Learning visual cognition from children's visual experiences

Khai Loong Aw Wanhee Lee Klemen Kotar
Rahul Mysore Venkatesh Honglin Chen Michael Frank Daniel LK Yamins
Stanford University

Abstract

Even in infancy, children display sophisticated visual reasoning abilities, prompting long-standing debates over whether they are innate or learned. To study this, recent studies have trained computational models on egocentric video datasets from children, e.g., BabyView. However, they focused on perceptual tasks rather than more sophisticated visual reasoning, and used labels to fine-tune model readout layers, thus limiting the strength of their claims. In this work, we apply the recently developed Local Random Access Sequence (LRAS) framework, which progressively trains a series of self-supervised models. We train LRAS on 800 hours (approximately 0.2 child-years) of BabyView data. Our models successfully perform a range of 3D perceptual tasks for objects, depth, and scenes, as well as cognitive tasks such as simulation of future object motion and viewpoint changes, and physical reasoning about object cohesion, solidity, agent-object motion, and multi-object interactions. Notably, our models perform tasks in a unified, zero-shot manner, thus providing stronger evidence for the learning-based hypothesis. Overall, we establish a computational proof-of-concept that visual cognitive abilities can emerge from developmentally realistic experience through statistical learning with minimal innate priors.

Keywords: vision, cognition, baby, child, machine learning

Introduction

Children show meaningful knowledge and reasoning about objects, places, and agents (Spelke & Kinzler, 2007; Spelke, 2022). For example, experimental evidence shows they organize perceptual input into distinct objects with boundaries, and reason about object cohesion, solidity, and interactions with other objects. They recognize agents as beings that generate their own motion and cause changes to the state of the world. Early emergence of these capacities has been taken as evidence for innate systems of "core knowledge", but this assumption may not be valid if these capacities can be learned from data accessible to the developing child.

We address this question by training deep neural networks with self-supervised learning objectives on BabyView (Long et al., 2024), a dataset of egocentric videos from children.

Prior work training models on child-egocentric videos has mostly focused on visual representation and recognition learning, e.g., object recognition, action recognition, and segmentation (Bambach, Crandall, Smith, & Yu, 2017; Orhan, Gupta, & Lake, 2020; Zhuang et al., 2021; Sheybani, Hansaria, Wood, Smith, & Tiganj, 2023). In contrast, in addition to perceptual tasks such as segmentation and depth estimation, we focus on visual cognition, the broad set of capacities that include physical simulation and reasoning about object behaviors, interactions between objects, and interactions between animate agents and objects. In addition, prior work required fine-tuning model readout layers using task-specific datasets with labels to perform downstream tasks, whereas we do not. This is crucial as non-human animals also exhibit visual cognition but do not receive explicit labeled supervisory signals.

Methods

The **Local Random Access Sequence (LRAS)** framework (Lee et al., 2025) (Figure 1, left) is a framework for progressively learning a series of self-supervised models from video frames. We train the following models. (1) A tokenizer that converts each image into patches of discrete tokens with patch indices, allowing prediction in any desired order. (2) An RGB model: (RGB0, sparse RGB1 $\rightarrow$ dense RGB1), which can then be used as a flow extractor to extract motion (Flow) between pairs of frames: (RGB0, RGB1 $\rightarrow$ Flow). Flow is used to train an (3) RGB-Flow-to-RGB model: (RGB0, Flow $\rightarrow$ RGB1) and an (4) RGB-to-Flow model: (RGB0, sparse Flow $\rightarrow$ dense Flow). BabyView contains videos with corresponding inertial measurement unit (IMU) data, which records the child's head motion in acceleration and rotation units. We train an RGB-IMU model: (RGB0, IMU $\rightarrow$ RGB1) that can predict novel views of scenes based on specified head motions, e.g., predict how a scene looks after panning rightwards (Figure 1). The tokenizer is a convolutional neural network with 40M parameters. The other models are autoregressive transformer models; 1B for the IMU model and 7B for the others.

Each trained model has a single, unified interface to perform tasks. For example, we can probe the RGB model to predict what happens if an object moves by providing RGB0 and a single RGB1 patch for the desired new location for the object. The model will complete its RGB1 prediction of the scene including how the object (and interacting objects) move.

Data. We train on BabyView, an ongoing dataset of $\sim$ 800 hours of egocentric videos. The majority are longitudinal recordings from the homes of children aged 6–36 months; $\sim$ 110 hours are from 3- to 5-year-old children in a preschool.

3DEditBench-Cog. We build a small benchmark to evaluate simulation and reasoning about agent-generated motion, object cohesion & motion, containment, support, and solidity & force transfer. Each example contains both an initial and ground truth frame (where objects are moved to new locations). Each model is run with 400 evaluations: 8 random seeds for 10 examples in each of the 5 categories.

Comparisons. Comparison models are trained on 6000+ hours of diverse, non-developmental videos (Big Video

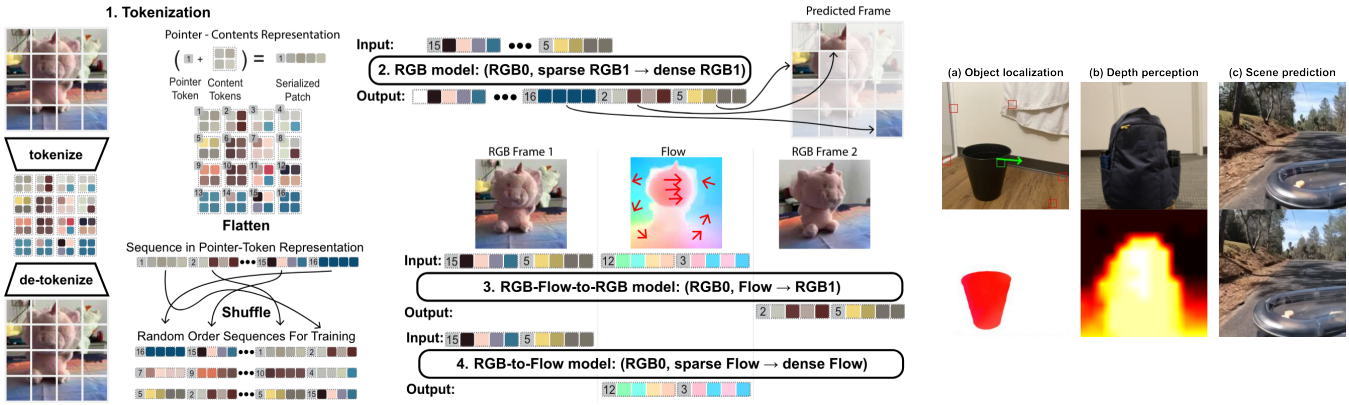

Figure 1: (Left) The Local Random Access Sequence (LRAS) framework progressively learns a series of self-supervised models from videos. (Right) Our model learns (a) object localization and boundaries, (b) depth perception, and (c) predicts how the scene will look if the head/camera is panned rightwards.

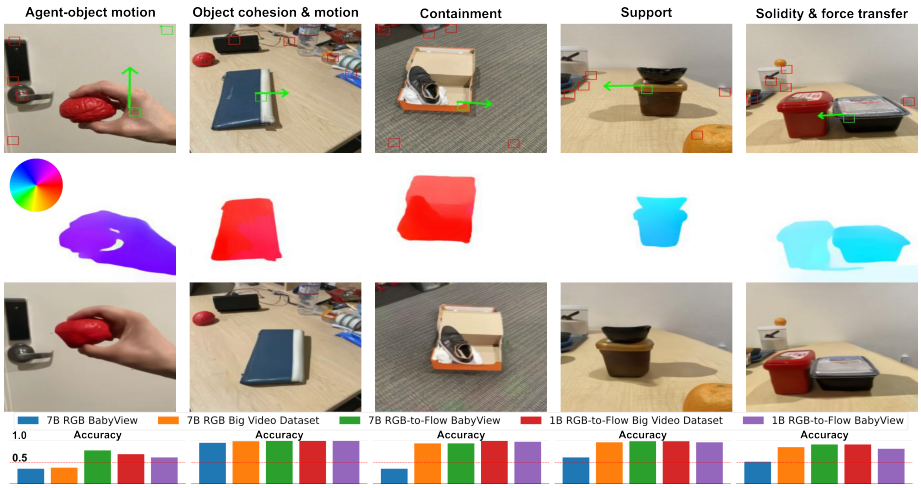

Figure 2: Our benchmark, 3DEditBench-Cog, evaluates model's simulation and reasoning abilities in five categories (see column labels). Top row shows model inputs: green arrow specifies a single patch to move, while red squares specify patches fixed with no motion. Row 2 shows RGB-to-Flow model predictions. Colors represent the direction of motion predicted (see color wheel). Row 3 shows RGB model predictions. Row 4 shows accuracy, defined as the percentage of examples where the model prediction is more similar to the target frame than the initial frame.

Dataset). We also compare to 1B parameter models.

Results

Perception. From a young age, children can segment objects from the background, perceive depth, and predict motion (Arterberry & Kellman, 2016). Our model shows all of these abilities in a single framework. Given a single object patch, our model localizes all related object patches and boundaries (Figure 1, right), providing localization and segmentation of objects. Our model performs stereoscopic depth estimation: given two laterally-shifted views of a scene, our model matches corresponding patches and uses their displacement to infer depth–objects that are further away shift more. Finally, our IMU-trained model can predict novel scene views based on different potential camera motions.

Reasoning. Simulating hypothetical and counterfactual possibilities is a core component of physical and causal reasoning (Gerstenberg, 2024). Our models succeed at simulating multi-object motions and interactions (Figure 2). In addition, infants show physical reasoning that objects do not spontaneously break into pieces (cohesion), two rigid objects cannot occupy the same space (solidity), and hands (agents) can cause other objects to move. Our models, capture these intuitions. All models predict single-object cohesion, but only the flow-based models and the models trained on non-developmental data succeeded in other tasks. This suggests motion (flow) is a useful intermediate representation.

Discussion

Human infants show remarkable visual cognition from a young age, reasoning about objects, agents, and their interactions (Spelke, 2022). Here we take a first step towards investigating how this kind of simulation and reasoning can emerge from a unified computational architecture trained on videos from human children. This work is an initial proof-of-concept and some work is ongoing; in particular, we train our models on processed BabyView representations extracted using two pretrained models (tokenizer and flow extractor); these are currently being replaced with BabyView-trained models. Nevertheless, this work is a step towards a computational proof-of-concept showing that visual cognitive abilities can be learned from developmental data with minimal innate priors.

References

Arterberry, M. E., & Kellman, P. J. (2016). *Development of perception in infancy: The cradle of knowledge revisited*. Oxford University Press.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2017, September). An egocentric perspective on active vision and visual object learning in toddlers. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 290–295). Lisbon: IEEE. Retrieved 2025-04-06, from http://ieeexplore.ieee.org/document/8329820/ doi: 10.1109/DEVLRN.2017.8329820

Gerstenberg, T. (2024, January). *Counterfactual simulation in causal cognition.* Retrieved 2025-01-08, from https://osf.io/72scr doi: 10.31234/osf.io/72scr

Lee, W., Kotar, K., Venkatesh, R. M., Watrous, J., Chen, H., Aw, K. L., & Yamins, D. L. K. (2025, April). *3D Scene Understanding Through Local Random Access Sequence Modeling.* arXiv. Retrieved 2025-04-08, from http://arxiv.org/abs/2504.03875 (arXiv:2504.03875 [cs]) doi: 10.48550/arXiv.2504.03875

Long, B., Xiang, V., Stojanov, S., Sparks, R. Z., Yin, Z., Keene, G. E., ... Frank, M. C. (2024, June). *The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences*. arXiv. Retrieved 2024-07-18, from http://arxiv.org/abs/2406.10447 (arXiv:2406.10447 [cs])

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020, December). *Self-supervised learning through the eyes of a child.* arXiv. Retrieved 2024-08-10, from http://arxiv.org/abs/2007.16189 (arXiv:2007.16189 [cs])

Sheybani, S., Hansaria, H., Wood, J. N., Smith, L. B., & Tiganj, Z. (2023). Curriculum Learning with Infant Egocentric Videos.

Spelke, E. S. (2022). *What babies know*. Oxford University Press.

Spelke, E. S., & Kinzler, K. D. (2007, January). Core knowledge. *Developmental Science*, *10*(1), 89–96. doi: 10.1111/j.1467-7687.2007.00569.x

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021, January). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118. doi: 10.1073/pnas.2014196118