Early Visual Degradation Facilitates Few-Shot Word Learning from a Child's Egocentric Input

Jeonghwan Cheon[1] **Marin Vogelsang**[2] **Lukas Vogelsang**[2] **Pawan Sinha**[2]
[1]Independent Researcher, Seoul, Republic of Korea,
[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
{cheon, ozaki, lvogelsa, psinha}@mit.edu

Abstract

Children learn words from continuous egocentric experience while their visual system is still developing. Recent work has demonstrated that neural network models can learn word–referent mappings from longitudinal head-mounted camera recordings of a single child, but it assumed adult-like visual input quality. Here, we show that incorporating developmentally motivated early visual degradations facilitates word grounding from naturalistic egocentric input. Specifically, we trained models on either full-fidelity inputs throughout or on initially blurred and color-depleted inputs during early training phases. Our results reveal that the developmentally inspired model learns more compact visual representations and outperforms the standard one on a novel word-learning task based on few-shot visual examples. The benefits of early visual degradations are particularly pronounced for basic-level categories and strengthen shape biases akin to those observed in infant learning. These findings suggest that early visual degradations can provide an inductive bias for sample-efficient grounded word learning.

1 Introduction

Children learn object words by linking snippets of speech to referents in a continuous, cluttered visual world, and by extending those mappings beyond individual exemplars. This learning process is facilitated by children's actions and attentional mechanisms that create moments in which certain visual objects are particularly dominant [1, 2]. By two years of age, children typically know about 300 words, many of which refer to common objects. The mechanisms underlying word learning have long been debated, particularly whether certain innate processes are required to acquire this important capacity or whether associative learning from a child's visual and linguistic experience is sufficient to do so. Longitudinal head-camera corpora capturing this experience now make it possible to test these mechanisms [3]. Recently, the Child's View for Contrastive Learning (CVCL) model showed that word–referent mappings can indeed be learned from the egocentric experience of a single child [4].

However, unlike the head-mounted camera videos used for CVCL, which exclusively capture high-fidelity inputs, children's visual inputs are initially severely degraded, for instance in visual acuity and chromatic content, and only improve gradually over the developmental time course [5–7]. Notably, the Adaptive Initial Degradation (AID) hypothesis has recently proposed that these transient degradations have functional significance [8]. Specifically, incorporating initial visual degradations into model training has been shown to facilitate holistic processing [9, 10], resilience to color shifts and removal [11], robustness to perturbations [12] and shape-based object recognition [12–14]. In addition, low-pass input improves basic-level categorization in computational models [15, 16]. Together, these studies suggest that AID can facilitate the development of useful visual representations. Against this background, we ask whether the benefits of AID also extend to word grounding in a naturalistic setting based on a child's egocentric visual experience.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Developmental Perspectives on AI.

Our results reveal that AID can indeed serve as an inductive bias for such word grounding. We trained matched visual encoders on SAYCam [3] either with or without an initial phase of blurred and color-depleted input during self-supervision, and then trained a contrastive vision–language model on the resulting visual representations. Notably, the developmentally inspired model that includes an initial degradation phase outperforms the standard model on a few-shot novel-word learning task. AID's few-shot benefit is most pronounced at the basic level, and AID shifts one-shot word generalization toward shape, similar to biases reported in infant word learning [17, 18]. Together, these results identify developmentally ordered perceptual input as an inductive bias that facilitates grounding novel words in visual concepts.

2 Results

2.1 Adaptive initial degradation improves few-shot novel-word learning

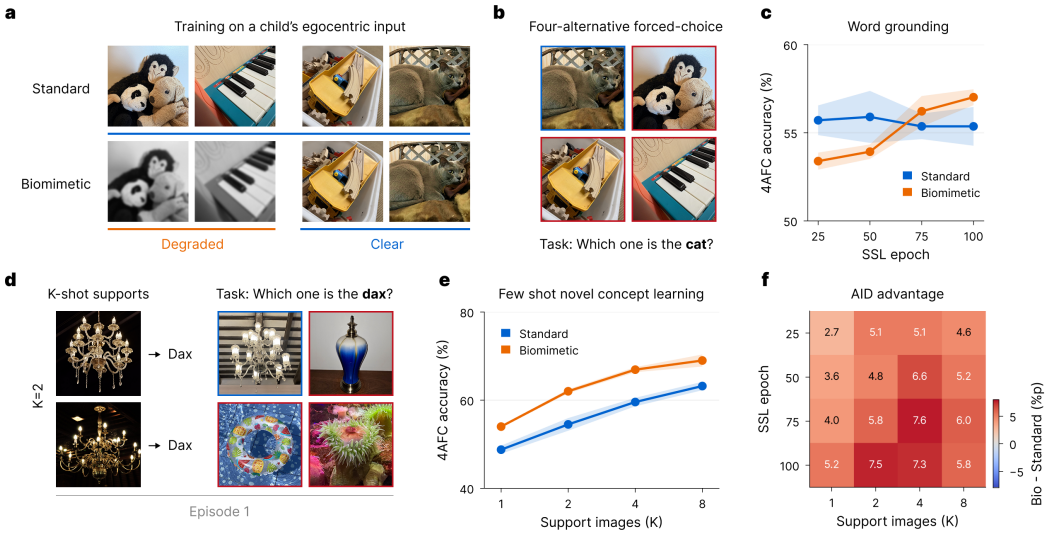


Figure 1: **Adaptive initial degradation improves few-shot novel-word learning. a,** Standard and biomimetic curricula for learning from a child's egocentric input. The biomimetic curriculum transitions from blurred, grayscale to clear, full-color input. **b,** 4AFC task on the SAYCam-S evaluation set. Images are illustrative and were captured for this figure rather than taken from the original SAYCam dataset. **c,** 4AFC accuracy across visual checkpoints. **d,** Few-shot novel-word learning task. The task is illustrated with CC0-licensed WordPress images replacing the originals. **e,** Few-shot learning accuracy across support sizes. **f,** Biomimetic–Standard differences across checkpoints and support sizes. Shaded areas indicate means ± 1 SD across five seeds.

We compared full-fidelity training (*standard*) with a training curriculum incorporating initial visual degradations (*biomimetic* or *AID*) on the same SAYCam stream [3, 4]. Inspired by prior developmental curricula with initially degraded input [9–13, 16, 19, 20], the biomimetic model received blurred and grayscale input during the first half of visual self-supervision and then received the same clear input as the standard model in the second half (Figure 1a). After visual training, we trained both standard and biomimetic models using vision–language contrastive learning following CVCL [4] and evaluated the word-grounding performance of each model using a four-alternative forced-choice (4AFC) task on the SAYCam-S evaluation set (Figure 1b). Although the biomimetic model initially trailed the standard model in word grounding, it surpassed the standard model after clear-input training (Figure 1c).

Next, we investigated the models' ability to learn novel words from few-shot examples (Figure 1d). Specifically, each model learned a novel pseudoword (e.g., *dax*) from K visual examples, and we evaluated subsequent word grounding using a 4AFC paradigm. Notably, the biomimetic model consistently outperformed the standard model on the few-shot novel-word learning task across different numbers of support images (Figure 1e). This benefit generalized across visual checkpoints and support sizes (Figure 1f). Taken together, AID facilitates word grounding, particularly for the rapid formation of novel visual–lexical associations from limited visual examples.

2.2 Early degradation compresses visual representations

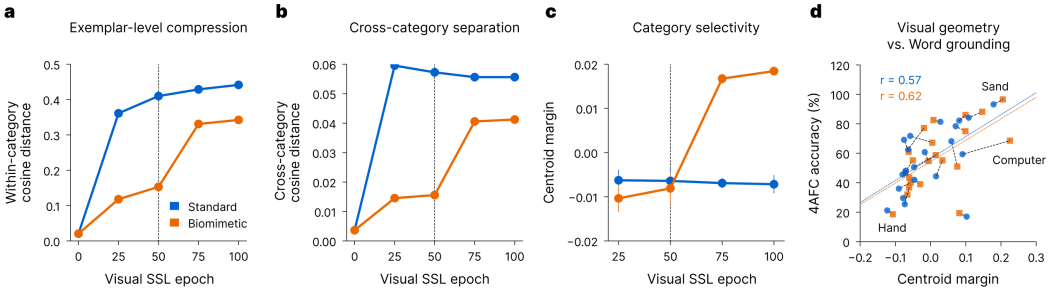


Figure 2: **Early degradation compresses visual representations and enhances category selectivity. a,** Within-category cosine distance; lower values indicate greater compression. **b,** Between-centroid cosine distance. **c,** Centroid margin (target minus closest competitor) in CVCL embeddings. **d,** Relationship between centroid margin and 4AFC accuracy at the final checkpoint.

Next, we further investigated how biomimetic training changes visual representations. Before vision–language alignment, we found that the AID model produced tighter within-category representations during the degradation phase (Figure 2a). Category centroids also moved closer together (Figure 2b), indicating broad compression rather than cleaner categorical separation. After vision–language alignment, the biomimetic model's embeddings separated target queries more strongly from their closest competing categories, largely by reducing competitor similarity (Figure 2c). Centroid margin also tracked word-grounding accuracy across categories within each condition (Figure 2d). Together, AID produced a distinctive representational profile characterized by compressed visual representations followed by stronger query-relative selectivity after language alignment.

2.3 AID's hierarchical benefit is most consistent at the basic level

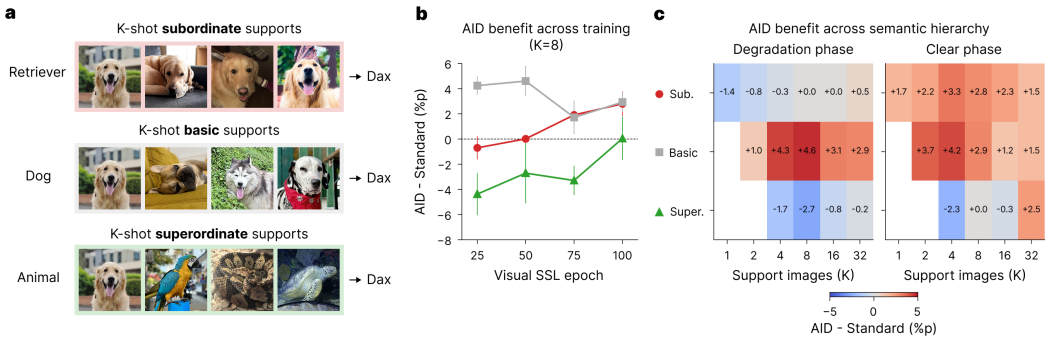


Figure 3: **Few-shot novel-word learning across different levels of abstraction. a,** Subordinate-, basic-, and superordinate-level novel-word tasks. The example is illustrated using CC0-licensed WordPress images in place of the originals. **b,** Biomimetic–Standard differences in 4AFC accuracy across visual checkpoints at $K = 8$. **c,** Differences across hierarchy levels and support sizes at the end of degradation and after clear training.

We then asked whether the AID-related benefits for few-shot word grounding reflect inductive biases previously reported in infant learning, such as a basic-level advantage or shape bias. First, we examined whether the novel-word benefit depended on the level of abstraction. We assigned pseudowords to K-shot visual examples representing subordinate, basic, or superordinate concepts, with higher-level queries drawn from hierarchy branches absent from the support set (Figure 3a). Notably, initial degradation most consistently benefited basic-level concept learning, and this advantage persisted during the subsequent clear-input phase (Figure 3b,c). Benefits for subordinate-level concept learning emerged only during the later clear-input phase. Benefits at the superordinate level also emerged later in training, but only under conditions with larger support sets. Together, these results suggest that AID preferentially supports basic-level novel-word learning, with benefits extending to subordinate concepts later in training and to superordinate concepts under more data-rich conditions.

2.4 AID shifts one-shot word generalization toward shape

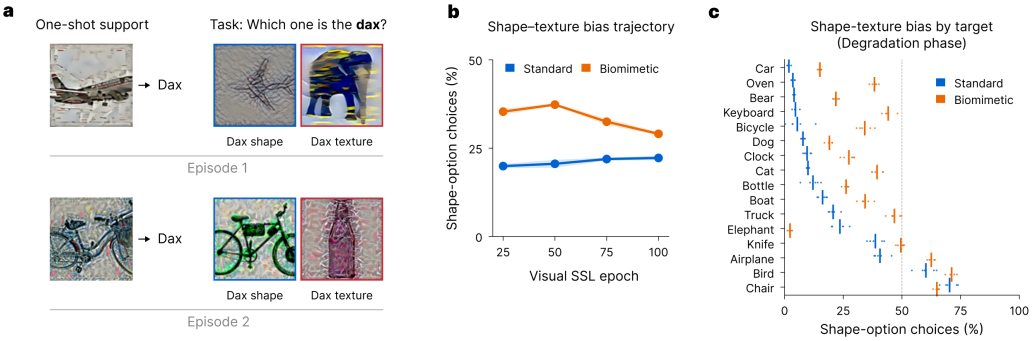


Figure 4: **Shape–texture bias in a few-shot word-learning task. a,** One-shot novel-word task using shape–texture cue conflicts. The support example contains consistent shape and texture cues, and the 2AFC task evaluates which cue the model learns to associate with the novel word. **b,** Shape-choice rates across visual checkpoints. Values above 50% indicate a predominant shape bias, whereas values below 50% indicate a predominant texture bias. **c,** Shape choices by target.

Finally, we investigated whether our trained models rely more on shape-based or texture-based cues when learning novel concepts. Each model learned a pseudoword from a single cue-consistent example containing both shape and texture information. We then presented reciprocal, source-disjoint cue-conflict options that directly contrasted the two cues, allowing us to evaluate each model's relative bias toward shape or texture (Figure 4a). Our biomimetic model selected the shape-consistent object more often than the standard model across visual self-supervised learning checkpoints (Figure 4b). This difference was largest during the degradation phase and appeared across most target categories (Figure 4c). These results suggest that initial degradation shifts novel-word generalization toward shape-based cues relative to texture-based cues. However, although the biomimetic model exhibited a stronger relative shape bias, both models still favored texture overall.

3 Discussion

Here, we incorporated a developmentally motivated visual trajectory into word grounding from the egocentric experience of a single child. Building on previous computational studies that model word grounding from a child's egocentric view [4], we applied adaptive initial degradation during visual learning. Our findings extend previously reported benefits of initial degradations in the visual domain [8, 9, 11–14, 19, 20] to grounded lexical learning by showing that developmentally ordered visual experience can serve as an inductive bias that facilitates few-shot word grounding.

These results align with two patterns observed in developmental psychology. First, our biomimetic model most consistently favored basic-level novel-word learning, consistent with the privileged status of basic categories in perception and early acquisition [17, 21]. By balancing within-category similarity with between-category distinctiveness, basic-level categories are thought to support efficient naming and generalization. Second, AID shifted one-shot word generalization toward shape, consistent with shape-directed effects of developmental visual curricula [12–14]. Children likewise tend to generalize novel object names by shape [18], and this tendency strengthens with early noun learning [22, 23]. Together, the basic-level benefit and shape-directed shift bring the model closer to qualitative signatures of early word learning.

Although our results provide insights into word learning during development, there are several important limitations. First, our protocol is not a chronological simulation of infant learning. Visual representations were learned through self-supervision before language grounding, with a separate CVCL model trained on the frozen representations from each checkpoint. In infants, perceptual development and word learning instead unfold alongside action and attention, with each process structuring subsequent input [1, 2]. In addition, we used a simplified two-stage curriculum, which does not fully reproduce the developmental trajectories of visual acuity and color sensitivity. AID should therefore be viewed as a useful perceptual inductive bias rather than a complete account of infant learning. Future work should interleave perceptual and linguistic learning [24] and test these trajectories across children and independent pretraining runs using richer naturalistic streams [25, 26].

Acknowledgments and Disclosure of Funding

The authors acknowledge the MIT Office of Research Computing and Data for providing high performance computing resources that have contributed to the research results reported within this paper. Marin Vogelsang is supported by the Japan Society for the Promotion of Science (JSPS) Overseas Research Fellowship and the Yamada Science Foundation, and Lukas Vogelsang is supported by a grant from the Simons Foundation International to the Simons Center for the Social Brain at MIT.

References

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

[2] Smith, L. B., Jayaraman, S., Clerkin, E. & Yu, C. The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences* **22**, 325–336 (2018).

[3] Sullivan, J., Mei, M., Perfors, A., Wojcik, E. & Frank, M. C. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind* **5**, 20–29 (2021).

[4] Vong, W. K., Wang, W., Orhan, A. E. & Lake, B. M. Grounded language acquisition through the eyes and ears of a single child. *Science* **383**, 504–511 (2024).

[5] Atkinson, J., Braddick, O. & Moar, K. Development of contrast sensitivity over the first 3 months of life in the human infant. *Vision Research* **17**, 1037–1044 (1977).

[6] Morrone, M. C., Burr, C. D. & Fiorentini, A. Development of contrast sensitivity and acuity of the infant colour system. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **242**, 134–139 (1990).

[7] Braddick, O. & Atkinson, J. Development of human visual function. *Vision Research* **51**, 1588–1609 (2011).

[8] Vogelsang, L., Vogelsang, M., Pipa, G., Diamond, S. & Sinha, P. Butterfly effects in perceptual development: A review of the 'adaptive initial degradation'hypothesis. *Developmental Review* **71**, 101117 (2024).

[9] Vogelsang, L. *et al.* Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences* **115**, 11333–11338 (2018).

[10] Jang, H. & Tong, F. Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of Vision* **21**, 6–6 (2021).

[11] Vogelsang, M. *et al.* Impact of early visual experience on later usage of color cues. *Science* **384**, 907–912 (2024).

[12] Cai, Y., Lin, Q., Nunna, B. S. & Zhang, M. Learning to see through a baby's eyes: Early visual diets enable robust visual intelligence in humans and machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13766–13780 (2026).

[13] Vogelsang, M., Vogelsang, L., Pipa, G., Diamond, S. & Sinha, P. Potential role of developmental experience in the emergence of the parvo-magno distinction. *Communications Biology* **8**, 987 (2025).

[14] Lu, Z., Thorat, S., Cichy, R. M. & Kietzmann, T. C. Adopting a human developmental visual diet yields robust and shape-based AI vision. *Nature Machine Intelligence* 1–14 (2026).

[15] French, R. M., Mermillod, M., Chauvin, A., Quinn, P. C. & Mareschal, D. The importance of starting blurry: Simulating improved basic-level category learning in infants due to weak visual acuity. In Gray, W. D. & Schunn, C. D. (eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (Lawrence Erlbaum Associates, 2002).

[16] Jinsi, O., Henderson, M. M. & Tarr, M. J. Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *Plos One* **18**, e0280145 (2023).

[17] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology* **8**, 382–439 (1976).

[18] Landau, B., Smith, L. B. & Jones, S. S. The importance of shape in early lexical learning. *Cognitive Development* **3**, 299–321 (1988).

[19] Cheon, J. & Paik, S.-B. Gradual sensory maturation promotes abstract representation learning. *bioRxiv* 2025–06 (2025).

[20] Vogelsang, M., Vogelsang, L. & Sinha, P. Computationally probing the role of time-limited neuronal plasticity in early visual development. *Developmental Science* **29**, e70099 (2026).

[21] Mervis, C. B. & Crisafi, M. A. Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development* **53**, 258–266 (1982).

[22] Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L. & Samuelson, L. Object name learning provides on-the-job training for attention. *Psychological Science* **13**, 13–19 (2002).

[23] Gershkoff-Stowe, L. & Smith, L. B. Shape and the first hundred nouns. *Child Development* **75**, 1098–1114 (2004).

[24] Schaumlöffel, T., Aubret, A., Roig, G. & Triesch, J. Caregiver talk shapes toddler vision: A computational study of dyadic play. In *2023 IEEE International Conference on Development and Learning (ICDL)*, 67–72 (IEEE, 2023).

[25] Long, B. L. *et al.* The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. In *8th Annual Conference on Cognitive Computational Neuroscience* (2025).

[26] Lin, D. *et al.* EgoBabyVLM: Benchmarking cross-modal learning from naturalistic egocentric video data. *arXiv preprint arXiv:2605.19130* (2026).

[27] Konkle, T., Brady, T. F., Alvarez, G. A. & Oliva, A. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General* **139**, 558 (2010).

[28] Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640 (IEEE, 2021).

[29] Hebart, M. N. *et al.* THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS One* **14**, e0223792 (2019).

[30] Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).

[31] Miller, G. A. WordNet: a lexical database for english. *Communications of the ACM* **38**, 39–41 (1995).

[32] Muttenthaler, L. *et al.* Aligning machine and human visual representations across abstraction levels. *Nature* **647**, 349–355 (2025).

[33] Geirhos, R. *et al.* ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations* (2019).

A Supplementary results

A.1 Category-level word grounding on the SAYCam-S evaluation set

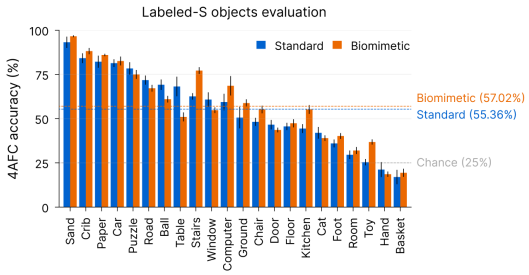


Figure S1: **Category-wise evaluation of word grounding on the SAYCam-S set.** Accuracy is measured by a 4AFC task across 22 different object categories. The orange and blue dashed lines indicate the mean 4AFC performance of the biomimetic and standard models, respectively. The gray dashed line indicates chance performance (25%).

In addition to the averaged overall word grounding performance visualized in Figure 1c, we report category-wise grounding performance (Figure S1). Based on the category means across five seeds, AID outperformed Standard in 14 categories. The largest gains were observed for *stairs*, *toy*, *kitchen*, *computer*, and *ground*, whereas performance for *table*, *ball*, *window*, and several other categories remained lower under AID. Thus, the late aggregate advantage reflected a broad but heterogeneous shift across familiar categories rather than a gain confined to a single target.

A.2 Familiar-word grounding with novel exemplars

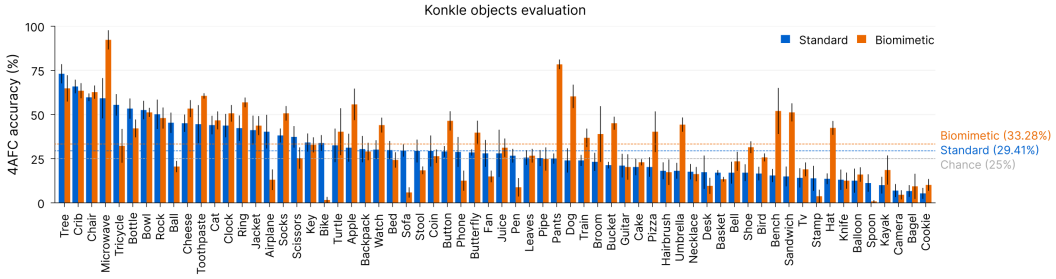


Figure S2: **Category-wise evaluation of word grounding on the Konkle Objects evaluation dataset.** Accuracy is measured using a 4AFC task across 64 different object categories. The orange and blue dashed lines indicate the mean 4AFC performance of the biomimetic and standard models, respectively. The gray dashed line indicates chance performance (25%).

To test whether the benefit of AID for word grounding presented in Figure 1 generalizes beyond SAYCam-S evaluation set, we evaluated word grounding on independently collected images from the Konkle object database [27]. AID achieved higher mean 4AFC accuracy than Standard (Figure S2). This result on novel exemplars indicates that the benefits of AID generalize to out-of-distribution images.

A.3 Few-shot novel-word learning under joint lexical competition

The main analysis (Figure 1e) learned each novel word independently from positive examples. As a control, we instead optimized all 64 lexical prototypes jointly, allowing the other novel concepts to compete during adaptation (Figure S3). AID yielded higher mean accuracy in every checkpoint-by-support-size comparison. These results suggest that the benefit of AID for few-shot word grounding is not limited to learning a single object concept at a time, but generalizes to learning multiple concepts under lexical competition, a setting that more closely resembles naturalistic word learning.

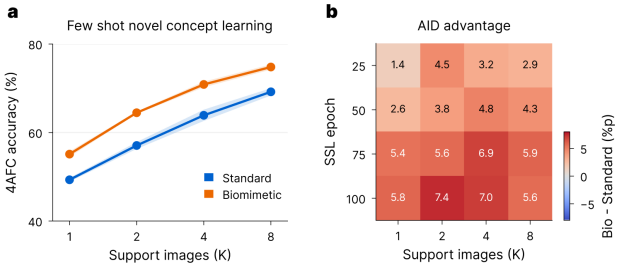


Figure S3: **Adaptive initial degradation improves few-shot novel-word learning under joint lexical competition. a,** Few-shot learning accuracy across support sizes. **b,** Bio–Standard differences across checkpoints and support sizes. Shaded areas indicate means ± 1 SD across five seeds.

A.4 Persistence of the shape shift after clear training

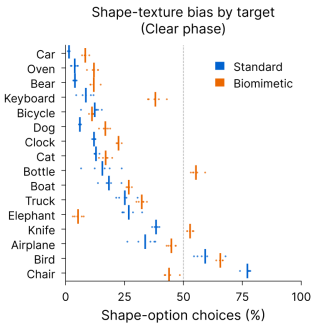


Figure S4: **Shape–texture bias in a few-shot word-learning task after clear phase.** Shape choices by target at the final checkpoint after the clear phase in biomimetic models. Points and bands indicate means ± 1 SD across five seeds.

The shape-texture bias evaluation across object categories in Figure 4c was conducted after the first half of training, when the biomimetic model had just completed the degradation phase. We additionally evaluated the final checkpoint, at which the biomimetic model had completed the subsequent clear phase (Figure S4). The target-level pattern observed in the main 2AFC analysis persisted after clear training. AID produced more shape-consistent choices for 13 of 16 targets and increased the category-averaged shape-choice rate from 22.24% to 29.06% (Table S5). Both conditions remained texture-dominant, indicating a durable relative shift in cue weighting.

A.5 Shape–texture preferences with non-cue response options

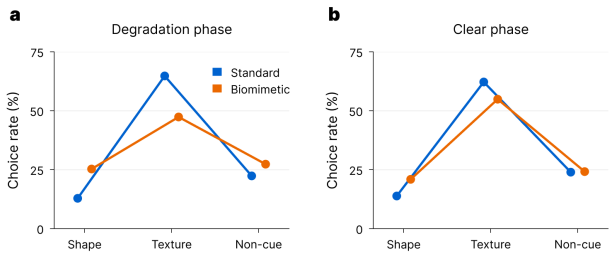


Figure S5: **Raw response count of model decision across shape-, texture-cue and non-cue. a,** Decision fraction after the first half of training, where the biomimetic model finishes the degradation phase. **b,** Decision fraction after full training, where the biomimetic model finishes the clear phase.

The main word-to-image 2AFC evaluation in Figure 4 necessarily assigns every response to either shape or texture. In a complementary image-to-word control, we instead compared each cue-conflict image against four one-shot category prototypes: its shape category, its texture category, and two balanced categories matching neither cue (Figure S5). The pattern observed in this 4AFC evaluation was consistent with that of the 2AFC evaluation presented in the main results: AID increased shape choices and reduced texture choices, while the two conditions showed nearly identical non-cue choice rates.

B Numerical values for main results

Table S1: Word grounding accuracy on the SAYCam-S evaluation set (complement to Figure 1c). Values are mean accuracies (%) $\pm$ 1 sample SD across five downstream CVCL seeds. Δ denotes the mean paired AID–Standard difference in percentage points (pp).

Epoch	Standard	AID	Δ (pp)
25	55.71 ± 0.84	53.39 ± 0.48	-2.32
50	55.90 ± 1.47	53.93 ± 0.41	-1.97
75	55.36 ± 0.72	56.22 ± 0.87	$+0.85$
100	55.36 ± 1.10	57.02 ± 0.45	$+1.65$

Table S2: Few-shot word grounding accuracy (complement to Figure 1e,f). Values are mean accuracies (%) $\pm$ 1 sample SD across five downstream CVCL seeds. K denotes the number of support examples. Δ denotes the mean paired AID–Standard difference in percentage points (pp).

Epoch	K	Standard	AID	Δ (pp)
25	1	50.94 ± 0.54	53.63 ± 0.40	$+2.69$
25	2	54.94 ± 0.52	60.02 ± 0.34	$+5.08$
25	4	60.78 ± 0.81	65.84 ± 0.43	$+5.06$
25	8	64.34 ± 0.39	68.91 ± 0.52	$+4.56$
50	1	48.55 ± 0.56	52.13 ± 0.29	$+3.58$
50	2	55.91 ± 0.99	60.70 ± 0.61	$+4.80$
50	4	59.38 ± 1.12	66.00 ± 0.81	$+6.62$
50	8	63.48 ± 0.64	68.67 ± 0.59	$+5.19$
75	1	49.80 ± 0.49	53.84 ± 0.71	$+4.05$
75	2	55.39 ± 1.04	61.22 ± 0.57	$+5.83$
75	4	59.39 ± 0.85	67.03 ± 1.05	$+7.64$
75	8	63.19 ± 0.75	69.20 ± 1.36	$+6.02$
100	1	48.77 ± 0.91	54.00 ± 0.25	$+5.23$
100	2	54.50 ± 1.43	62.02 ± 0.45	$+7.52$
100	4	59.58 ± 0.96	66.92 ± 0.57	$+7.34$
100	8	63.20 ± 0.96	69.00 ± 1.36	$+5.80$

Table S3: Hierarchy-dependent AID effects on 4AFC accuracy (complement to Figure 3b,c). Values are mean paired AID–Standard differences in percentage points (pp) across five downstream CVCL seeds, using 16 queries per target. Positive values indicate higher accuracy under AID. Dashes indicate unreported combinations.

Epoch	Level	$K = 1$	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$
25	Subordinate	–	–	–	-0.70	–	–
	Basic	–	–	–	$+4.24$	–	–
	Superordinate	–	–	–	-4.38	–	–
50	Subordinate	-1.39	-0.78	-0.30	$+0.01$	$+0.04$	$+0.48$
	Basic	–	$+1.04$	$+4.29$	$+4.60$	$+3.09$	$+2.94$
	Superordinate	–	–	-1.67	-2.71	-0.79	-0.21
75	Subordinate	–	–	–	$+1.91$	–	–
	Basic	–	–	–	$+1.72$	–	–
	Superordinate	–	–	–	-3.29	–	–
100	Subordinate	$+1.68$	$+2.18$	$+3.26$	$+2.79$	$+2.30$	$+1.51$
	Basic	–	$+3.71$	$+4.15$	$+2.93$	$+1.21$	$+1.50$
	Superordinate	–	–	-2.29	$+0.04$	-0.33	$+2.46$

Table S4: One-shot word-to-image 2AFC shape–texture generalization across visual SSL epochs (complement to Figure 4b). Values are mean shape-choice rates (%) $\pm$ 1 sample SD across five downstream CVCL seeds. Δ denotes the mean paired AID–Standard difference $\pm$ 1 sample SD, in percentage points (pp).

Epoch	Standard	AID	Δ (pp)
25	19.93 ± 0.61	35.33 ± 0.82	$+15.40 \pm 0.66$
50	20.58 ± 1.11	37.32 ± 0.39	$+16.73 \pm 0.99$
75	21.92 ± 0.42	32.48 ± 0.85	$+10.56 \pm 0.97$
100	22.24 ± 0.62	29.06 ± 0.57	$+6.82 \pm 1.07$

Table S5: Source-disjoint one-shot word-to-image 2AFC shape–texture generalization (complement to Figure 4c). Values are target-specific mean shape-choice rates (%) across five downstream CVCL seeds at the end of degradation (epoch 50) and after clear training (epoch 100). Δ denotes the mean paired AID–Standard difference in percentage points (pp).

	Epoch 50			Epoch 100		
Item	Standard	AID	Δ (pp)	Standard	AID	Δ (pp)
Car	1.85	15.05	$+13.20$	1.45	8.35	$+6.90$
Oven	3.55	38.29	$+34.74$	3.88	12.04	$+8.15$
Bear	4.05	21.85	$+17.80$	3.96	11.96	$+8.00$
Keyboard	4.64	44.14	$+39.50$	8.59	38.09	$+29.50$
Bicycle	5.44	34.17	$+28.73$	12.37	11.20	-1.17
Dog	7.84	19.10	$+11.26$	6.07	16.87	$+10.80$
Clock	9.61	27.39	$+17.78$	12.06	22.46	$+10.39$
Cat	9.97	39.40	$+29.43$	12.91	17.00	$+4.09$
Bottle	12.08	26.21	$+14.12$	15.61	55.32	$+39.71$
Boat	16.20	34.32	$+18.11$	18.40	26.81	$+8.41$
Truck	20.69	46.80	$+26.10$	25.14	32.33	$+7.19$
Elephant	23.58	2.32	-21.27	26.82	5.29	-21.53
Knife	38.68	49.53	$+10.85$	38.38	52.82	$+14.44$
Airplane	40.65	62.50	$+21.85$	33.77	44.89	$+11.12$
Bird	60.21	71.17	$+10.96$	59.24	65.66	$+6.42$
Chair	70.30	64.88	-5.42	77.15	43.90	-33.25

C Methods

C.1 Experimental overview

We used a two-stage training procedure. First, we pretrained matched Standard and adaptive initial degradation (AID) visual encoders on the complete video stream recorded from child S in SAYCam [3]. Second, we trained a separate Child's View for Contrastive Learning (CVCL) model on each visual checkpoint using the speech-aligned portion of the same child's experience [4]. The visual teacher backbone was retained after logical epochs 25, 50, 75, and 100. For each condition and checkpoint, we trained five independent CVCL models, for a total of 40 vision–language models. Standard and AID used the same training data and evaluation manifests. All stimuli were bicubically resized to 224×224 pixels and embedded by the CVCL visual branch in evaluation mode.

C.2 Egocentric visual input

Visual pretraining used the complete Child-S release, comprising 224 recording sessions and 1,237 videos totaling 201.9 hours. Frames were sampled at a nominal rate of 5 fps. The shorter image dimension was resized to 256 pixels, after which we extracted a 224×224 crop whose center was shifted 16 pixels upward and rotated the result by 180°, following the released preprocessing. This produced 3,209,928 training frames. Frames were drawn from resampled shards and shuffled with a 10,000-sample buffer, so visual pretraining did not preserve the chronological order of the recordings.

C.3 Visual self-supervision

Both conditions used self-distillation with no labels (DINO) [28], implemented with randomly initialized ResNeXt-50 student and teacher encoders. Each frame produced two 224×224 global crops sampled at scales $[0.2, 1.0]$ and eight 96×96 local crops sampled at scales $[0.05, 0.2]$. The student received all crops, and the teacher received the two global crops. A projection head mapped the 2,048-dimensional backbone output through 2,048-dimensional hidden layers and a 256-dimensional bottleneck to 65,536 output units. Training matched centered, sharpened teacher predictions to student predictions across views. Common augmentations included horizontal flipping, color jitter, and solarization of the second global crop. We excluded random grayscale and Gaussian blur from the common augmentations to isolate the effect of adaptive initial degradation.

Each condition was trained in FP32 on four NVIDIA A100 GPUs with a global batch size of 512. Training comprised 330,000 updates, grouped into 100 logical epochs of 3,300 updates. We used AdamW with a constant learning rate of 10^{-4}, zero weight decay, and per-parameter gradient clipping at 1.0. Student and teacher temperatures were 0.1 and 0.04, respectively, and the output center used momentum 0.9. The teacher was updated by exponential moving average using a cosine schedule from base momentum 0.9995 toward 1.0. Downstream analyses used the teacher ResNeXt-50 backbone rather than the DINO projection head.

C.4 Adaptive initial degradation

The conditions differed only in the visual curriculum, which combined low-to-high acuity and grayscale-to-color manipulations used in prior AID simulations [9, 11, 13, 20]. The standard model was trained on clear, full-color images throughout training. The biomimetic model was trained on blurred, grayscale images, generated using Gaussian blur with a fixed radius of 4 and grayscale conversion, during the first 165,000 updates (logical epochs 0–49), and on clear, full-color images during the remaining 165,000 updates (logical epochs 50–99). All other training conditions were identical.

C.5 Vision–language grounding

Language grounding used the released 61-hour speech-aligned SAYCam-S subset, comprising 37,486 transcribed utterances and 600,267 frame references across the original data splits [4]. A 2,350-entry vocabulary, including special tokens, was constructed from the training split by retaining tokens observed at least three times. During training, one frame was sampled from each utterance-aligned frame set and paired with the corresponding transcript; evaluation used the first aligned frame. Start- and end-of-sequence tokens were added before truncation to 25 tokens.

The 2,048-dimensional ResNeXt-50 output was mapped to 512 dimensions by a trainable visual projection. Utterances were represented by averaging learned 512-dimensional embeddings over all nonpadding tokens, including the start and end tokens. Visual and linguistic representations were ℓ_2-normalized and optimized using a symmetric image-to-text and text-to-image contrastive loss with temperature 0.07. Backbone parameters remained fixed, although BatchNorm running statistics continued to update as in the released implementation; only the visual projection and token embeddings were optimized. Each CVCL model was trained for 400 epochs with AdamW (learning rate 10^{-4}, weight decay 0.1, batch size 8) and a reduce-on-plateau schedule (factor 0.1, patience 20). We retained the checkpoint with the lowest validation loss.

C.6 Word grounding evaluation

Word grounding was evaluated on the fixed Labeled-S test split used by CVCL [4]. We excluded *carseat*, *couch*, *greenery*, and *plushanimal* from the 26 released directories, leaving 22 target categories; images were split 50/50 into development and test sets within category. For each target word, the test manifest contained 100 four-alternative forced-choice (4AFC) trials, each pairing one target image with images from three distinct foil categories. Queries consisted of the target word bracketed by start- and end-of-sequence tokens, and the model selected the image with the highest cosine similarity to this embedding. The same 2,200 trials were used for all models, and performance was measured by accuracy (chance $= 25\%$). The training-vocabulary label *kitty* was used for the test category *cat*, following the released evaluation.

C.7 Konkle object generalization

We evaluated categories whose labels appeared in the CVCL vocabulary using the public Konkle object database [27], following the CVCL evaluation [4]. Using 953 images from 64 categories, we generated 4,765 4AFC trials by combining target and foil images. Images were bicubically resized to 224×224 pixels, and each target label was embedded as a single vocabulary token without start- or end-of-sequence tokens, matching the released evaluation command.

C.8 Few-shot novel-word learning

Few-shot learning used naturalistic images from THINGS [29]. We selected 64 concepts—eight concepts from each of eight higher-level groups—with at least 12 images per concept. A concept was included only if none of its name tokens appeared in either the CVCL vocabulary or the SAYCam-S training transcripts. Five deterministic mappings assigned novel consonant–vowel–consonant pseudowords and image orders. For each mapping, nested support sets contained $K \in \{1, 2, 4, 8\}$ of eight designated support images, and four image-disjoint images served as queries. Each query was presented with foils from three other concepts in the same higher-level group.

All existing model parameters were frozen during adaptation. In each logical episode, a single 512-dimensional pseudoword prototype was optimized using only its concept's K positive support embeddings; other concepts, queries, and foils did not enter the loss. The objective was negative mean cosine similarity divided by 0.07, optimized for 500 Adam steps at learning rate 10^{-3} across 20 adaptation runs. This positive-only objective converged toward the normalized support centroid and is distinct from CVCL's contrastive training objective. At test, the adapted prototype was compared with the four candidate embeddings by cosine similarity. We measured 4AFC accuracy and the cosine margin between the target and the most similar foil. Standard and AID used identical support, query, and label-mapping manifests.

For the joint-objective control, all 64 pseudoword prototypes were optimized simultaneously. Each support embedding was classified against the full prototype set using cross-entropy over cosine logits with temperature 0.07. The optimization schedule and test trials matched the independent analysis; only lexical competitors were added to the adaptation loss.

C.9 Representation geometry

Geometry was evaluated on the 6,828 unique images from the 22 Labeled-S categories used by the familiar-word test. Before vision–language alignment, we extracted the 2,048-dimensional global-average-pooled output of each visual teacher backbone, before both the DINO head and CVCL

projection, and ℓ_2-normalized it for cosine analyses. Within-category distance was the exact mean pairwise cosine distance among images in a category. Between-centroid distance was the mean cosine distance from that category's normalized centroid to the other 21 normalized category centroids.

After vision–language alignment, we analyzed the unit-normalized 512-dimensional CVCL visual embeddings. For each held-out fold, category centroids were computed from the other four folds, ensuring that a query image never contributed to its scoring centroid. The centroid margin for a query was its cosine similarity to the target centroid minus its greatest similarity to any non-target centroid. The final-checkpoint association was computed across the 22 category means separately for each condition using Pearson correlation.

C.10 Hierarchical novel-word learning

Hierarchical episodes used ILSVRC-2012 validation images [30] and WordNet 3.0 synsets [31]. We transcribed the basic labels and seven explicit superordinate groupings illustrated in Muttenthaler et al. [32], then linked ImageNet leaves to fixed basic-level WordNet anchors. This produced a grouped taxonomy with seven superordinate, 96 basic, and 385 subordinate categories. The query-expanded analysis retained all 385 subordinate leaves, 23 basic categories with at least three subordinate branches, and six superordinate categories with at least eight basic branches. We retained 48 images per subordinate leaf. Subordinate episodes drew support and query images from the same leaf. Basic-level episodes drew balanced support from two subordinate branches and queries from a held-out third branch. Superordinate episodes drew balanced support from four basic branches and queries from four held-out basic branches.

Support sizes were $K \in \{1, 2, 4, 8, 16, 32\}$ for subordinate, $K \in \{2, 4, 8, 16, 32\}$ for basic, and $K \in \{4, 8, 16, 32\}$ for superordinate learning. Supports were nested across K, and the 16 target queries and their candidate sets were fixed across support sizes. Subordinate and basic trials used three foils from other basic categories within the target superordinate category; superordinate trials used three other eligible superordinate categories. Within each episode, support, target-query, and foil images were disjoint. Adaptation and 4AFC scoring followed the independent positive-only prototype procedure above using five pseudoword mappings.

C.11 Shape–texture word generalization

We used the public style-transfer cue-conflict set of 1,280 images spanning 16 categories, comprising 80 cue-consistent controls and 1,200 cue conflicts [33]. For a target category c and another category d, a novel word was grounded from one cue-consistent support image with shape and texture both equal to c. The two test options formed a reciprocal conflict: the shape-consistent option had shape c and texture d, whereas the texture-consistent option had shape d and texture c. To enforce source disjointness, we excluded supports that reused either the shape option's exact content source or the texture option's exact style source.

The visual encoder and all existing CVCL parameters remained frozen. For $K = 1$, the optimum of the positive-cosine objective is the normalized support embedding, which was therefore used as the novel-word prototype. We compared its cosine similarity with the two option embeddings and assigned half credit to ties. Shape choices were averaged first over valid supports, then over reciprocal option pairs, other categories, and finally the 16 target words, giving each target word equal weight.

In the complementary image-to-word 4AFC control, each cue-conflict image was compared with one-shot prototypes from its shape category, texture category, and two non-cue categories. Seven balanced foil pairs covered all 14 non-cue categories once per query. Shape prototypes excluded the query's exact content source, texture prototypes excluded its exact style source, and every valid four-way support combination was evaluated. We averaged first over support combinations and foil pairs, then over the 1,200 equally weighted queries.