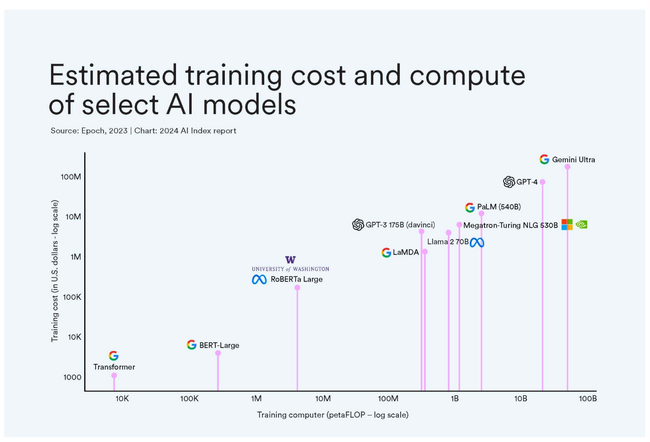
Establishing a Benchmark for Baby-Level Vision Foundation Models

Andrew Zhu[1, 2], David Li [2], Shawn Li[2], Jeffrey Li[2], Helen Chen[2], Andrew Zagula[2], Mimi Zhao[2], Joey Huang[2], Shengao Wang[2], Wenqi Wang[2], Zecheng Wang[2], Michael Wakeham[2], Boqing Gong[2]

Illinois Math and Science Academy, Aurora, IL[1]; Boston University, Boston, MA[2]

INTRODUCTION

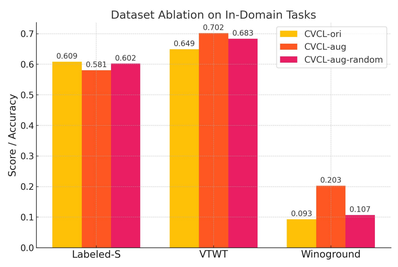
Modern Vision-Language Models (VLMs) that accept both image and text input have been become prohibitively large to train for researchers at university [1].



Because of the massive scale of these models, training state of the art models is not possible for researchers not in industry. As a result, researchers have shifted focus to efficiency of training these models.

One way is using data from babies; they develop incredibly fast in comparison to the costly VLMs.

Synthetic child-directed data helps the scores of the VLMs significantly [2].



Efforts to benchmark VLMs have been primarily focused on the large-scale models produced by industry.

These benchmarks are not appropriate for assessing VLMs training solely off of a baby [2].

Benchmark	Task Diversity	Baby-like	In-domain
General purpose (VQA [2], Winoground [48], etc.)	✓	✗	✗
DevBench [45]	✓	✓	✗
Labeled-S [31]	✗	✓	✓
ModelVsBaby [41]	✗	✓	✗
MEWL [16]	✓	✓	✗
BabyVLM	✓	✓	✓

To address this gap in the literature, we establish a novel benchmark that is founded in developmental psychology in order to accurately assess the strength of the VLMs.

Using the Mullen Learning Scales of Early Development and NIH baby toolbox [3], we adapt the following tasks:

BENCHMARKS

Visual Delayed Response (VDR)
Baby Version
A toy appears at the center of the screen and then moves off to be covered on either the left or right. After some time passes with a distraction covering the object, the baby determines where the object is hidden.

Visual Delayed Response (VDR)
Model Version
An object appears at the center of the screen and then moves off-screen in one of four directions – left, right, top, or bottom. The VFM must determine the direction of movement.

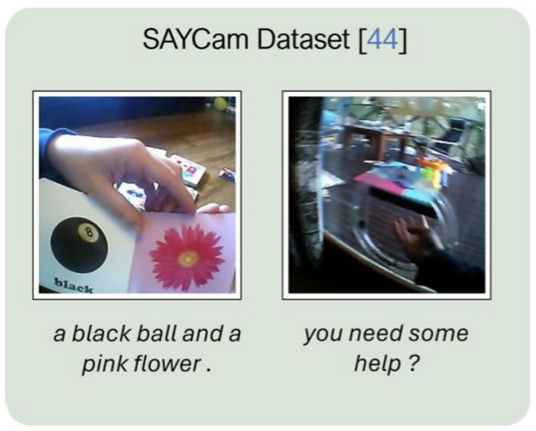
Memory Task (MT)
Baby Version
- Learning: The baby sees two animals; one they saw on the last page and one new one.
- Testing: The baby sees two animals; one that they saw in learning and one that is totally new. The baby must click on the new one.

Memory Task (MT)
- Learning: The model views overlapping image pairs.
- Testing: It sees randomized pairs containing one learned image and one novel image and must select the learned image.

Picture Vocabulary (PV)
The VFM is presented with an image that contains 2–4 objects with labeled bounding boxes. The VFM must select the correct bounding box based on a referring prompt (e.g., “identify the car”)

Picture Vocabulary (PV)
The VFM is presented with an image that contains 2–4 objects with labeled bounding boxes. The VFM must select the correct bounding box based on a referring prompt (e.g., “identify the car”)

METHODS

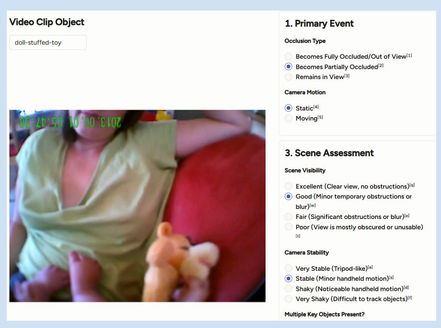


We use ChatGPT annotation as well as open-set object detection in order to create the candidates for our tasks.

Firstly, we use ChatGPT to create annotations from 500+ hours of the SayCAM dataset.

Using a sliding window, ChatGPT identified clips where the object became occluded.

Further, YOLO-E, an open-set object detector verified that the generated clips were not hallucinated.




```
frame 0005: key_object=Book
frame 0006: key_object=Book
frame 0007: key_object=Book
frame 0008: key_object=Chair
frame 0009: key_object=Chair
frame 0010: key_object=Hand
```

In total, we identified 3543 possible clips, which were then manually labeled. We place the statistics below. In total we identified 2220 usable clips.

Camera Stability	Count
Very Stable	16
Stable	2646
Shaky	691
Very Shaky	190

Occlusion Type	Count
Fully Occluded	2480
Partially Occluded	572
Remains in View	491

Scene Visibility	Count
Excellent	410
Good	1892
Fair	931
Poor	310

REFERENCES

[1] AI Index Steering Committee. The 2025 AI Index Report; Stanford Institute for Human-Centered Artificial Intelligence: Stanford, CA, 2025.
[2] Wang, S.; Chandra, A.; Liu, A.; Saligrama, V.; Gong, B. BabyVLM: Data-Efficient Pretraining of VLMs Inspired by Infant Learning; 2025.
[3] Gershon, R., Novack, M. A., & Kaat, A. J. (2024). The NIH Infant and Toddler Toolbox: A new standardized tool for assessing neurodevelopment in children ages 1–42 months. Child development, 95(6), 2252–2254.
Marchman, V. A.; Dale, P. S. The MacArthur-Bates Communicative Development Inventories: Updates from the CDI Advisory Board. Frontiers in Psychology 2023, 14. https://doi.org/10.3389/fpsyg.2023.1170303.

ACKNOWLEDGEMENTS

I thank Professor Boqing Gong for his guidance and support, our PhD mentors Shengao, Wenqi and Victor Wang for their mentorship, and the RISE program for facilitating this opportunity.