Evaluating Baby-Scale Vision-Language Models with a Developmentally-Inspired Vocabulary Task

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

Canyon Crest Academy, 5951 Village Center Loop Rd, San Diego, CA 92130[1], West Windsor-Plainsboro High School South, 346 Clarksville Rd, Princeton Junction, NJ 08550[2], Boston University, Commonwealth Ave, Boston, MA 02215[3]

Introduction

- **Vision-language models (VLMs)** are models that take text and visual input
  - Usually require a huge amount of data and computing resources to train
- In comparison, babies develop visual learning skills quickly with little input
- Recent studies introduced baby-level frameworks to train VLMs with more efficiency [1]
  - Existing benchmarks are mostly designed for large-scale models and don't align with testing baby-level VLMs
- Targeted fundamental skills of cognitive and language development by adapting **NIH Baby Toolbox** Tasks for testing VLMs [2]
  - Standardized tool for assessing cognitive development in children 1-42 months old
  - We specifically focused on the **Picture Vocabulary Test**
  - A word is verbally given to the child who has to match it to one of four images

Prompt: Banana

Figure 1: Example test question from NIH Baby Toolbox

Methods

Finding Initial Set

- Selected Picture Vocabulary prompts that are included in the **MAB-CDI** (MacArthur-Bates Communicative Development Inventories) [3], a baby-level vocabulary database
- Used Large Language Models such as ChatGPT to generate baby-level labels for the three distractor images
- Matched prompts from Picture Vocabulary and generated labels to **SAYCam** annotations, a video dataset of babies aged 6-32 months daily activities [4]
- Used GroundingDINO and ChatGPT with open-source object detection models to generate pre labels and extract object croppings from relevant SAYCam frames
- Manually screened images at each step to ensure quality of the resulting problem sets

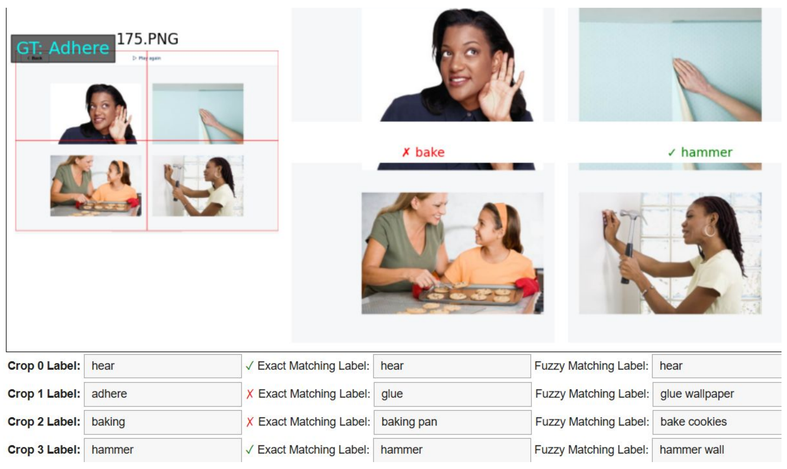


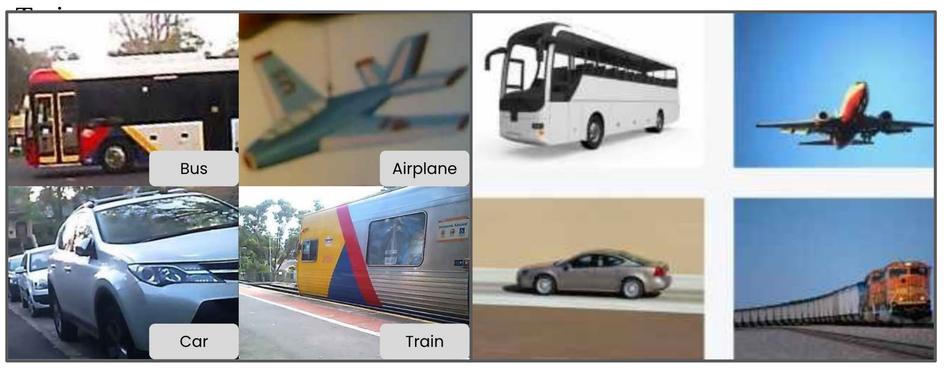
Figure 2: Manual Screening Interface

Creating New Test Sets

- Found distractors that semantically, categorically, and phonologically similar to the target word
  - Semantical: CLIP (Contrastive Language–Image Pretraining) Similarity Scoring
  - Categorical: K-Means clustering algorithm over CLIP embeddings
  - Phonological: Soundex algorithm
  - Other distractors were considered "unrelated"
- Reflected statistical properties of the original test set from NIH Baby Toolbox

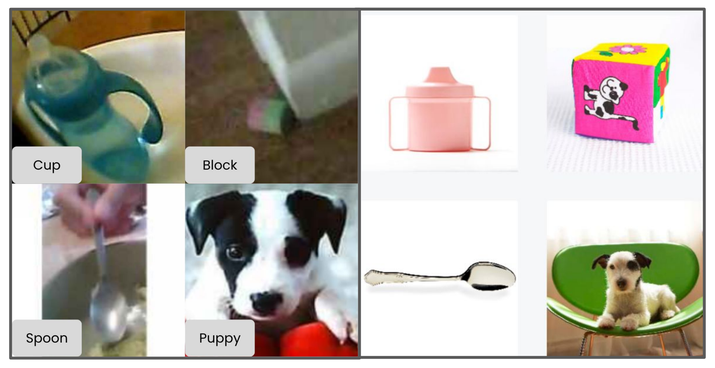
Results

Instructional Prompt:



Extracted from SAYCam — Original Problem

Instructional Prompt: Spoon



Extracted from SAYCam — Original Problem

Figure 3: Examples of gathered initial test samples

Generated Problems:

Total questions: 409
Total distractors: 1227

Generated Problems				Original Problems (Baby-Level)			
Unrelated:	24.78%	Semantic:	15.08%	Unrelated:	25.64%	Semantic:	14.72%
Phonological:	2.85%	Category:	57.29%	Phonological:	3.21%	Category:	56.43%

Discussion

Impact

- This benchmark:
  - Allows a systematic evaluation of how much baby-scale vision-language models approximate the characteristics of actual infant-like intelligence
  - Guides researchers in developing models that better simulate early human learning, potentially leading to safer, more human-aligned AI

Limitations and Future Works

- Use this benchmark to test foundation models
- Find more sources of baby-level vocabulary outside of MAB-CDI
- Improve annotations on original SAYCam dataset
- Develop ways to automate matching and extracting problems better to skip manual screening at each stage
- Refine method for finding new distractors to increase effectiveness
  - Some selected categorical distractors lack clear categorical relevance to the ground truth (eg., "album" and "crisp" for ground truth "man")
  - Phonological distractors often do not have enough sound similarity to the ground truth
  - Develop a more targeted algorithm for assigning distractor types to each target word, as some words may be better suited for certain types based on their linguistic properties

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. 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.
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. 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.

Acknowledgements

I would like to extend my deepest gratitude to Professor Boqing Gong and our mentors Zecheng Wang, Shengao Wang, and Wenqi Wang for their tremendous support, patience, and guidance throughout this research opportunity. You have made this experience so much more valuable to me and I am incredibly grateful for everything I learned from you this summer.