Benchmarking Attribute Discrimination in Infant-Scale Vision-Language Models

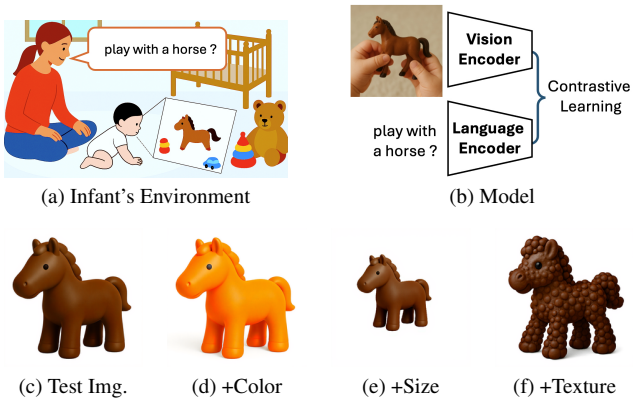
Patrick Batsell[1] Satoshi Tsutsui[2] Bihan Wen[2]
[1]Rice University [2]Nanyang Technological University
pb52@rice.edu satoshi.tsutsui@ntu.edu.sg bihan.wen@ntu.edu.sg

Abstract

Infants learn not only object categories but also fine-grained visual attributes such as color, size, and texture from limited experience. Prior infant-scale vision–language models have mainly been evaluated on object recognition, leaving open whether they support within-class attribute discrimination. We introduce a controlled benchmark that varies color, size, and texture across 67 everyday object classes using synthetic rendering to control attribute values while holding object identity fixed. We evaluate infant-trained models (CVCL and an infant-trained DINO baseline) against web-scale and ImageNet models (CLIP, SigLIP, ResNeXt) under two complementary settings: an image-only prototype test and a text–vision test with attribute–object prompts. We find a dissociation between visual and linguistic attribute information: the evaluated SAYCam-trained models form strong visual representations for size and texture but perform poorly on visual color discrimination, and in the text–vision setting, they struggle to ground color and show only modest size grounding. In contrast, the web-trained vision–language models strongly ground color from text while exhibiting weaker visual size discrimination.

1. Introduction

Infants demonstrate remarkable efficiency in learning visual concepts from limited data. Within their first two years, they acquire not only object categories but also fine-grained attributes such as color, size, and texture [3, 10, 12]. This developmental ability has inspired computational models that learn under similarly constrained conditions [2, 5, 13]. A recent milestone is Child's View for Contrastive Learning (CVCL) [19], which trains a CLIP-style model [11] on approximately 37,000 infant egocentric video frames from the SAYCam dataset [17] paired with parental speech transcripts. Despite training on only about 0.01% of CLIP's data, CVCL achieves impressive object recognition accuracy, consistent with developmental psychology findings that infants acquire concepts from sparse but structured experiences.



(a) Infant's Environment (b) Model

(c) Test Img. (d) +Color (e) +Size (f) +Texture

Figure 1. (a, b): Infants are efficient learners, inspiring models trained on infant egocentric views with parental speech [19]; prior evaluation is limited to object category discrimination. (c–f): We benchmark these models on infant-level attribute discrimination, controlling color, size, and texture without changing other visual characteristics.

However, prior evaluation of CVCL has focused exclusively on object-level classification, leaving open whether infant-scale learning also supports fine-grained attribute discrimination. Developmental psychology emphasizes that children's visual understanding depends not only on identifying objects but also on perceiving within-class differences—distinguishing a red ball from a green one, or recognizing that two cups differ in size [9, 21]. While prior attribute datasets exist, they typically focus on narrow domains or lack systematic control over attributes, making it difficult to disentangle object identity from attribute variation [7, 8, 20]. Consistent with experimental paradigms in perceptual psychology, attribute learning is often probed using isolated objects on uniform backgrounds or simplified synthetic stimuli to control confounding variables [14, 15].

In this work, we introduce a controlled benchmark that systematically varies color, size, and texture across 67 object classes, enabling direct evaluation of attribute discrimination. We compare six models: CVCL, DINO (Infant), DINO (ImageNet), CLIP, SigLIP, and ResNeXt [4, 11, 23,

24]. Our results reveal a striking dissociation: the evaluated SAYCam-trained models encode size more strongly than web-trained models but fail at color discrimination, while web-trained models show the opposite pattern. Inspection of the SAYCam vocabulary indicates that color words appear in child-directed speech, yet CVCL cannot ground them—suggesting that the linguistic supervision available in this specific dataset may be insufficient to support robust color–object grounding.

Contributions.

- We introduce a controlled benchmark for *within-class* attribute discrimination that systematically varies color, size, and texture across 67 everyday object categories while holding object identity fixed.
- We provide the first systematic evaluation of CVCL on controlled within-class attribute discrimination, revealing a sharp asymmetry: CVCL outperforms web-trained models on visual size discrimination but substantially underperforms on color discrimination.
- We identify a consistent dissociation across training regimes: infant-trained models encode size more strongly yet fail on color, whereas web-trained vision–language models strongly ground color but show no visual size advantage.

2. Benchmark Design

2.1. Dataset Construction

We construct a controlled benchmark that systematically varies three visual attributes—color, size, and texture—across multiple everyday object classes. These attributes were chosen because they are among the earliest perceptual features that infants reliably recognize [3, 10, 12]. Following the original CVCL paper, we draw 67 object classes from the Konkle "things" dataset [6], a collection of high-quality object images spanning everyday categories that infants are likely to encounter.

Rather than relying on real-world photographs, we employ the OmniGen2 generative model [22] to synthesize images. Synthetic generation provides precise control over attribute variation, balanced coverage of all attribute combinations, and minimization of confounds such as background clutter or lighting.

For each class, images are generated by systematically crossing three attribute dimensions:

- 9 distinct colors (blue, brown, gray, green, orange, pink, purple, red, yellow)
- 3 size bins (small: $< 20\%$ pixel occupancy, medium: 20–70%, large: $> 70\%$) Pixel occupancy captures apparent size in the image; with fixed camera and background settings, it provides a simple, controlled size cue aligned with perceptual judgments.
- 2 texture variants (smooth, bumpy)

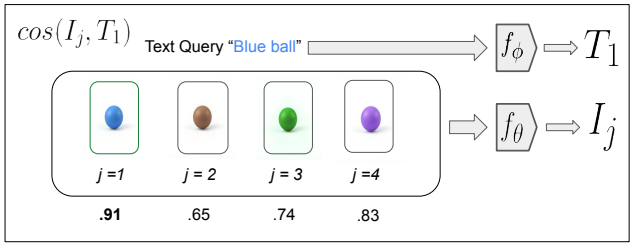


Figure 2. Text-vision test procedure. A query image embedding is compared against text embeddings of attribute-object phrases (e.g., "red cup", "blue cup"). The highest cosine similarity determines the predicted attribute.

Images are rendered at 224×224 resolution on a uniform white background, following the methodology of the Konkle dataset [6]. We restrict texture to two variants to align with developmental psychology studies showing that infant texture discrimination is typically tested in binary form [1]. We also create two distinct style variants per class to increase within-class diversity, yielding $9 \times 3 \times 2 \times 2 = 108$ images per class (7,236 total).

2.2. Evaluation Protocol

Our evaluation focuses on within-class discrimination using a 4-way forced-choice task. We test three conditions:

1. **Same-class, different color (SCDC)**: Can the model distinguish colors within the same object class?
2. **Same-class, different size (SCDS)**: Can the model distinguish sizes within the same object class?
3. **Same-class, different texture (SCDT)**: Can the model distinguish textures within the same object class?

Each condition is tested under two complementary modes:

Prototype tests (image-only). A prototype embedding is computed as the mean of ℓ_2-normalized image embeddings sharing a target attribute (e.g., all green balls), excluding the query [16]. The query embedding is compared to each prototype using cosine similarity and labeled according to the most similar prototype. This evaluates whether the model forms consistent attribute-based clusters without language.

Text-vision tests (image-text). Attribute-based text prompts (e.g., "red cup") are used in place of prototypes. The query image embedding is compared with text embeddings using cosine similarity, and the predicted attribute is given by the highest-similarity candidate. This evaluates whether the model can ground linguistic descriptions in visual features. We evaluate text–vision alignment only for color and size; texture is omitted because relevant adjectives are largely absent from SAYCam.

Candidate set construction. For each query, we form a 4-way forced-choice set by sampling three distractors from the same object class that differ only in the tested attribute.

Chance accuracy is 25%. We repeat trial construction over three independent seeds and report mean accuracy ± standard deviation.

3. Models

We evaluate six models with contrasting training paradigms:

- **CVCL**: ResNeXt-50 trained on SAYCam infant egocentric data with CLIP-style contrastive learning [19].
- **DINO (Infant)**: ResNeXt-50 trained on SAYCam using self-supervised DINO [4].
- **DINO (ImageNet)**: ResNeXt-50 trained on ImageNet using self-supervised DINO [4].
- **CLIP**: ResNet-50 trained on 400M web image-text pairs [11].
- **SigLIP**: ViT-B/16 trained on web-scale data with sigmoid loss [24].
- **ResNeXt**: ResNeXt-50 pre-trained on ImageNet with supervised classification [23].

CVCL and DINO (Infant) share the same infant visual data but differ in language supervision, enabling isolation of the language contribution. CLIP, SigLIP, and ResNeXt represent web-scale training with varying architectures and objectives.

4. Results

4.1. Overall Classification

We first evaluate overall *class* discrimination to verify that models trained on naturalistic data generalize to our synthetic renders and to establish a baseline before moving to fine-grained attribute discrimination. Results are shown in Fig. 3 (prototype) and Fig. 4 (text–vision).

Why text–vision is harder with our benchmark. While the prototype setting only requires that images of the same category form a coherent cluster, the text–vision setting additionally requires grounding *attribute words* to visual evidence. Our benchmark deliberately makes this harder by breaking real-world attribute priors—e.g., by synthesizing non-canonical color realizations (atypically colored apples) while holding object identity fixed [8, 18], removing shortcuts and forcing attribute recognition to depend on actual visual content.

4.2. Prototype Tests

Fig. 5 presents the prototype test results. First, **the evaluated SAYCam-trained models (CVCL, DINO Infant) excel at size discrimination** compared to web-trained models, suggesting the egocentric visual experience captured in SAYCam encodes size information that ImageNet and web data do not emphasize. Second, **the SAYCam-trained models fail at color discrimination**, while web-trained

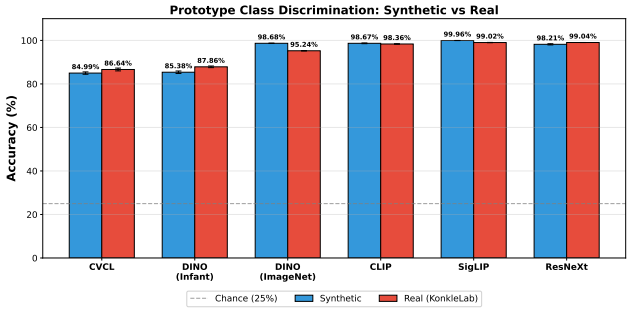


Figure 3. Prototype class discrimination on real Konkle images vs. our synthetic renders. Accuracy remains high and trends are consistent, suggesting the synthetic benchmark retains object identity while enabling controlled attribute variation.

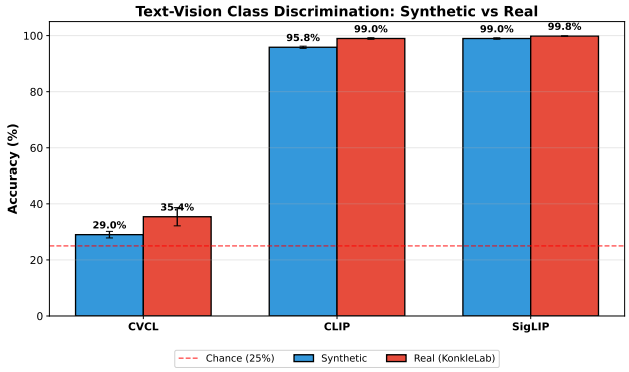


Figure 4. Text–vision class discrimination for models with text encoders (CVCL, CLIP, SigLIP). High accuracy on synthetic renders with non-canonical appearances confirms models recognize object identity under counterfactual colors, validating downstream attribute tests.

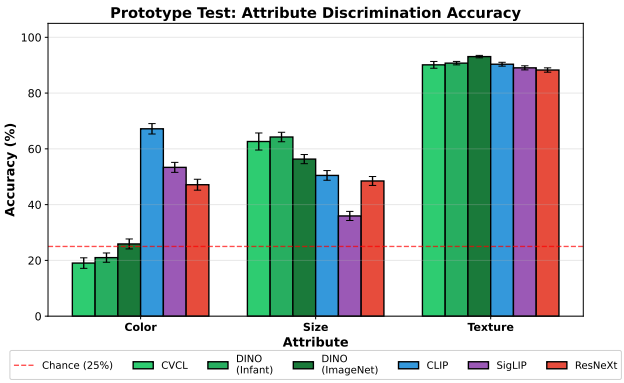


Figure 5. Prototype test results across six models and three attribute discrimination tasks. Infant-trained models (CVCL, DINO Infant) excel at size discrimination but fail at color, while web-trained models show the opposite pattern. All models perform similarly on texture. Chance is 25% (dashed).

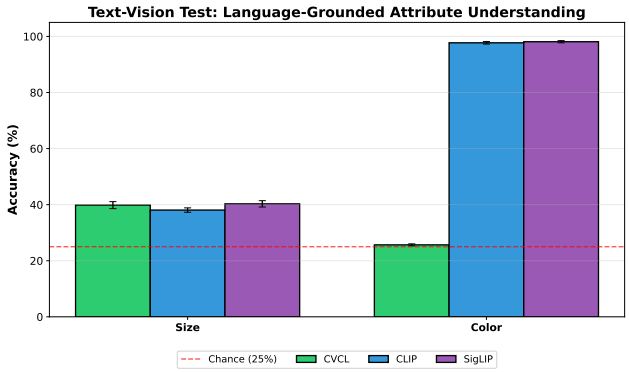


Figure 6. Text-vision test results for vision-language models. CLIP and SigLIP achieve near-perfect color grounding, while CVCL performs at chance. All models show modest performance on size grounding. Chance is 25% (dashed).

models achieve substantially higher accuracy. Third, **all evaluated models perform similarly on texture**, indicating texture discrimination is robustly learned regardless of these specific training data differences.

4.3. Text-Vision Tests

Fig. 6 shows the text-vision test results. Both CLIP and SigLIP achieve near-perfect color grounding, while CVCL performs at chance. Per-color analysis confirms this failure is not color-specific: CVCL's predictions are approximately uniform across target colors, consistent with random guessing. For size, all models perform modestly above chance, suggesting size words are poorly grounded across all training regimes.

All tested color words appear in SAYCam transcripts (e.g., red, blue, green), yet CVCL cannot ground them, suggesting that the limited scale of infant-directed linguistic supervision is insufficient for robust attribute–object grounding. Synonym substitution tests (varying size adjectives: *big*/*huge*/*little*/*tiny*) confirm that the overall model ordering is stable across prompts, though CVCL exhibits the highest prompt sensitivity (24.3 pp range vs. 17.2 pp for CLIP). CVCL's color confusion is uniform across predictions, consistent with guessing rather than collapse to a subset of colors; in contrast, size errors show more structure, suggesting partial size grounding.

5. Discussion

Our findings reveal qualitatively different attribute representations from infant-scale vs. web-scale training:

Infant visual experience encodes size. Both CVCL and DINO (Infant) outperform web-trained models on visual size classification. The source of this advantage remains an open question.

Infant experience does not encode color. Despite color words appearing in child-directed speech, infant-trained models fail at color discrimination. CVCL's performance matching DINO (Infant) suggests this is a weakness of the training data, not architectural differences.

Web-scale language grounds color but not size. CLIP and SigLIP achieve near-perfect color grounding but show no advantage for size.

These results highlight complementary strengths: infant-scale learning captures size, while web-scale learning captures color that is frequently described in text. Our benchmark isolates attribute representations under controlled conditions, analogous to laboratory perception studies that minimize confounds [14, 15].

Limitation. Our stimuli are controlled synthetic renders on uniform backgrounds; absolute accuracies may differ in natural scenes. The benchmark is designed to isolate attribute representations under counterfactual appearances rather than to measure in-the-wild performance.

6. Related Work

Our work lies at the intersection of infant-scale visual learning, multimodal grounding, and controlled attribute benchmarking. Prior work has shown that infants acquire visual attributes early [1, 3, 10, 12, 21], and models such as CVCL demonstrate category learning under extreme data scarcity [19]. However, existing attribute benchmarks either lack systematic control or focus on domains misaligned with infant perception [7, 8, 20].

By generating synthetic variants of infant-relevant object categories with controlled attribute manipulations, our benchmark isolates within-class attribute representations under counterfactual appearances (e.g., non-canonical colors) while preserving object identity, analogous to controlled stimuli used in perception studies.

7. Conclusion

We introduced a controlled benchmark for within-class attribute discrimination over infant-relevant object categories, systematically varying color, size, and texture while holding object identity fixed. Across the evaluated models, we observe a consistent dissociation: the SAYCam-trained models visually encode size better than web-scale models but fail to discriminate or ground color, whereas web-trained vision–language models strongly ground color but show limited advantages on size. By breaking natural co-occurrence priors via counterfactual attribute realizations, our benchmark provides a principled tool for probing which perceptual attributes different training regimes capture. Future work can extend this framework to additional attributes and evaluate more recent infant-scale models as new training data becomes available. Benchmark dataset and code will be released publicly upon acceptance.

References

[1] Janette Atkinson and Oliver Braddick. Visual segmentation of oriented textures by infants. *Behavioural Brain Research*, 49(1):123–131, 1992. 2, 4

[2] Sven Bambach, David J. Crandall, Linda B. Smith, and Chen Yu. Toddler-inspired visual object learning. In *Advances in Neural Information Processing Systems*, 2018. 1

[3] Marc H Bornstein, William Kessen, and Sally Weiskopf. Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 2(1):115, 1976. 1, 2, 4

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 3

[5] Xueyi Ke, Satoshi Tsutsui, Yayun Zhang, and Bihan Wen. Discovering hidden visual concepts beyond linguistic input in infant learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4343–4352, 2025. 1

[6] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental Psychology: general*, 139(3):558, 2010. 2

[7] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 4

[8] Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Adu Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? In *Advances in Neural Information Processing Systems*, 2025. 1, 3, 4

[9] Amy Needham. Infants' use of featural information in the segregation of stationary objects. *Infant Behavior and Development*, 21(1):47–76, 1998. 1

[10] Anthony M Norcia, Christopher W Tyler, and Russell D Hamer. Development of contrast sensitivity in the human infant. *Vision research*, 30(10):1475–1486, 1990. 1, 2, 4

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3

[12] Özlem Sensoy, Jody C Culham, and Gudrun Schwarzer. Do infants show knowledge of the familiar size of everyday objects? *Journal of experimental child psychology*, 195: 104848, 2020. 1, 2, 4

[13] Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 2023. 1

[14] Kenny Smith, Andrew DM Smith, and Richard A Blythe. Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3):480–498, 2011. 1, 4

[15] Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008. 1, 4

[16] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 2017. 2

[17] Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5:20–29, 2021. 1

[18] Yingtian Tang, Yutaro Yamada, Yoyo Minzhi Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *The Conference on Empirical Methods in Natural Language Processing*, 2023. 3

[19] Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024. 1, 3, 4

[20] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 4

[21] Teresa Wilcox. Object individuation: Infants' use of shape, size, pattern, and color. *Cognition*, 72(2):125–166, 1999. 1, 4

[22] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2

[23] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 1, 3

[24] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11973–11985, 2023. 2, 3