplementary Material for this manuscript includes the following:

MDAR Reproducibility Checklist

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

S.1. The CVCL Model

The CVCL model (**C**hild's **V**iew for **C**ontrastive Learning) takes in child-directed utterances paired with temporally aligned image frames, with the goal of learning multimodal representations that align information from visual and linguistic modalities and thereby grounding words to their visual referents. Significant advances in multimodal learning using contrastive approaches (*24, 25, 35*) were spurred by the discovery that such models can learn generalizable representations that enable *zero-shot classification*. Contrastive approaches learn by treating an image and its corresponding utterance as a matching pair, while treating any other image or utterance as a mismatching pair, aiming to push embeddings for the former closer together while pushing away embeddings for the latter. This approach is represented in Figure 1B.

Vision Encoder. We use a ResNeXt-50 32x4d convolutional neural network (CNN) architecture as our vision encoder f_θ (*68*). The CNN was initialized with pre-trained weights obtained via self-supervised learning on 194 hours of egocentric visual data **only from this child** using the DINO algorithm (*69*). The pre-trained model is available from the following public GitHub repository: `https://github.com/eminorhan/silicon-menagerie` with the model identifier string '`dino_sfp_resnext50`'. Previous work from (*32, 47*) showed that self-supervised approaches on rich, developmental datasets could learn high-quality representations of visual features. The backbone of this vision encoder was frozen, and combined with a trainable linear projection layer that embedded images into a shared multimodal space, with the embedding size set to 512 for all models.

Because each utterance was associated with multiple frames extracted around a short temporal window, during training we randomly sampled one of these frames as the matching image frame associated with a given utterance. We also applied data augmentation to images during training, using the same set of augmentations as described in (*32*), with the exception of ColorJitter as it breaks the correspondence between color words and color in images.

Language Encoder. We used a simple Embedding layer as our language encoder f_ϕ. In the **Embedding** encoder, the model learns a separate embedding (of size *D*=512) for each word in its vocabulary. To obtain a single embedding for an utterance containing multiple words, an embedding for each word was separately retrieved via the trained embedding layer and then averaged across each of the resulting word embeddings.

Contrastive Loss. For a given mini-batch of frames and their corresponding child-directed utterances, image embeddings are obtained by passing image frames separately through the vision encoder:

$$\mathbf{v_i} = f_\theta(\mathbf{x_i}). \tag{1}$$

Likewise, text embeddings are obtained by passing utterances separately through the language encoder:

$$\mathbf{u_i} = f_\phi(\mathbf{w_i}). \tag{2}$$

After obtaining the image and text embeddings, they were both normalized, and we calculated two different contrastive losses (based on the cross-entropy) that either tried to match each frame with its corresponding utterance in the mini-batch:

$$L_{\text{frame}} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp(\mathbf{v_i}^T \mathbf{u_i}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{v_i}^T \mathbf{u_j}/\tau)}, \tag{3}$$

or match each utterance to its corresponding frame:

$$L_{\text{utterance}} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp(\mathbf{u_i}^T \mathbf{v_i}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{u_i}^T \mathbf{v_j}/\tau)}. \tag{4}$$

All other possible pairings besides the true pairing within a given mini-batch were treated as mismatches. Finally, these two losses were combined and served as our joint contrastive loss to train our multimodal neural network:

$$L = \frac{1}{2}L_{\text{frame}} + \frac{1}{2}L_{\text{utterance}}. \tag{5}$$

S.2. Alternative Models

In addition to the main CVCL model we present in the main text, we also evaluated a number of variants and control models:

CLIP. We utilized CLIP ViT-L/14 (*25*) as one upper bound for zero-shot classification performance. Images and category labels were pre-processed and tokenized using the same respective procedures as in the original CLIP model. Because CLIP is also trained via contrastive learning with the ability to perform zero-shot classification, no other modifications were required for it to make predictions on our main Labeled-S evaluation dataset. In our case, we evaluate CLIP directly using the same text prompt consisting only of the target category label, rather than averaging across many different textual prompts as in the original work.

Linear Probe. As another upper bound estimate, we trained a series of linear probes, aiming to measure the effect of direct supervision for the target concepts, using either 100%, 10% and 1% of the available labeled data. These linear probes were constructed by taking and freezing the original pre-trained vision encoder obtained from self-supervision, and adding a trainable linear output classification head, mapping visual features to a set of class labels. The use of linear probes is common in self-supervised learning, as it provides a measure of the representational quality of the visual features (*54, 70*). The linear probes were trained to directly predict category labels either from the training split of the original Labeled-S evaluation dataset (22-way classification), or from the Konkle Objects evaluation dataset (64-way classification). Each linear probe was trained for 100 epochs. We obtained predictions from the Linear Probe models by passing each of the four evaluation images per trial through the trained probe separately, and outputting a response based on the logit corresponding to the target category label which was largest across the set of four images.

CVCL (Shuffled). As a lower bound, we trained a variant of the CVCL model, but where the set of video frames associated with each utterance were randomly shuffled and paired with new utterances. Shuffling the utterances breaks the consistency of co-occurrence information across modalities, while retaining the same visual and linguistic inputs passed into the CVCL model. This shuffling procedure was performed once before training commenced.

CVCL (Random Features). As a second lower bound estimate, this model was the same as the CVCL model, but where the features from vision encoder were randomly initialized and frozen during training, rather than leveraging pre-trained visual features from self-supervision. This model captures whether words can be grounded from random visual features, rather than learned visual features.

CVCL (LSTM). In this variant, the Embedding language encoder was replaced with a single-layer LSTM (which was also randomly initialized and trained from scratch), to investigate whether richer sequential language processing would be helpful for learning word-referent mappings. The embedding and hidden sizes of the LSTM were set to 512, and Dropout was set to 0.5 and applied after the input embedding. The text embedding for a given utterance was obtained by passing the entire utterance through the LSTM and returning the final hidden state of the LSTM.

CVCL (Single Frame). In this variant, instead of randomly sampling one of the multiple associated video frames with each utterance during training, only the first frame associated with each utterance is used. This reduced the available paired visual information the model associated with each utterance during training.

CVCL (Scratch). In this variant, rather than initializing the vision encoder with pre-trained weights obtained via self-supervision (*69*), the vision encoder was randomly initialized and fully trained from scratch using only the multimodal contrastive objective, allowing us to examine whether visual pre-training was necessary for the acquisition of grounded word-referent mappings.

CVCL (Transformer). To examine whether we can push CVCL even further towards domain-general, generic learning, we replaced both the vision and language encoders with Transformers (*71*). For the vision encoder, we used a ViT-B/14 Transformer that was pre-trained via self-supervision on the SAYCam-S visual data alone (*47*). For the language encoder, we used a single-layer text Transformer with 8 attention heads. Both encoders used learned rather than fixed positional embeddings, thereby minimizing any domain-specific assumptions regarding the ordering of information within each modality. Otherwise, the training procedure was the same as CVCL.

Results. We evaluated these variants of the CVCL model on the Labeled-S dataset. In the CVCL (LSTM) model, we replaced the Embedding language encoder with a Long Short-Term Memory (LSTM) network (*72*), resulting in a slight decrease in classification performance (M=58.9%). This finding suggests that visual concepts in Labeled-S can be meaningfully grounded without necessarily requiring sequential language processing capabilities as implemented in an LSTM. In the CVCL (Single Frame) model, the model was trained using only the first frame from each paired utterance, rather than sampling one of multiple video frames around a short temporal window. This also led to a slight, but negligible decrease in classification performance (M=59.3%), suggesting that the additional visual temporal context provided a limited benefit for classification performance. In the CVCL (Scratch) model, rather than initializing the vision encoder with pre-trained weights obtained from self-supervision on the SAYCam-S video data, we randomly initialized and trained it from scratch via contrastive learning exclusively. This variant also resulted in only a small drop in classification performance (*M*=58.3%), suggesting that prior high-level visual features are not a pre-requisite for grounding. Finally, our evaluations on the CVCL (Transformer) model were comparable to CVCL, achieving an accuracy of 55.5% (vs. 61.6% for CVCL) on the Labeled-S evaluation, and 33.0% (vs. 34.7% for CVCL) on the Konkle Objects evaluation. These results show how the ideas behind CVCL can be taken further to utilize even more generic forms of learning, without sacrificing much in terms of learnability or performance.

S.3. Training Details

Each model was trained for up to 400 epochs using a batch size of 8, and the AdamW optimizer (*73*) with a learning rate of 1e-4. The learning rate was adjusted using ReduceLROnPlateau with a factor of 0.1 and a patience of 20, based on the validation loss. Weight decay of 0.1 was used for all models. In all of our simulations, we set the temperature parameter τ to be fixed at 0.07. Early stopping was performed using the joint contrastive loss on the validation set. For each model, we trained three models using different random seeds. Minimal hyperparameter tuning was performed using a hyperparameter sweep.

S.4. Training Dataset (SAYCam-S)

To study the learnability of word-referent mappings, we required a developmentally representative source of data. We used the SAYCam dataset (*27*), a longitudinal dataset consisting of egocentric head-mounted camera recordings from 3 children (S, A, and Y). Recordings took place for a few hours each week over the course of a few years, with 100-200 hours of recorded video data per child.

Our multimodal training dataset was constructed from the data from one of the three children (S), because it contained the largest proportion of naturalistic speech transcribed (61 hours, spanning 6 to 25 months of age). The vision encoder was pre-trained using the videos from this portion of the dataset, along with 133 hours of additional video-only egocentric data from the same child that was not transcribed. The relevant information from the transcripts were the transcribed utterance, the speaker and the timestamp of the spoken utterance (in seconds). As an initial pre-processing step, due to instances of long annotations spanning multiple minutes, we split annotated utterances into multiple shorter utterances using spaCy (accessed from https://spacy.io). The timestamps for each set of split-up utterances were marked by linearly interpolating between the original starting timestamp and the timestamp of the next original utterance (*74*). Next, we filtered the data to only include child-directed utterances, excluding any utterances produced by the child themselves. We excluded child-produced utterances, since many early utterances from the child contained a lot of babbling, and as our focus was on learning exclusively from the input that the child receives. We also applied the spaCy tokenizer on the filtered utterances from the training set to build the vocabulary, replacing anything annotated as `inaudible` and any tokens with a frequency less than 3 in the dataset with an `<UNK>` token, resulting in a vocabulary size of 2350. All transcripts were lowercased (although in some of our figures some child-directed utterances are capitalized for ease of reading). All utterances were truncated to a maximum length of 25 tokens.

Finally, using the timestamp information associated with each child-directed utterance, we extracted multiple video frames at a rate of 5fps from the beginning of the timestamp, either until the timestamp of the next utterance or until 32 frames were extracted (corresponding to 6.4s of video). Each frame was extracted by resizing the minor edge of the original 640 × 480 video frame to 256, and then applying a 224 × 224 square center crop positioned at 16 pixels lower than the center of the frame (to remove timestamp information visible in the video frame from the head-mounted camera).

The resulting dataset consists of 600,285 image frames paired with 37,486 child-directed utterances. The dataset was split by utterances into a 90%, 5%, 5% split for training, validation and testing purposes respectively. We randomly split the dataset, rather than preserving the temporal ordering of utterances, so that the model treats each frame-utterance pair as independent. In this work, we only use the training and validation splits, and leave the test split for future use. Table S.1 contains some additional descriptive statistics about this dataset.

Relative to image-text datasets like MS COCO (*75*), where the correspondence between images and their paired captions is relatively strong, the correspondence between utterances and image frames in the SAYCam-S dataset is much noisier, as shown in Figure 1A. Among datasets used to study language development, CHILDES (*76*) is primarily text-only, and other egocentric datasets collected from infants consist of shorter videos (measured in minutes) aggregated across multiple infants (*77, 78*). Thus, our dataset provides a unique opportunity for questions about the *learnability* of word-referent mappings from a developmentally representative, temporally extended, longitudinal source of data obtained from a single child.

S.5. Evaluation Datasets

Labeled-S Evaluation. We adapted the Labeled-S dataset from (*32*) to create our main evaluation dataset. The original Labeled-S Dataset consisted of ∼60K frames from 26 common visual categories extracted from videos of baby S. In order to examine the acquisition of word-referent mappings, we re-purposed the images and category labels from this dataset to generate a large set of evaluation trials to evaluate the model's ability to perform zero-shot visual classification.

As a pre-processing step, images from 4 out of the 26 categories (carseat, couch, greenery and plushanimal) were excluded from our evaluation since their category labels were not present in our model's vocabulary, leaving 22 categories for our evaluation. Note that all but a few of these categories are present in the MacArthur-Bates Communicative Development Inventories, the gold standard for tracking infant vocabulary (*5*) (the exceptions are floor, ground, road, sand and computer). Next, we split the dataset so that half of the images were used as a training set for a series of Linear Probes (see S.2 Alternative Models), while the other half of the images were used for testing by constructing a set of evaluation trials. These trials consisted of a target image from one category and its corresponding category label, and three other randomly sampled images from three other randomly sampled foil categories, as shown in Figures 1C and S.1 (left-pane). For each category, we generated 100 distinct evaluation trials, for a total of 2200 trials.

Model predictions were generated by first obtaining a text embedding of the category label via the models' language encoder (*79*). Separately, we passed each image into the trained vision encoder to obtain a series of image embeddings. Then, the cosine similarity was computed between each image embedding and the target text embedding, and the model's prediction based on the image whose cosine similarity with the target text embedding was largest, as depicted in Figure 1E. Note that in this evaluation, the category text embeddings are taken directly from the contrastive training procedure without any additional fine-tuning required.

Although these frames were originally extracted from the same videos as SAYCam-S, the differences between the objectives in training and evaluation tasks mean that visual-linguistic overlaps are not a substantive issue. Examining the two datasets more closely, we found 26 instances (1%) of direct overlap, and approximately 5% of indirect overlap, see Figure S.8 for additional details and examples.

Labeled-S (Filtered) Evaluation. We also evaluated our model on a separate set of evaluation trials using images from a manually cleaned-up subset of Labeled-S, which we call Labeled-S (Filtered). Some of the images from the Labeled-S dataset contained a mixture of object and scene classes within the same image, for example, a chair in the kitchen, or a ball on the ground. This had the unintended effect of making certain evaluation trials more ambiguous than expected, making it difficult to determine whether low performance for certain categories was driven by a failure of learning, or ambiguity in the evaluation trials. To alleviate some of these issues, we performed a two-step filtering procedure to create a cleaner evaluation dataset. First, as an initial automated sweep, we used CLIP ViT-B/16 (*25*) to classify each image using a 22-way classification (based on the categories in Labeled-S), and retained only the images that were correctly classified. Second, we removed the scene classes (ground, floor, kitchen, road, room, sand), as well as any superordinate classes (toy), leaving 15 unique classes. For the remaining set of 15 classes, we performed a second filtering step by manually removing any images that did not contain the target class within the image, or if multiple classes were present, or if the image was too blurry, leaving us with a set of $\sim$2.4K images. From this set, we generated a separate evaluation dataset of 1500 trials (corresponding to 100 trials per category), and re-evaluated CVCL on this set to check whether performance on the original versus filtered evaluations were comparable. The results of this analysis are shown in Figure S.3, indicating that the performance of CVCL increases modestly for the filtered set (64.7% to 72.6%). Nevertheless we see no observed increase in performance for some of the lowest accuracy categories (basket, hand, foot and table), suggesting that the model genuinely was unable to acquire these concepts.

Konkle Objects Evaluation. We generated an additional set of evaluations based on an image dataset containing a large set of common object categories (*33*). This second evaluation was designed to address two limitations of our Labeled-S evaluation dataset. First, the images in the Labeled-S evaluation were from the same visual distribution that CVCL was trained on, so it cannot address whether our models can generalize zero-shot to novel instances for learned visual concepts. Second, the total number of visual categories (22) was limited due to the small set of object categories that were previously annotated for baby S. Yet, many additional concepts are present both visually and linguistically that were not included in these annotations, but that our model could still have acquired over the course of training.

Similar to the process for creating evaluations from the Labeled-S dataset, we filtered the set of object categories to only include ones that were present in the model's vocabulary, from 200 object categories to 64 object categories. For each image, we resized it to be 50% of its original size, and then resized the overall image to be 224 × 224. For each exemplar in each object category, we generated 5 independent evaluation trials in the same manner as above, with examples shown in Figures 1D and S.1. This resulted in a second evaluation dataset consisting of a total of 4763 evaluation trials. Surprisingly, there was little overlap between the visual categories used in both evaluation datasets, with only five overlapping categories (which were ball, crib, basket, cat and chair). Model predictions were obtained using exactly the same procedure as the Labeled-S evaluation, again by taking the category text embeddings directly from the contrastive training procedure without any additional fine-tuning required.

S.6. Supplementary Figures and Tables

	Train	Validation	Test
Number of utterances	33,737	1,874	1,875
Mean (SD) utterance length	6.67 (5.49)	6.59 (5.46)	6.62 (4.95)
Number of tokens	225,001	12,355	12,418
Number of frames	540,681	29,686	29,918
Mean frames per utterance	16.0	15.8	16.0
Out-of-vocabulary rate	1.99%	2.42%	2.79%

Table S.1: **Descriptives for the SAYCam-S Dataset.**

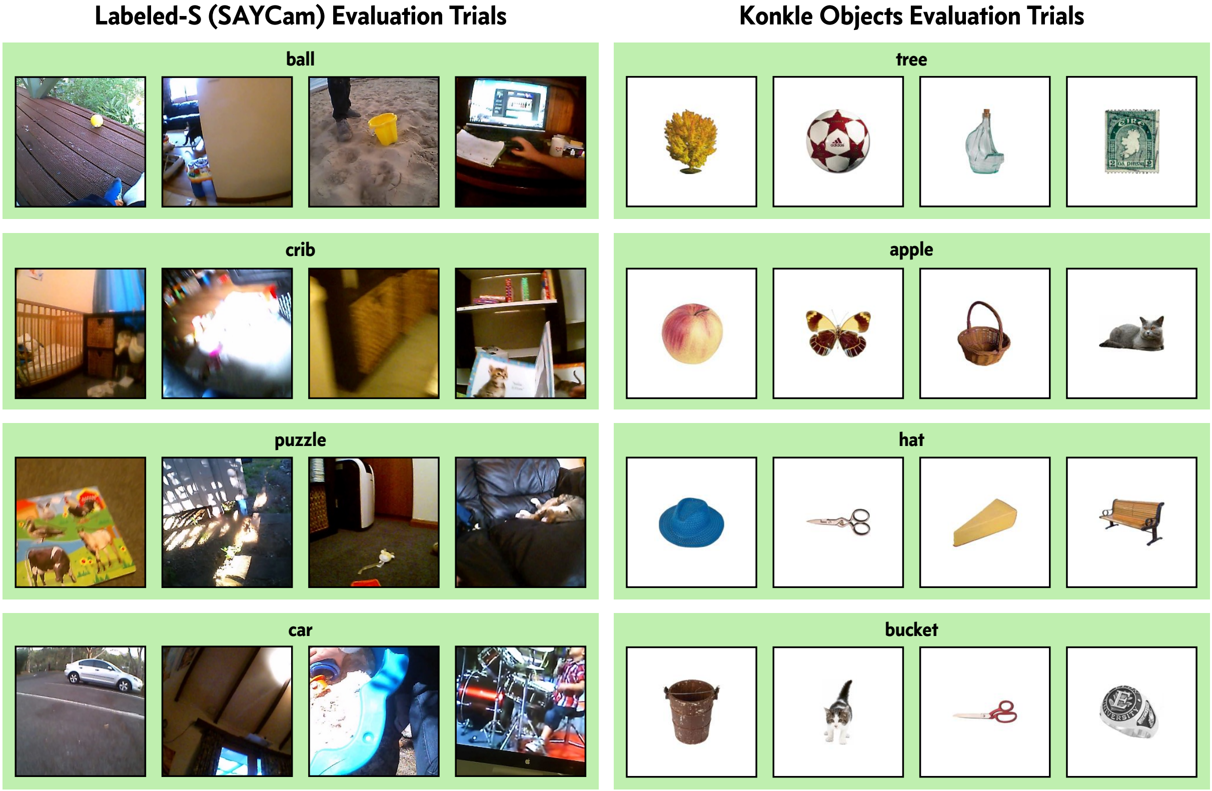
Labeled-S (SAYCam) Evaluation Trials
Konkle Objects Evaluation Trials
ball
tree
crib
apple
puzzle
hat
car
bucket

Figure S.1: **Additional examples from the Labeled-S (left) and Konkle Objects (right) evaluations**. In each example depicted, the target referent is presented on the left, alongside three other randomly sampled foil referents. For the category label, the model was only presented with the single token corresponding to the target category word (e.g. “ball”).

Category	SAYCam-S (Training)	Linear Probe (100%)	Linear Probe (10%)	Linear Probe (1%)
Ball	481	2106	211	22
Basket	19	74	8	1
Car	176	645	65	7
Cat	416	751	76	8
Chair	54	535	54	6
Computer	24	840	84	9
Crib	51	459	46	5
Door	44	1267	127	13
Floor	24	3572	358	36
Foot	114	407	41	5
Ground	18	1090	109	11
Hand	118	1546	155	16
Kitchen	12	1079	108	11
Paper	51	715	72	8
Puzzle	71	1529	153	16
Road	15	1740	174	18
Room	62	1979	198	20
Sand	85	318	32	4
Stairs	36	477	48	5
Table	21	1323	133	14
Toy	37	4307	431	44
Window	32	1188	119	12
Total	1961	27947	2802	291

Table S.2: **Frequency of examples for the 22 Labeled-S categories for training CVCL and Linear Probes.** Each row represents either the word frequency of the category during training (from the SAYCam-S dataset), or the number of supervised examples used to train the corresponding Linear Probe. The bottom row shows the total frequency across all 22 categories.

Category	SAYCam-S	Category	SAYCam-S
Ball	481	Umbrella	18
Cat	416	Phone	17
Train	235	Knife	16
Socks	129	Bagel	11
Bottle	110	Bench	10
Camera	92	Cheese	10
Pants	91	Clock	10
Apple	77	Key	10
Watch	73	Hairbrush	9
Balloon	68	Rock	9
Bucket	59	Turtle	9
Chair	54	Airplane	8
Spoon	53	Ring	7
Crib	51	Sofa	7
Jacket	49	Broom	6
Juice	48	Stool	6
Bowl	46	Bell	5
Tree	46	Cookie	5
Backpack	44	Microwave	5
Bed	43	Scissors	5
Bird	36	Stamp	5
Button	34	Tv	5
Shoe	34	Coin	4
Dog	31	Necklace	4
Hat	28	Sandwich	4
Pen	27	Toothpaste	4
Leaves	26	Desk	3
Bike	23	Fan	3
Butterfly	23	Kayak	3
Cake	22	Pipe	3
Guitar	21	Pizza	3
Basket	19	Tricycle	3

Table S.3: **Frequency of examples per category in the Konkle Objects evaluation.** Each row represent the word frequency of the 64 categories during training (from the SAYCam-S dataset).

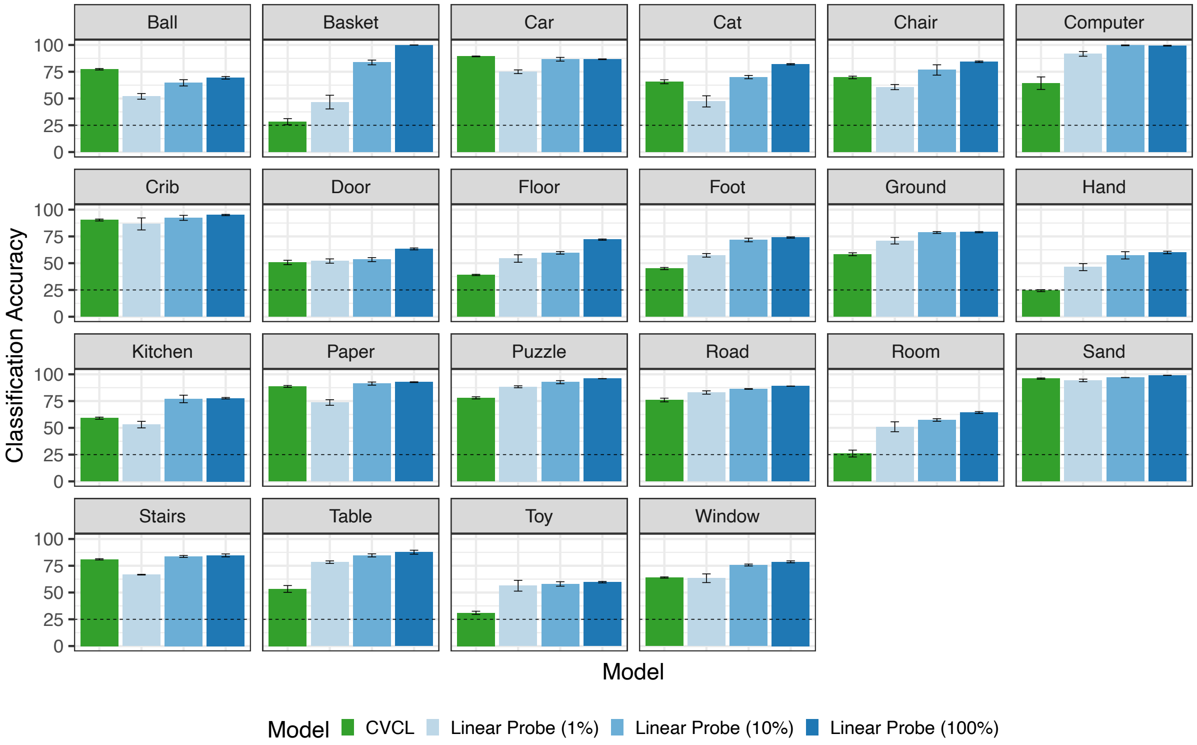

Figure S.2: **CVCL vs. Linear Probe image classification accuracy for Labeled-S evaluations by target category.** In this figure, we compare the performance of the CVCL model to the three Linear Probe models trained with varying levels of direct supervision. Direct supervision is most helpful for categories where the CVCL model is close to chance, while in some cases the CVCL model outperforms any of the Linear Probes. Error bars represent standard error across three models trained with different random seeds, and the dashed line represents chance accuracy.

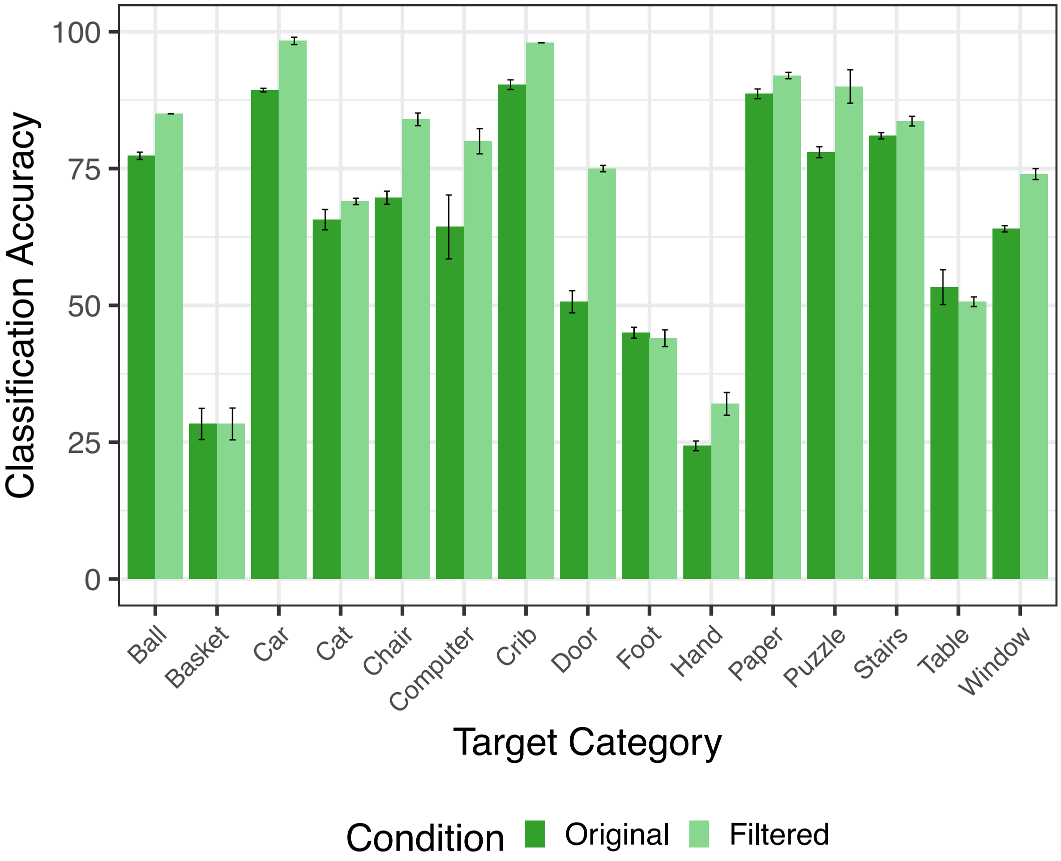

Figure S.3: **Comparison between CVCL by category on a subset of clean images from Labeled-S**. We generated a second version of the Labeled-S evaluation dataset, using a subset of non-overlapping categories (15 out of the 22 concepts), and manually filtering the dataset from ∼60K to ∼2.4K images to ensure they contained the target referents. Our results show that for the concepts that CVCL had originally acquired, performance on this clean evaluation dataset led to even stronger performance (*M*=72.3%), compared to 64.7% on the original evaluation dataset for these 15 concepts. There was no substantial change in performance for the concepts the model had originally failed to acquire.

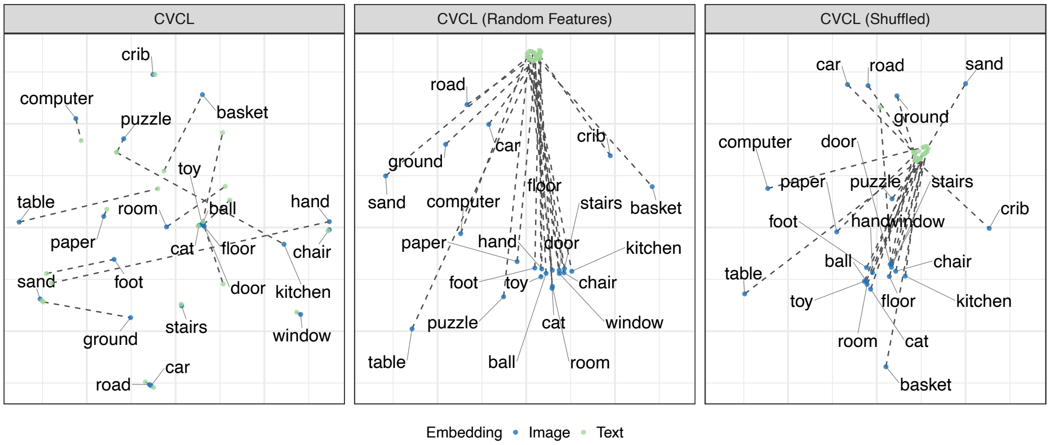

Figure S.4: **Comparison of t-SNE plots derived from the cosine similarities between the mean image embeddings and text embeddings for CVCL and two lower bound variants**. While CVCL's image and text embeddings show conceptual alignment between the visual and linguistic modalities, by computing the correlation between all pairwise cosine similarities across modalities, no conceptual alignment was observed in either the CVCL (Random Features) ($r = -0.002, p = 0.97$) or the CVCL (Shuffled) models ($r = -0.01, p = 0.88$).

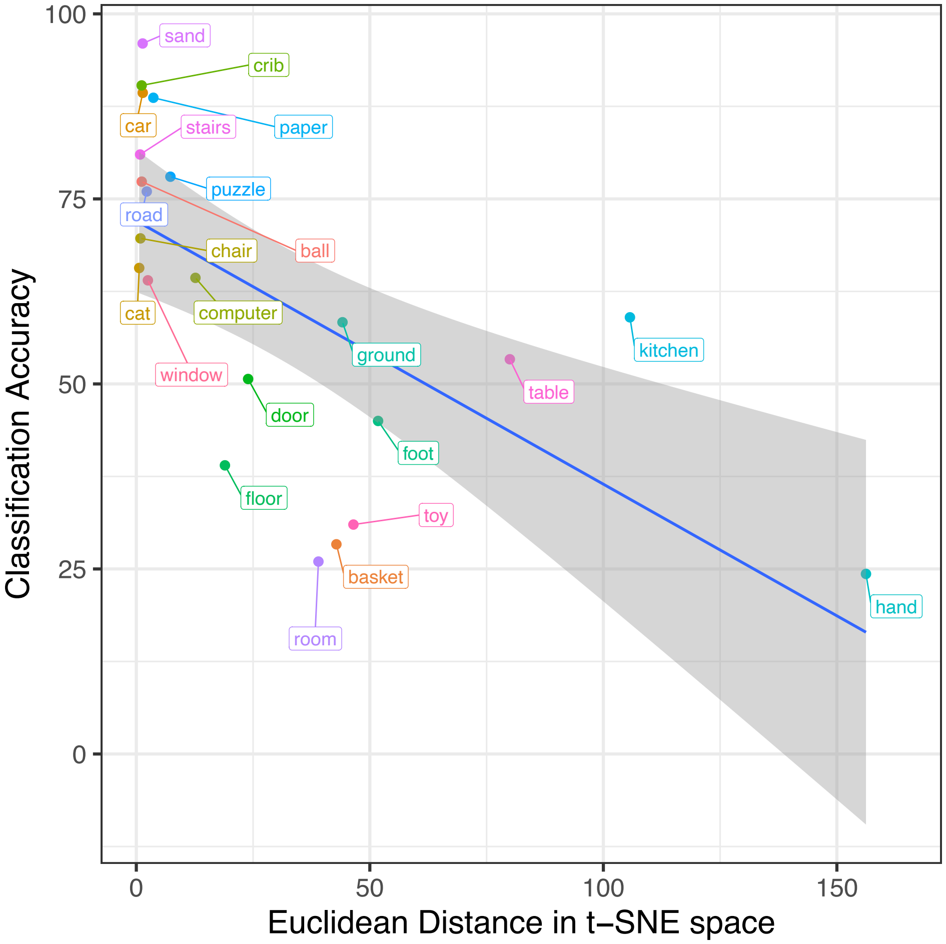

Figure S.5: **Correlation between alignment distance in t-SNE space vs. CVCL's classification performance**. There was a strong negative correlation observed between the alignment distance of concepts (calculated as the Euclidean distance between a concept's word embedding and the mean image embeddings), to its classification performance on the Labeled-S evaluation ($r = -0.65, p = 0.001$), suggesting that concepts whose image and text embeddings were closer to one another were easier to classify. To check for the robustness of this finding, we also performed this analysis in the original embedding space (without t-SNE), which also demonstrated a significant negative correlation ($r = -0.84, p < 0.001$).

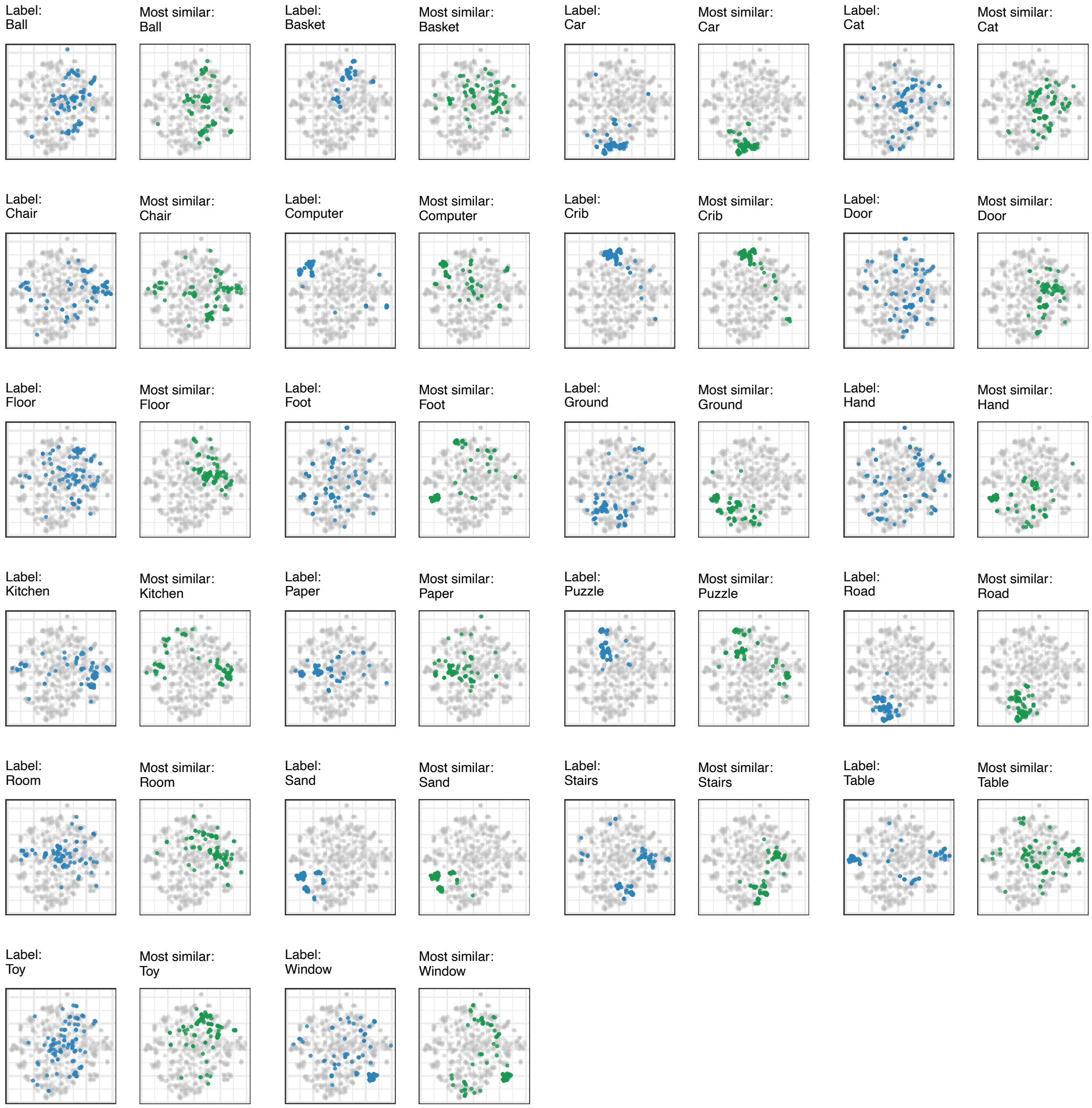

Figure S.6: **Comparison between labeled frames versus most similar frames for all Labeled-S concepts**. In each plot, we visualize a subset of frame embeddings from the Labeled-S evaluation using t-SNE, where the blue points on the left correspond to 100 randomly sampled frames per category, while the green points on the right correspond to the 100 most similar frames (based on the cosine similarity to each concept's word embedding using CVCL), which roughly can be viewed as the extension of the word. Across many concepts, we see a strong overlap between the true labeled points and the points predicted to be most similar from CVCL.

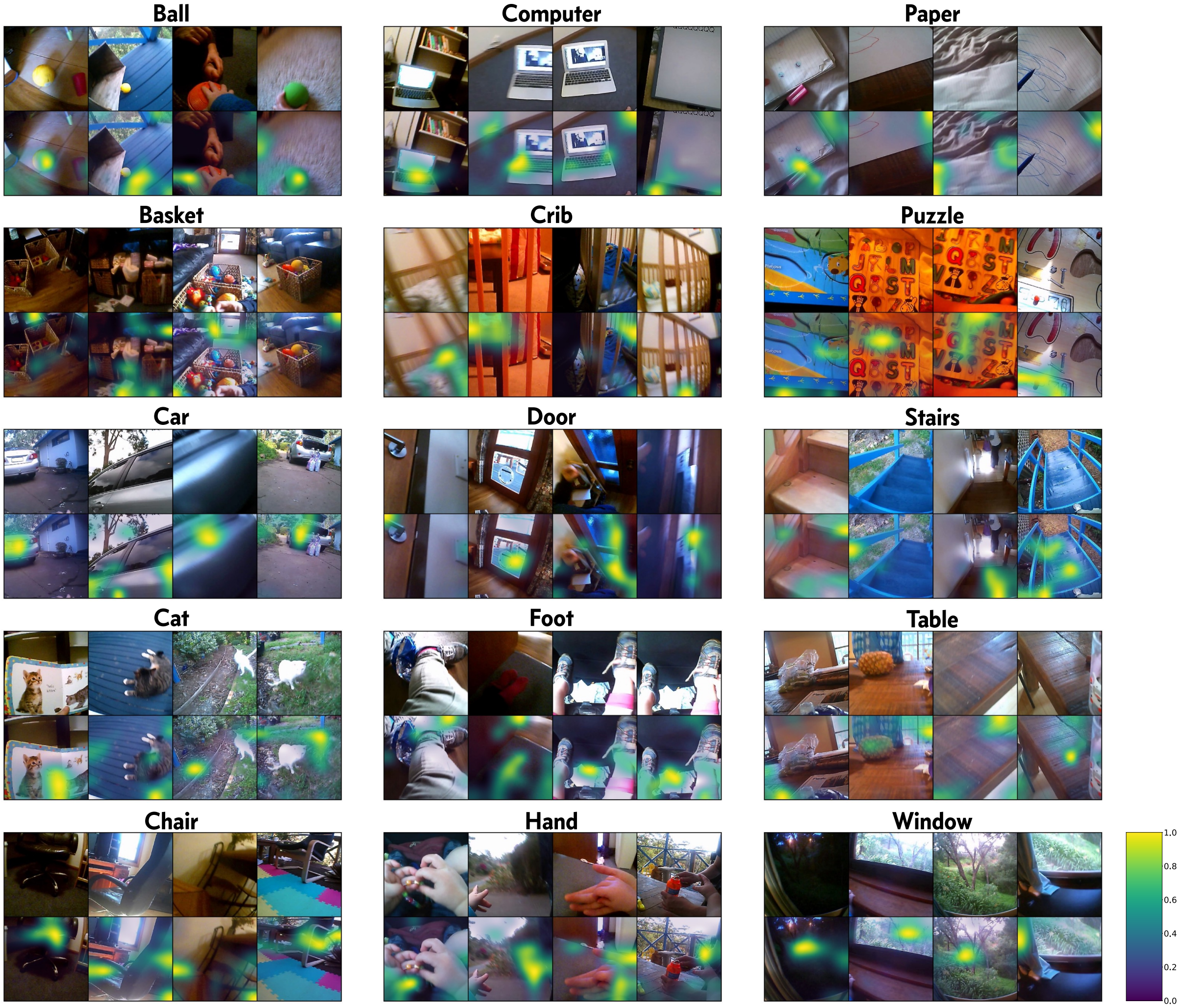

Figure S.7: **Additional attention maps generated via Grad-CAM for fifteen different categories showing object localization capabilities in CVCL.** Each plot contains 4 different examples from a category, with the corresponding normalized attention map below, where yellow indicates regions with the highest attention. Images from each category were randomly selected from the set of manually filtered images from Labeled-S.Across the different categories, we see a mix of positive and negative evidence for CVCL's ability to localize referents within a scene.

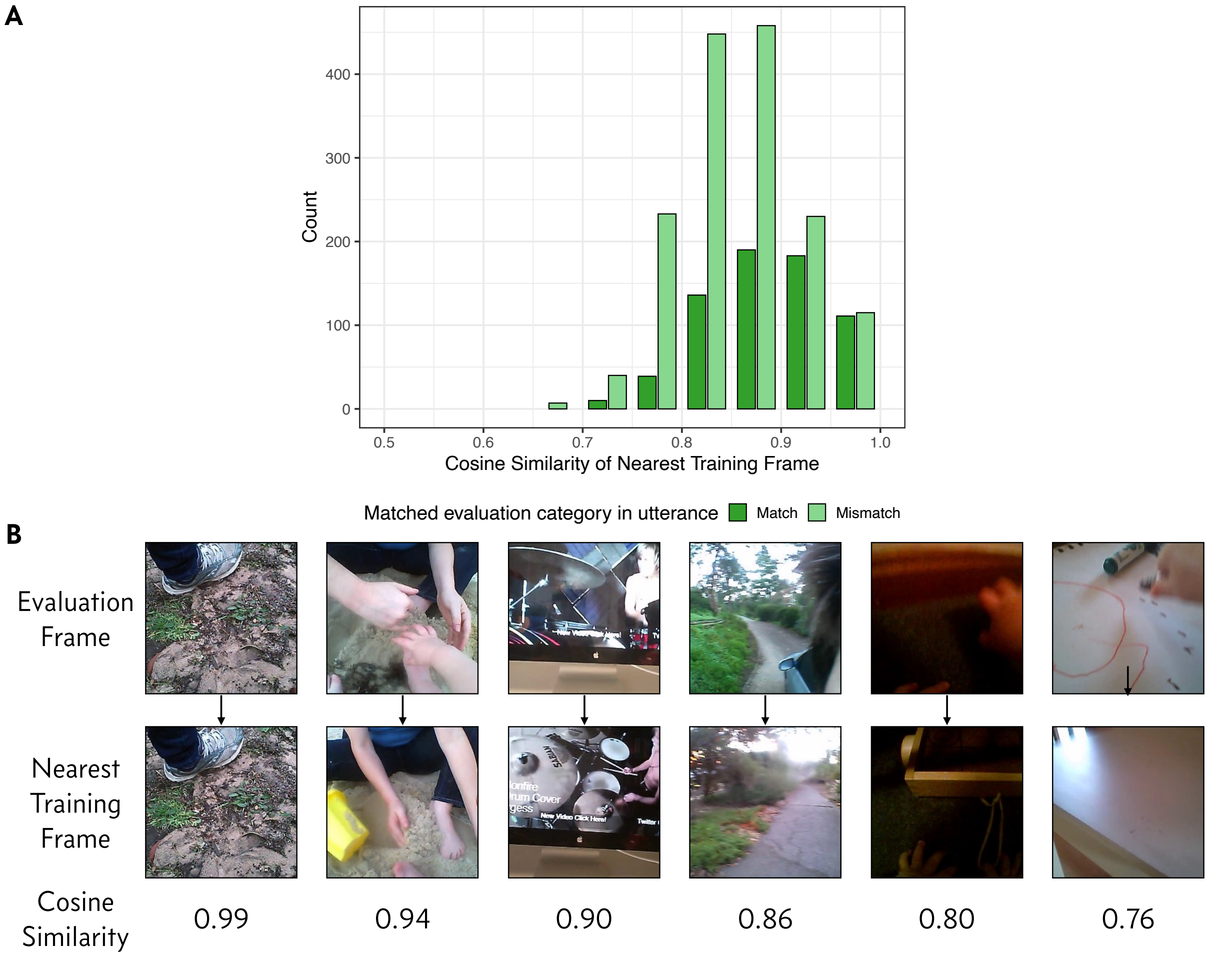

Figure S.8: **Examining dataset overlap between training and evaluation frames.** (A) A histogram of the cosine similarity between each evaluation frame and its corresponding nearest neighbor in the training set (calculated using image embeddings from the self-supervised CNN on frames from baby S (*47*)). Results are split based on whether the training frames' corresponding utterance matched the evaluation category or not. (B) Example evaluation frames and their closest training frames (with matched utterances) for varying levels of cosine similarity. Indirect overlap was considered as matched frames with a cosine similarity score greater than 0.95.

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, At 6-9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3253–3258 (2012). doi:10.1073/pnas.1113380109 Medline

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, Wordbank: An open repository for developmental vocabulary data. *J. Child Lang.* **44**, 677–694 (2017). doi:10.1017/S0305000916000209 Medline

7. T. Regier, The emergence of words: Attentional learning in form and meaning. *Cogn. Sci.* **29**, 819–865 (2005). doi:10.1207/s15516709cog0000_31 Medline

8. E. Colunga, L. B. Smith, From the lexicon to expectations about kinds: A role for associative learning. *Psychol. Rev.* **112**, 347–382 (2005). doi:10.1037/0033-295X.112.2.347 Medline

9. C. Yu, L. B. Smith, Rapid word learning under uncertainty via cross-situational statistics. *Psychol. Sci.* **18**, 414–420 (2007). doi:10.1111/j.1467-9280.2007.01915.x Medline

10. L. B. Smith, S. H. Suanda, C. Yu, The unrealized promise of infant statistical word-referent learning. *Trends Cogn. Sci.* **18**, 251–258 (2014). doi:10.1016/j.tics.2014.02.007 Medline

11. J. Fiser, R. N. Aslin, Statistical learning of new visual feature combinations by infants. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15822–15826 (2002). doi:10.1073/pnas.232472899 Medline

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, Ontological categories guide young children’s inductions of word meaning: Object terms and substance terms. *Cognition* **38**, 179–211 (1991). doi:10.1016/0010-0277(91)90051-5 Medline

15. M. Tomasello, M. J. Farrar, Joint attention and early language. *Child Dev.* **57**, 1454–1463 (1986). doi:10.2307/1130423 Medline

16. D. A. Baldwin, Infants’ contribution to the achievement of joint reference. *Child Dev.* **62**, 875–890 (1991). doi:10.1111/j.1467-8624.1991.tb01577.x Medline

17. M. Bohn, M. H. Tessler, M. Merrick, M. C. Frank, Predicting pragmatic cue integration in adults’ and children’s inferences about novel word meanings. *J. Exp. Psychol. Gen.* **151**, 2927–2942 (2022). doi:10.1037/xge0001216 Medline

18. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015). doi:10.1038/nature14539 Medline

19. D. K. Roy, A. P. Pentland, Learning words from sights and sounds: A computational model. *Cogn. Sci.* **26**, 113–146 (2002). doi:10.1207/s15516709cog2601_4

20. L. Smith, C. Yu, Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* **106**, 1558–1568 (2008). doi:10.1016/j.cognition.2007.06.010 Medline

21. M. C. Frank, N. D. Goodman, J. B. Tenenbaum, Using speakers' referential intentions to model early cross-situational word learning. *Psychol. Sci.* **20**, 578–585 (2009). doi:10.1111/j.1467-9280.2009.02335.x Medline

22. A. Fazly, A. Alishahi, S. Stevenson, A probabilistic computational model of cross-situational word learning. *Cogn. Sci.* **34**, 1017–1063 (2010). doi:10.1111/j.1551-6709.2010.01104.x Medline

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, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] (2021).

26. M. Nikolaus, A. Alishahi, G. Chrupała, Learning English with Peppa Pig. *Trans. Assoc. Comput. Linguist.* **10**, 922–936 (2022). doi:10.1162/tacl_a_00498

27. J. Sullivan, M. Mei, A. Perfors, E. Wojcik, M. C. Frank, SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective. *Open Mind (Camb.)* **5**, 20–29 (2021). doi:10.1162/opmi_a_00039 Medline

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, Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition* **173**, 43–59 (2018). doi:10.1016/j.cognition.2017.11.008 Medline

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, in *Methods for Assessing Children's Syntax*, D. McDaniel, C. McKee, H. S. Cairns, Ed. (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, Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol. Gen.* **139**, 558–578 (2010). doi:10.1037/a0019165 Medline

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, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision" in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021), pp. 4904–4916.

36. B. D. Roads, B. C. Love, Learning as the unsupervised alignment of conceptual systems. *Nat. Mach. Intell.* **2**, 76–82 (2020). doi:10.1038/s42256-019-0132-2

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, The adaptive nature of human categorization. *Psychol. Rev.* **98**, 409–429 (1991). doi:10.1037/0033-295X.98.3.409

40. B. C. Love, D. L. Medin, T. M. Gureckis, SUSTAIN: A network model of category learning. *Psychol. Rev.* **111**, 309–332 (2004). doi:10.1037/0033-295X.111.2.309 Medline

41. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization" in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 618–626.

42. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a Visual Language Model for Few-Shot Learning. *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, Variability, negative evidence, and the acquisition of verb argument constructions. *J. Child Lang.* **37**, 607–642 (2010). doi:10.1017/S0305000910000012 Medline

45. B. C. Roy, M. C. Frank, P. DeCamp, M. Miller, D. Roy, Predicting the birth of a spoken word. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12663–12668 (2015). doi:10.1073/pnas.1419773112 Medline

46. W. Wang, W. K. Vong, N. Kim, B. M. Lake, Finding Structure in One Child's Linguistic Experience. *Cogn. Sci.* **47**, e13305 (2023). doi:10.1111/cogs.13305 Medline

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, Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychol. Rev.* **119**, 831–877 (2012). doi:10.1037/a0029872 Medline

49. J. L. Elman, Finding Structure in Time. *Cogn. Sci.* **14**, 179–211 (1990). doi:10.1207/s15516709cog1402_1

50. T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997). doi:10.1037/0033-295X.104.2.211

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, Learning Through Processing: Toward an Integrated Approach to Early Word Learning. *Annu. Rev. Linguist.* **8**, 77–99 (2022). doi:10.1146/annurev-linguistics-031220-011146 Medline

56. B. Landau, L. B. Smith, S. S. Jones, The importance of shape in early lexical learning. *Cogn. Dev.* **3**, 299–321 (1988). doi:10.1016/0885-2014(88)90014-7

57. L. Gleitman, The Structural Sources of Verb Meanings. *Lang. Acquis.* **1**, 3–55 (1990). doi:10.1207/s15327817la0101_2

58. J. C. Trueswell, T. N. Medina, A. Hafri, L. R. Gleitman, Propose but verify: Fast mapping meets cross-situational word learning. *Cogn. Psychol.* **66**, 126–156 (2013). doi:10.1016/j.cogpsych.2012.10.001 Medline

59. K. Gulordava, T. Brochhagen, G. Boleda, “Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks” in *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, pp. 2089–2095.

60. R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, “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, Episodes of experience and generative intelligence. *Trends Cogn. Sci.* **26**, 1064–1065 (2022). doi:10.1016/j.tics.2022.09.012 Medline

62. T. M. Gureckis, D. B. Markant, Self-Directed Learning. *Perspect. Psychol. Sci.* **7**, 464–481 (2012). doi:10.1177/1745691612454304 Medline

63. W. K. Vong, B. M. Lake, Cross-Situational Word Learning With Multimodal Neural Networks. *Cogn. Sci.* **46**, e13122 (2022). doi:10.1111/cogs.13122 Medline

64. E. H. Wojcik, M. Zettersten, V. L. Benitez, The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdiscip. Rev. Cogn. Sci.* **13**, e1596 (2022). doi:10.1002/wcs.1596 Medline

65. B. M. Lake, G. L. Murphy, Word meaning in minds and machines. *Psychol. Rev.* **130**, 401–431 (2023). doi:10.1037/rev0000297

66. J. Gilkerson, J. A. Richards, S. F. Warren, J. K. Montgomery, C. R. Greenwood, D. Kimbrough Oller, J. H. L. Hansen, T. D. Paul, Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *Am. J. Speech Lang. Pathol.* **26**, 248–265 (2017). doi:10.1044/2016_AJSLP-15-0169 Medline

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.

68. S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1492–1500.

69. M. Caron, H. Touvron, I. Misra, J. Jégou, M. Mairal, P. Bojanowski, A. Joulin, “Emerging Properties in Self-Supervised Vision Transformers” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9650–9660.

70. A. d. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG] (2018).

71. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).

72. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997). doi:10.1162/neco.1997.9.8.1735 Medline

73. I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] (2017).

74. Although this interpolation procedure did not lead to timestamps that were exactly matched with each of the spoken utterances, the relative stability of visual information across short periods of time (seconds to minutes) meant that this was not a major issue.

75. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick,, in *European Conference on Computer Vision* (Springer, 2014), pp. 740–755.

76. B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk* (Psychology Press, 2014).

77. E. M. Clerkin, L. B. Smith, Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2123239119 (2022). doi:10.1073/pnas.2123239119 Medline

78. C. M. Fausey, S. Jayaraman, L. B. Smith, From faces to hands: Changing visual input in the first two years. *Cognition* **152**, 101–107 (2016). doi:10.1016/j.cognition.2016.03.005 Medline

79. We used the original category labels, except the category "cat", which we replaced with the label "kitty", because that matched the word used in the training dataset to refer to cats.