A Developmentally-Inspired Vocabulary Benchmark for Infant-Level Vision-Language Models

Helen Chen[1], Mimi Zhao[2], Zecheng Wang[3], Shengao Wang[3], Wenqi Wang[3] , Boqing Gong[3]

West Windsor-Plainsboro High School South, Princeton Junction, NJ[1], Canyon Crest Academy, San Diego, CA[2], Boston University, Boston, MA[3]

Introduction

Problem Setup

- **Vision-Language Models (VLMs)** are foundation models which take both text and visual input
  - However, they currently take vast amounts of data and computational power to train
- Compared to VLMs, babies learn quickly with minimal stimuli exposure
- Inspired by baby learning, "baby" VLMs aim to use this fact by training on baby datasets like **SAYCam**[1,2]
- Lack of appropriate benchmarks available to test baby-level VLMs

Our Task

- Adapted developmental psychology tests (i.e. **NIH Baby Toolbox**[4]) to VLM benchmark
  - NIH Toolbox measures baby development
  - Adapted the **Picture Vocabulary** test from the NIH Baby Toolbox
  - In this task, babies listen to a word and are required to select the correct corresponding image out of 4 images

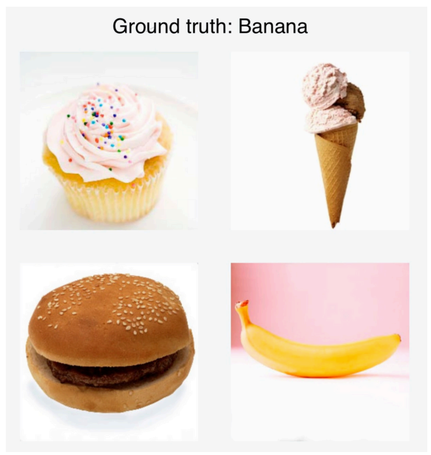

Figure 1: Example test question from NIH Baby Toolbox

Results

Gathered 47 initial test samples right from Baby Toolbox

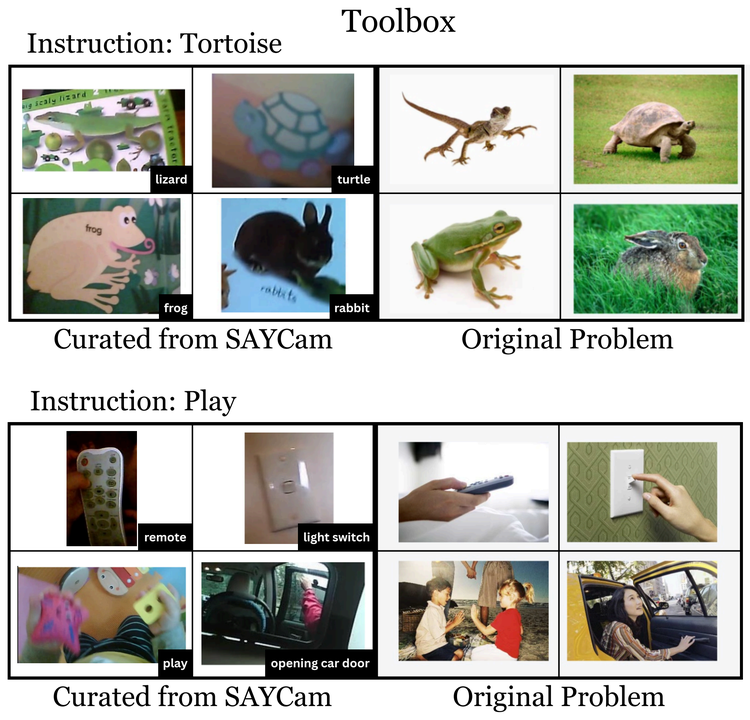

Figure 3: Examples of gathered test samples

Generated 409 questions and 1227 total distractors

Generated Problems			
Unrelated:	24.78%	Semantic:	15.08%
Phonological:	2.85%	Category:	57.29%

Original Problems (Baby-Level)			
Unrelated:	25.64%	Semantic:	14.72%
Phonological:	3.21%	Category:	56.43%

Figure 4: Statistics for generated problems compared to original problems

Discussion

Impact:

- Benchmark will help evaluate infant-level VLMs
  - Aid the work in developing smaller scale models mimicking babies that perform well on this benchmark
  - Helps in finding a way to require less resources for VLM pre-training
- Lab plans to host a BabyVLM competition using this benchmark as an evaluation metric for submitted models

Limitations and future work:

- Test foundation models on this benchmark
- Improve annotations on the original dataset
  - Current annotations are GPT generated
  - May not fully capture all frame objects and might affect matching quality
- Find more sources of baby vocabulary outside of MAB-CDI
  - What defines baby vocabulary?
    - Conduct literature review
- CLIP score is unreliable for this application
  - Try detecting every object from every frame of dataset to skip the CLIP score step for future passes in the pipline
  - Select object using a combination of clip score and blurriness rating

Methods

Finding Initial Test Sets

- Collected **test examples** from the NIH Baby Toolbox iPad App and words from the MacArthur-Bates Communicative Development Inventory **(MAB-CDI)**[3] vocabulary list.
- Selected appropriate level questions by matching groundtruth to CDI words
- Generated labels for each image with GPT and manually screened
- Matched exact labels with SayCam annotations
- Used embedding based similarity matching to find top five frames per label
- Generated prelabels for cropping with **GroundingDINO**
- Used Label Studio to select frames and crop
- Sorted images and visualized final examples

Creating New Test Sets

- Gathered statistics on distractors
  - Soundex for phonologically similar
- Direct embendding and CLIP score for semantically similar
- K means clustering (K = 100) for same category
- All other "distractors" were considered unrelated

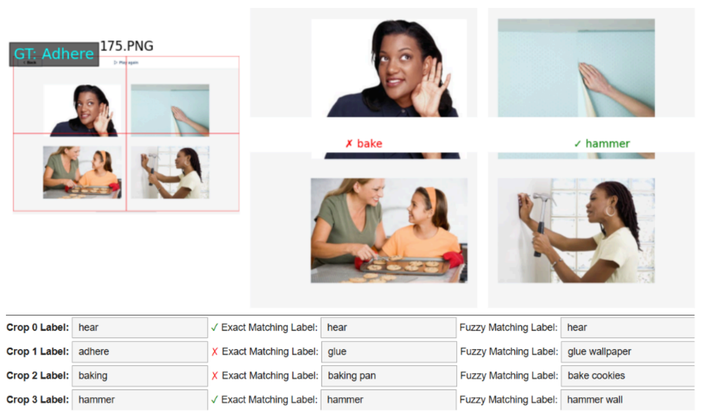

Figure 2: Manual labeling interface

References

1. Wang, S., Chandra, A., Liu, A., Saligrama, V., & Gong, B. (2025). BabyVLM: Data-Efficient Pretraining of VLMs Inspired by Infant Learning. arXiv preprint arXiv:2504.09426.
2. 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 2021, 1–10.
3. 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.
4. Gershon, R.; Novack, M. A.; Kaat, A. J. The NIH Infant and Toddler Toolbox: A New Standardized Tool for Assessing Neurodevelopment in Children Ages 1–42 Months. Child Dev. 2024, 95 (6), 2252–2254. https://doi.org/10.1111/cdev.14135.

Acknowledgements

I would like to thank to our PI, Professor Boqing Gong, and our mentors Zecheng Wang, Shengao Wang, and Wenqi Wang for all the support, guidance, and encouragement over the past few weeks. I've learned so much from you, and I really appreciate everything you've done to help us grow.