Tracing the Development of Syntax and Semantics in a Model trained on Child-Directed Speech and Visual Input

Nina Schoener*
Department of Psychology
University of California, Berkeley

Mahesh Srinivasan
Department of Psychology
University of California, Berkeley

Colin Conwell
CSAIL
MIT

Abstract

In contrast to most large language models, children learn in an remarkably data-efficient manner from their naturalistic environments. These environments are grounded in sensory perception, critically enabling learning from the alignment between visual and auditory input, among other types. In the current work, we identify three signatures of human language acquisition that highlight this ability for efficient, constructive learning: the identification of semantic hierarchies, the ability to bootstrap meaning on the basis of polysemy, and the generalization of syntactic frames. Using a testing set of the CHILDES Providence corpus [1], we create probes for each of these signatures of efficient learning, and test them on variations of the BabyLLaVA [2] model (trained on SAYCam [3]): the baseline model, one finetuned on younger children's data from the Providence corpus, one finetuned on older children's data from the Providence corpus, and one finetuned on equal samples of younger and older children's data as a more rigorous baseline. Preliminary findings suggest that finetuning on the Providence corpus improves performance on most probes, though we also observe interactions between finetuning data and probe type. This suggests that, while some signatures of efficient human language learning are present in a VLM trained on naturalistic data, not all types of data contribute equally to their emergence.

1 Introduction

Modern deep learning models have demonstrated unprecedented ability to acquire linguistic knowledge from general-purpose learning mechanisms without built-in language-specific biases. However, they are developmentally implausible in several ways: often trained on orders of magnitude more data than children receive, and primarily exposed to text alone, unlike human learners who experience and learn in rich multisensory input ecologies [4]. What happens when models learn from similar inputs?

Models trained on child-directed input Past work on training models with child-directed input has often focused on producing coherent adult-like output efficiently [5; 6; 7; 8]. While it has been hypothesized that training on child-directed input might lead to more data-efficient learning, prior studies comparing child- versus adult-directed corpora have provided mixed or inconclusive evidence [9; 7]. In this study, we investigate whether vision-language models (VLMs) exposed to naturalistic child-directed input exhibit signatures of efficient human-like language learning. Demonstrating such signatures would suggest that models can, at least in principle, extract linguistic knowledge from multimodal data in ways analogous to human language acquisition. In this investigation, we design novel probes of syntax and semantics and apply them to models finetuned either on infant-directed input or on toddler-directed input. We then evaluate which type of input best supports the emergence of child-like linguistic signatures long considered key indicators of efficient learning.

*Correspondence to: nina_schoener@berkeley.edu.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp.

Human language acquisition Young children acquire both word meanings and syntactic structures relatively quickly despite receiving far less input than would seem necessary – a classic example of the induction problem inherent to human language acquisition [10; 11]. Children are rarely told directly which uses are correct or incorrect, and yet they must infer the right generalizations from a vast hypothesis space [12]. Various accounts have proposed innate constraints, but others suggest that children can rely on distributional evidence and inductive principles: for example, using positive evidence to rule out implausible hypotheses [13; 14].

In this spirit, we focus on three signatures of efficient learning that may help children *and* language models solve inductive problems. First, *zero-shot polysemy extension* allows children to use knowledge of one word sense to constrain and extend another, reducing ambiguity in novel contexts. Second, *semantic hierarchy formation* enables children to organize meanings in structured ways, helping them to identify relevant generalizations and rule out unlikely ones. Third, *syntactic frame learning* captures how children abstract from limited exemplars to infer broader combinatorial rules—an ability that reflects an inductive solution to mapping input variability onto underlying regularities. In the current study, we test for these signatures in a VLM trained and finetuned on child-directed input.

We further test how the models finetuned on infant-directed and toddler-directed input compare to each other and to a model finetuned on an equal mix of both types of input. The goal of this comparison is to assess whether input directed towards younger children may be particularly conducive to learning some aspects of language while input directed towards older children may be conducive to others. Such effects have long been attested in the literature on human language acquisition [15; 16].

2 Methods

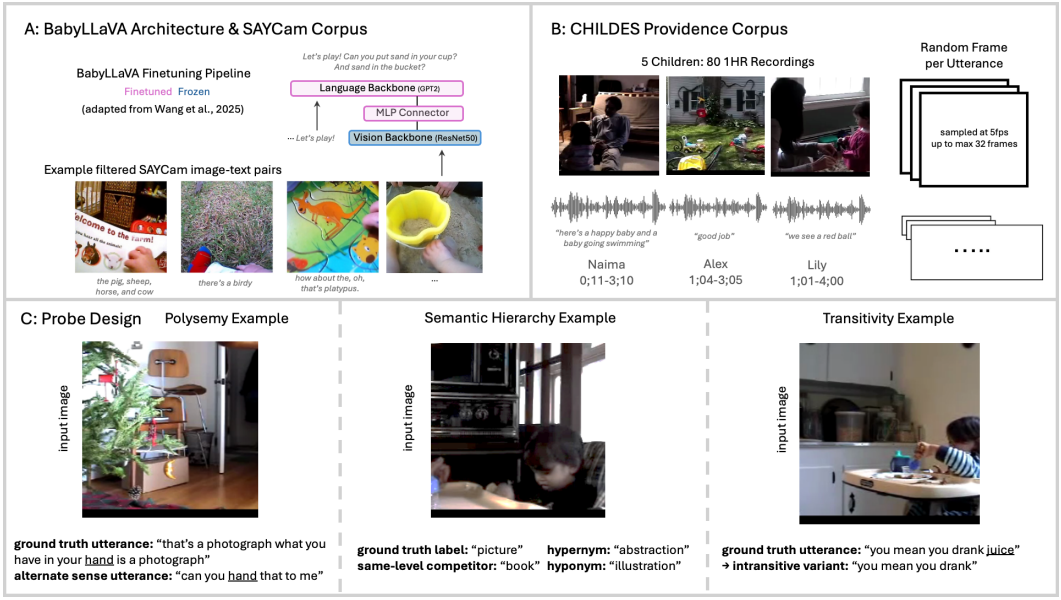

Figure 1: A: Information about the BabyLLaVA architecture and training pipeline, along with examples from the filtered SAYCam dataset (adapted from [2]). B: Description of CHILDES Providence corpus and preprocessing. C: Example probes of each type: polysemy, semantic hierarchy, and transitivity.

Dataset + Models The current study makes use of a previously introduced VLM (BabyLLaVA [2]) trained on child-directed input (the SAYCam dataset [3]). We finetune the out-of-the-box BabyLLaVA model on a second dataset of child-directed input: the CHILDES Providence corpus [1]. Our principal motivation for further finetuning the BabyLLaVA model before evaluating its performance was to include child-directed data from a broader age range of children than the SAYCam data provides. Additionally, the SAYCam dataset provides egocentric headcamera video, while the CHILDES Providence corpus provides a third-person fixed video perspective. We reasoned that because this

third-person perspective is likely to capture a broader view of the scene, it may be more conducive to extracting semantic and syntactic information across scenes. We also note that this means the CHILDES dataset is less directly naturalistic than the first-person SAYCam video data, though it has also been shown that children can learn from observation of third-person interactions and overheard speech [17; 18; 19] and may often find themselves in the position to do so (e.g., in daycare settings or when observing parent-sibling interactions).

BabyLLaVA BabyLLaVA [2] was created as part of a larger framework, BabyVLM [2], which consists of a developmentally-inspired vision-language model and complementary developmentally-inspired benchmarks. The model architecture and training pipeline resembles that of LLaVA, featuring a GPT-2 Language Backbone with around 7 million parameters, a ResNeXt-50 Vision Backbone with around 23 million parameters, and a lightweight MLP connector. The model training was done in three distinct phases, in which the language and vision backbones were first trained independently and then integrated via the MLP connector [2]. The language and vision backbones were trained on naturalistic developmental data sourced from the SAYCam dataset.

SAYCam dataset The SAYCam dataset, first introduced by Sullivan et al. [3], is a multimodal longitudinal dataset that includes headcamera video and transcript data from three children throughout multiple years of development (ranging from 6 to 38 months). BabyLLaVA was trained on data from one of these three children [6]. Subsequent data processing [6] yielded approximately 67K image-utterance pairs used for model training.

Providence dataset The Providence dataset, constructed by Demuth et al. [1], consists of longitudinal third-person video and transcript data from five children ranging in age from 11 months to 4 years. Sessions were typically recorded biweekly for one hour each (some children were recorded more or less frequently), yielding approximately 300 hours of transcribed video data. In the current work, we matched utterances from these transcripts (truncated to 25 tokens and edited to remove utterances produced by the child themselves, transcription notes, and punctuation) to segments of video recordings, sampled frames at a rate of 5fps up to a maximum of 32 frames, and then randomly selected one of these frames to be the image paired with the utterance for the purposes of model finetuning. This resulted in a set of approximately 122K image-utterance pairs, which were then split into train/validation/test sets of three types: one including half of the total data from when children were younger than 2 years, 3 months (Infant-finetuned model); one including half of the total data from when children were 2 years, 3.5 months and older (Toddler-finetuned model); and one including 50% of the total data sampled randomly from each half (forming a baseline against which to evaluate the other two finetuned models). This age cutoff between the infant and toddler finetuned models results in approximately equal numbers of utterances in each dataset.

Experimental Probes For the purposes of evaluating the two models finetuned on the younger and older halves of the Providence dataset, we designed three probe types, each testing for one of the three hallmarks of human language acquisition described in Section 1.

Polysemy Our first probe tested for the ability to represent polysemy – the phenomenon when a single word form denotes multiple, related meanings, such as in *chicken* (food) and *chicken* (animal). To do this, we extracted instances from the test set of each finetuning dataset in which a polysemous word was used (such as "what you have in your *hand* is a photograph"), and matched the image from that instance to both the original ground-truth utterance, and an utterance that included a different sense of the same word (such as "can you *hand* that to me"). Polysemous senses were extracted using a dataset of annotated CHILDES transcript word senses [20]. At evaluation, the input consisted of the ground truth image, on the basis of which the log probabilities for each token in each of the two utterance options were extracted and normalized for each utterance. The utterance with the higher mean log probability per token was taken to be the model's response and was counted correct if this corresponded to the ground truth utterance. A similar evaluation approach is taken for the semantic hierarchy and transitivity probes. Filtering to include only probes with at least 80% vocabulary overlap with the finetuning train data, this yielded 147 probes.

Semantic Hierarchy Our second probe type tested for the construction of semantic hierarchies. For example: in the case that a dog is the topic of discussion and perhaps even appears in-frame, does the model recognize that in some contexts it may be referred to as a *dog*, but in other cases as a *pet*, or as

a *Golden Retriever*? For this purpose, we identified instances of tokens in the test data bearing at least one hypernym (superordinate category), one hyponym (subordinate category), and one same-level competitor as identified in WordNet [21]. Given the input of the ground-truth image with which a basic-level label was used, we extracted the model's log probabilities for each of these four labels. This yielded 56 probes with at least 80% vocabulary overlap with the finetuning train data. Cases in which the ground-truth label received the highest log probability were counted as correct.

Transitivity Lastly, to probe the model's generalizations of syntax, we designed probes testing the model's representations of *transitivity*, the linguistic property of either having (transitive) or not having (intransitive) a direct object. Many English verbs can be used in both constructions: for example, "she's eating" and "she's eating the pie" are both grammatical, but may be used in different social and environmental contexts. We identified instances in the test data in which a common verb was used in a transitive syntactic frame (such as "she's eating the pie") and manipulated them to remove the direct object (yielding "she's eating"). Given the input of the ground-truth image, both utterances were then evaluated for mean log-probability per token. Cases in which the ground-truth utterance received the highest log probability were counted as correct. This process yielded 83 probes after filtering for 80% vocabulary overlap with the training data.

3 Results

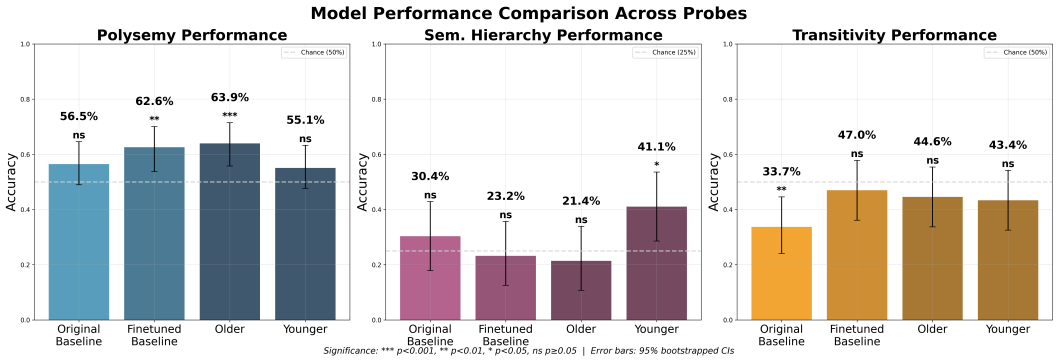

Figure 2: Comparison of performance across four models: the baseline BabyLLaVA model, a baseline model finetuned on 50% of the infant data and 50% of the toddler data, the infant finetuned model, and the toddler finetuned model, for each of three probe types. Error bars represent 95% bootstrapped confidence intervals over individual probes.

Baseline BabyLLaVA Performance on the polysemy and semantic hierarchy probes was slightly above chance for the baseline BabyLLaVA model, though not significantly so (56.5% and 30.4%, respectively). Performance on the transitivity probes, however, was significantly below chance levels. This suggests a baseline bias towards intransitive utterances which cannot be purely an effect of utterance length due to our normalization for utterance length at evaluation.

Baseline finetuned Performance on the polysemy and transitivity probes improved upon finetuning the BabyLLaVA model on half of the infant and half of the toddler data, to 62.6% and 47.0%, respectively. This resulted in performance significantly above chance level for polysemy probes. However, the baseline finetuned model performed at chance level on the semantic hierarchy probes (23.2%), a slight but statistically insignificant reduction in performance as compared to Baseline BabyLLaVA.

Infant finetuned BabyLLaVA As compared to the finetuned baseline, performance on the polysemy and transitivity probes was slightly lower (55.1% and 43.4%, respectively) resulting in chance performance on both probe types. In contrast, performance on the semantic hierarchy probes nearly doubled to 41.1%. This level of performance is significantly above the chance level of 25% and is the highest of all four models on this probe type.

Toddler finetuned BabyLLaVA As compared to the finetuned baseline, performance on all three probe types was similar, with a slight increase in polysemy to 63.9% and slight decreases in semantic hierarchy and transitivity to 21.4% and 44.6%, respectively. Thus, this model performed above chance levels only on polysemy probes.

4 Conclusion

Taken together, our results in this analysis provide preliminary evidence that the BabyLLaVA model finetuned on further naturalistic visual and linguistic input from the CHILDES Providence corpus shows some signatures of human language acquisition, as in the distinction of polysemous word senses, the construction of semantic hierarchies, and the varied use of syntactic frames based on context. However, it appears that there is a dissociation in terms of which type input is most useful for each hallmark of efficient learning. For example, while it is clear that the infant-finetuned model performed best at constructing semantic hierarchies, perhaps due to the prevalence of explicit object labeling in this data split, a *mix* of infant-directed and toddler-directed data actually yielded the highest reduction in the bias towards intransitive semantic frames (see 'Baseline Finetuned' performance on Transitivity probes). A further preliminary observation that can be made on the basis of these model comparisons is that the Baseline BabyLLaVA model (trained on SAYCam data) differed in performance on some probe types as compared to the Baseline finetuned model (which was finetuned on a mix of infant-directed and toddler-directed input from the CHILDES Providence corpus). One possible explanation for this difference in performance relates to the shift in perspective between first-person and third-person video data – it is possible that, because third-person video data captures more of the child's environment, it is more obvious which interactions involve a direct object and thus should be paired with a transitive utterance when viewed from this perspective. Future research should further investigate the presence of signatures of efficient human learning in VLMs trained on a variety of input types. One outcome of this research could be a much more efficient regime of curriculum learning that approximates the social and sensory ecologies young humans learn from so quickly, and which model developers may do well to mimic.

References

[1] Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. Word-minimality, Epenthesis and Coda Licensing in the Early Acquisition of English. *Language and Speech*, 49(2):137–173, June 2006. ISSN 0023-8309, 1756-6053. doi: 10.1177/00238309060490020201.

[2] Shengao Wang, Arjun Chandra, Aoming Liu, Venkatesh Saligrama, and Boqing Gong. BabyVLM: Data-Efficient Pretraining of VLMs Inspired by Infant Learning, April 2025. URL https://arxiv.org/abs/2504.09426.

[3] 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, May 2021. ISSN 2470-2986. doi: 10.1162/opmi_a_00039.

[4] Michael C. Frank. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992, November 2023. ISSN 1879-307X. doi: 10.1016/j.tics.2023.08.007.

[5] Bria Long, Robert Z. Sparks, Violet Xiang, Stefan Stojanov, Zi Yin, Grace E. Keene, Alvin W. M. Tan, Steven Y. Feng, Chengxu Zhuang, Virginia A. Marchman, Daniel L. K. Yamins, and Michael C. Frank. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences, July 2025. URL http://arxiv.org/abs/2406.10447. arXiv:2406.10447 [cs].

[6] 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, February 2024. doi: 10.1126/science.adi1374. URL https://www.science.org/doi/10.1126/science.adi1374. Publisher: American Association for the Advancement of Science.

[7] Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. Is Child-Directed Speech Effective Training Data for Language Models? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1231. URL https://aclanthology.org/2024.emnlp-main.1231/.

[8] Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, Adina Williams, Tal Linzen, et al. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2504.08165*, 2025.

[9] Seoyoung Ahn, Gregory J. Zelinsky, and Gary Lupyan. Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, 21(13):13, December 2021. ISSN 1534-7362. doi: 10.1167/jov.21.13.13. URL https://doi.org/10.1167/jov.21.13.13.

[10] Noam Chomsky. Persistent Topics in Linguistic Theory. *Diogenes*, 13(51):13–20, September 1965. ISSN 0392-1921. doi: 10.1177/039219216501305102. URL https://doi.org/10.1177/039219216501305102. Publisher: SAGE Publications Ltd.

[11] Ellen M. Markman. Constraints Children Place on Word Meanings. *Cognitive Science*, 14(1):57–77, 1990. ISSN 1551-6709. doi: 10.1207/s15516709cog1401_4. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1401_4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1401_4.

[12] Gary F. Marcus. Negative evidence in language acquisition. *Cognition*, 46(1):53–85, January 1993. ISSN 0010-0277. doi: 10.1016/0010-0277(93)90022-N. URL https://www.sciencedirect.com/science/article/pii/001002779390022N.

[13] Fei Xu and Joshua B. Tenenbaum. Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272, 2007. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.114.2.245. URL https://doi.apa.org/doi/10.1037/0033-295X.114.2.245.

[14] Terry Regier. The Emergence of Words: Attentional Learning in Form and Meaning. *Cognitive Science*, 29(6):819–865, 2005. ISSN 1551-6709. doi: 10.1207/s15516709cog0000_31. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_31. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog0000_31.

[15] Janellen Huttenlocher, Heidi Waterfall, Marina Vasilyeva, Jack Vevea, and Larry V. Hedges. Sources of variability in children's language growth. *Cognitive Psychology*, 61(4):343–365, December 2010. ISSN 1095-5623. doi: 10.1016/j.cogpsych.2010.08.002.

[16] Meredith L. Rowe. A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, 83(5):1762–1774, 2012. ISSN 1467-8624. doi: 10.1111/j.1467-8624.2012.01805.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8624.2012.01805.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8624.2012.01805.x.

[17] Yuriko Oshima-Takane, Elizabeth Goodz, and Jeffrey L. Derevensky. Birth Order Effects on Early Language Development: Do Secondborn Children Learn from Overheard Speech? *Child Development*, 67(2):621–634, 1996. ISSN 1467-8624. doi: 10.1111/j.1467-8624.1996.tb01755.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8624.1996.tb01755.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8624.1996.tb01755.x.

[18] Maricela Correa-Chávez and Barbara Rogoff. Children's attention to interactions directed to others: Guatemalan mayan and european american patterns. *Developmental Psychology*, 45(3):630–641, 2009. ISSN 1939-0599, 0012-1649. doi: 10.1037/a0014144. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0014144.

[19] Nameera Akhtar, Jennifer Jipson, and Maureen A. Callanan. Learning Words Through Overhearing. *Child Development*, 72(2):416–430, 2001. ISSN 0009-3920. URL https://www.jstor.org/stable/1132404. Publisher: [Wiley, Society for Research in Child Development].

[20] Stephan Meylan, Jessica Mankewitz, Sammy Floyd, Hugh Rabagliati, and Mahesh Srinivasan. Quantifying Lexical Ambiguity in Speech To and From English-Learning Children, February 2021. URL https://osf.io/zxkm2.

[21] George A. Miller. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL https://dl.acm.org/doi/10.1145/219717.219748.