BabyVLM: Data-Efficient Pretraining of VLMs Inspired by Infant Learning

Shengao Wang Arjun Chandra Aoming Liu Venkatesh Saligrama Boqing Gong

Boston University

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Introduction

The North Star Question:

Can we achieve infant-level intelligence using infants' sensory data only?

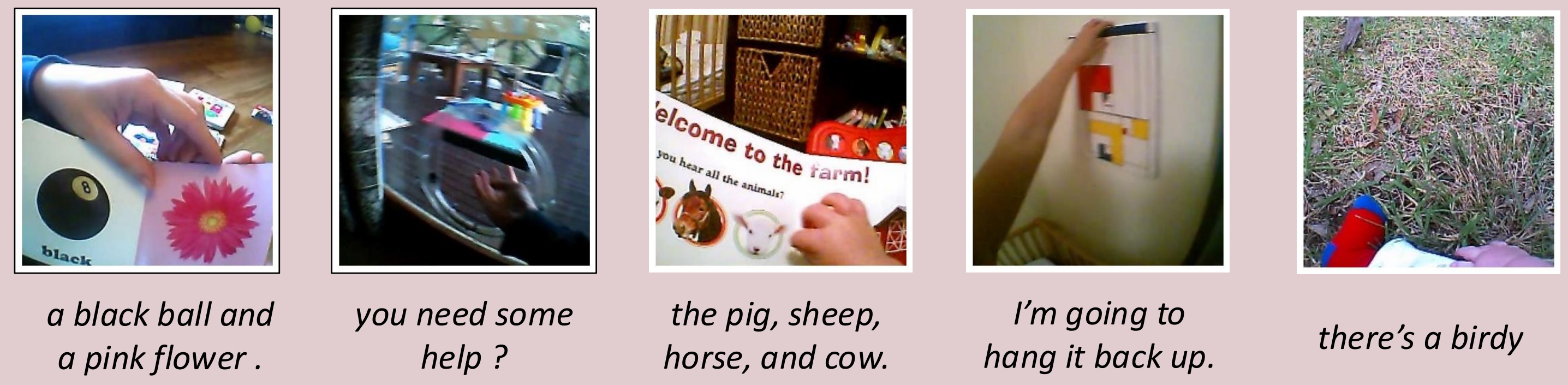
We provide a framework of…

- A developmental training dataset curated for Vision Language Model (VLM) pretraining.
- A set of in-domain VLM evaluation tasks.
- A O(M)-scale baseline VLM trained from scratch purely on infant data.

Datasets

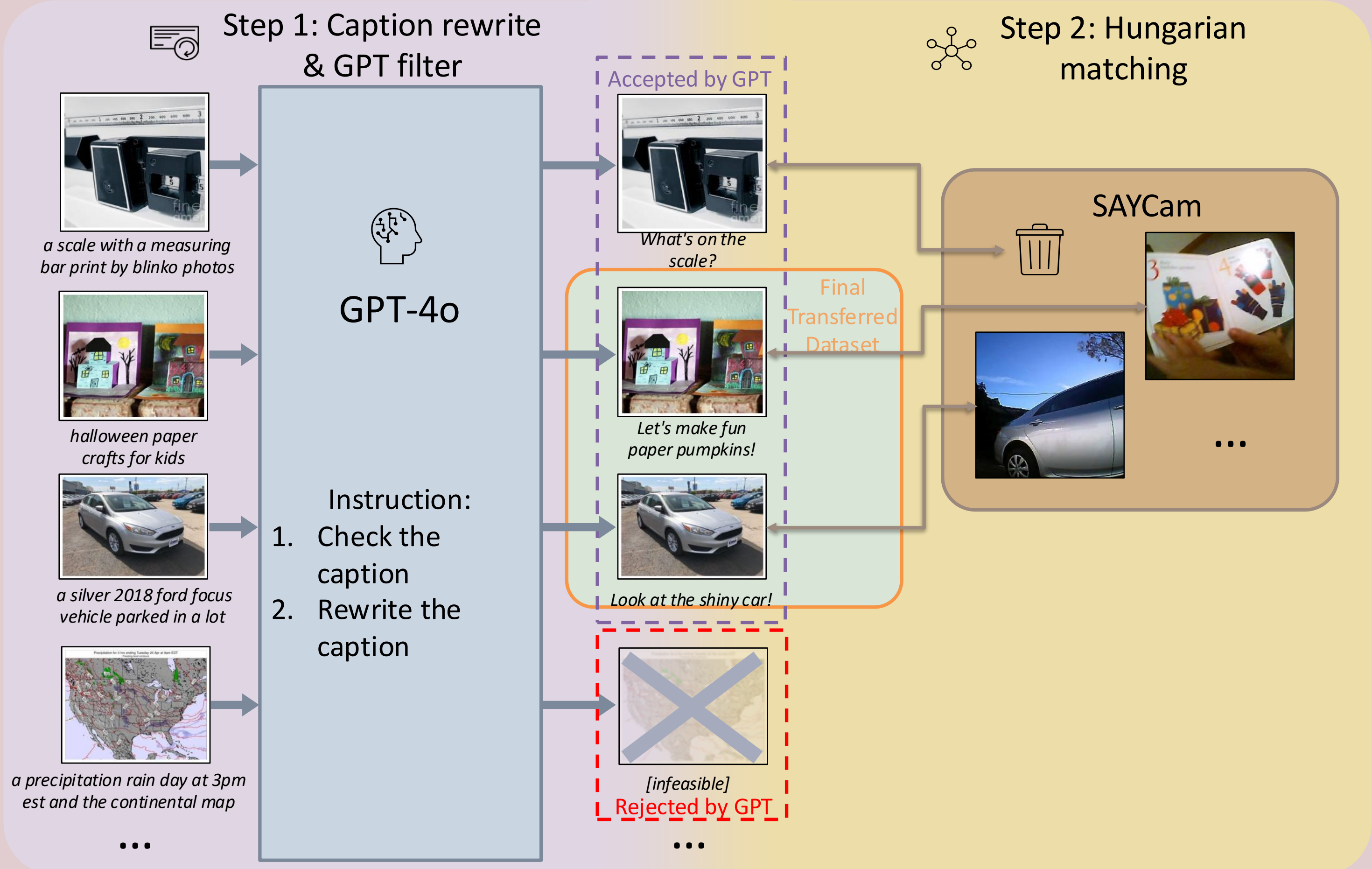
Filtered SAYCam dataset:

- Egocentric, longitudinal, audiovisual dataset collected by infants (6-32 months).
- Child-directed utterances + sample frames.
- High-quality image-utterance pairs (CLIP similarity > 0.2 → 67K pairs).


a black ball and a pink flower .
you need some help ?
the pig, sheep, horse, and cow.
I'm going to hang it back up.
there's a birdy

Transferred dataset:

- Start from CC3M / LAION / SBU, transfer the caption into child-directed utterances


Step 1: Caption rewrite & GPT filter
a scale with a measuring bar print by blinko photos
halloween paper crafts for kids
a silver 2018 ford focus vehicle parked in a lot
a precipitation rain day at 3pm est and the continental map
…
GPT-4o
Instruction:
1. Check the caption
2. Rewrite the caption
Accepted by GPT
What's on the scale?
Let's make fun paper pumpkins!
Look at the shiny car!
Final Transferred Dataset
[infeasible]
Rejected by GPT
…
Step 2: Hungarian matching
SAYCam
…

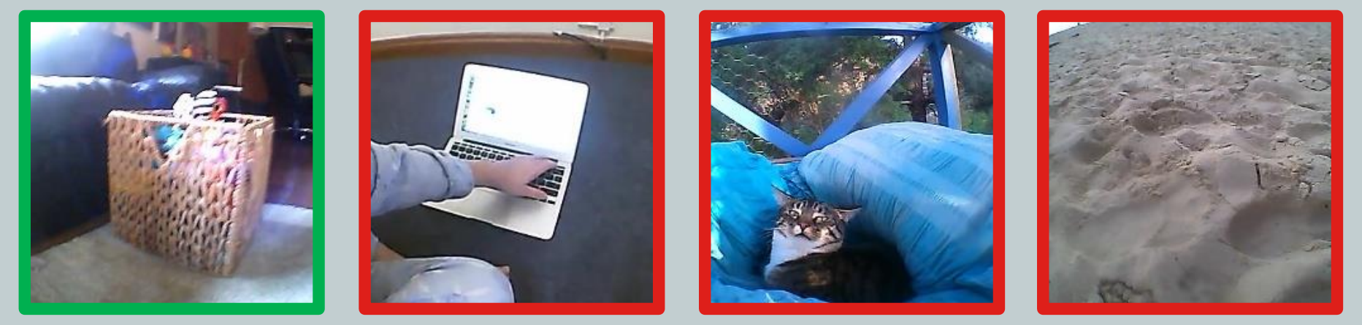
Evaluation Tasks

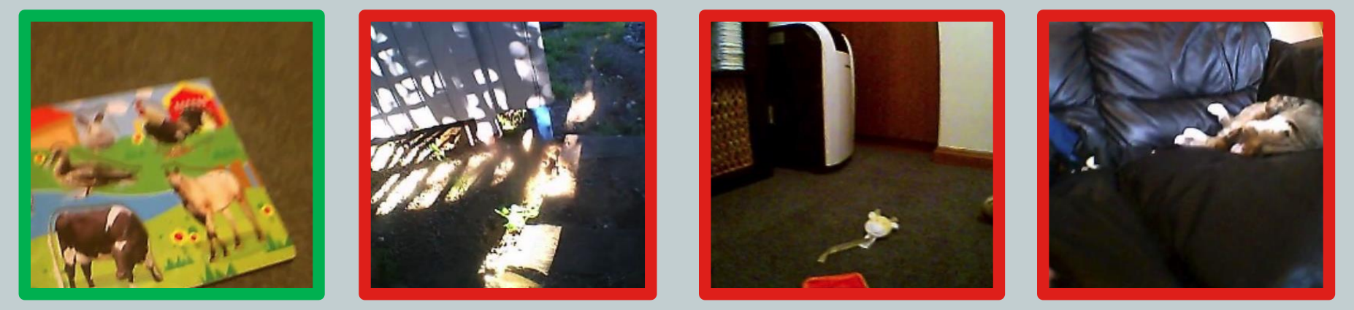
Principle:

- Comparable to developmental milestones of infants
- Testable for computer vision models
- In the same domain as the training data

Labeled-S:

Match target category with correct image (24 classes).


basket


puzzle

Visual Two-Word Test (VTWT)

Match SAYCam images with simple two-word phrases, reflecting the "two-word stage" typical of two-year-old children.

spread jam ✓

cut bread ✗

teddy eat ✓

teddy sleep ✗

Baby Winoground

Two images + two phrases; test compositional reasoning with synthetic distractors.

play keys

play drum

climb up

dig hole

SAYCam Caption

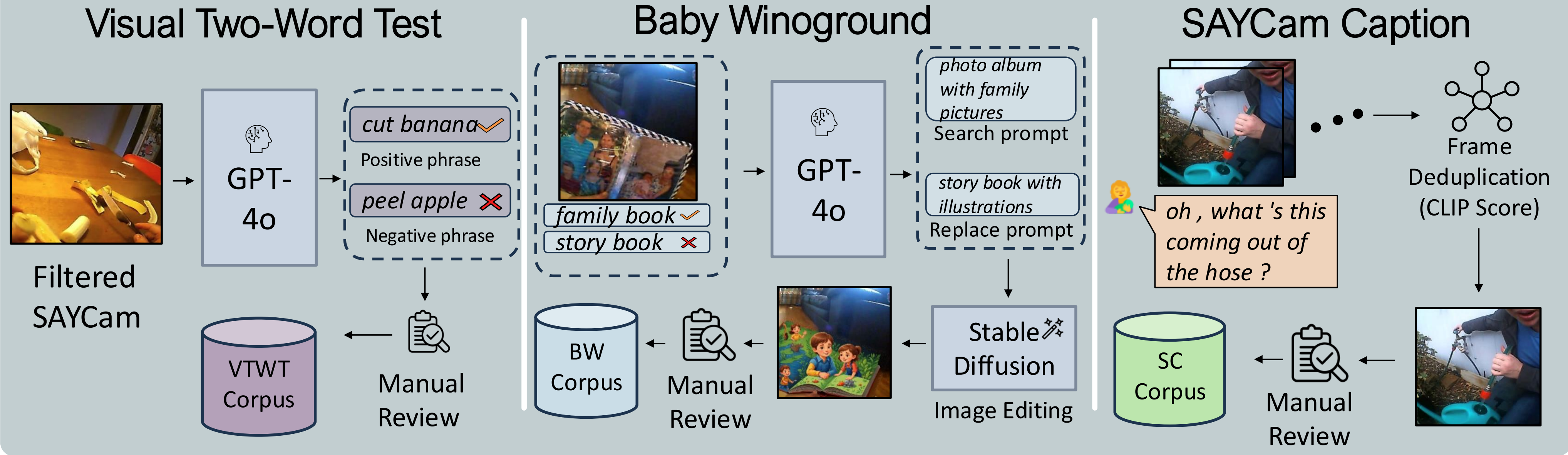
Generate child-directed captions for SAYCam frames, evaluated by METEOR score.

we have all these books.

METEOR score: 0.0820

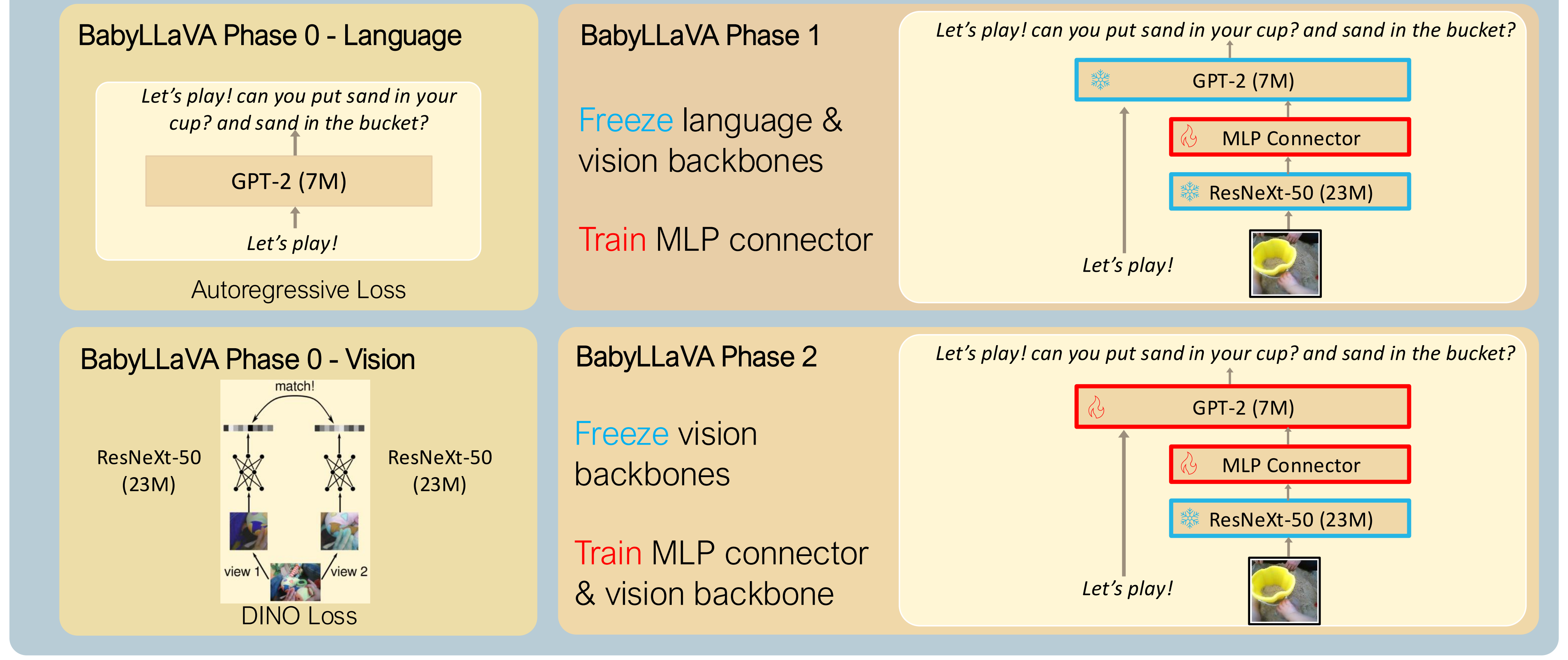
You want to read about books?

Task Generation Pipeline

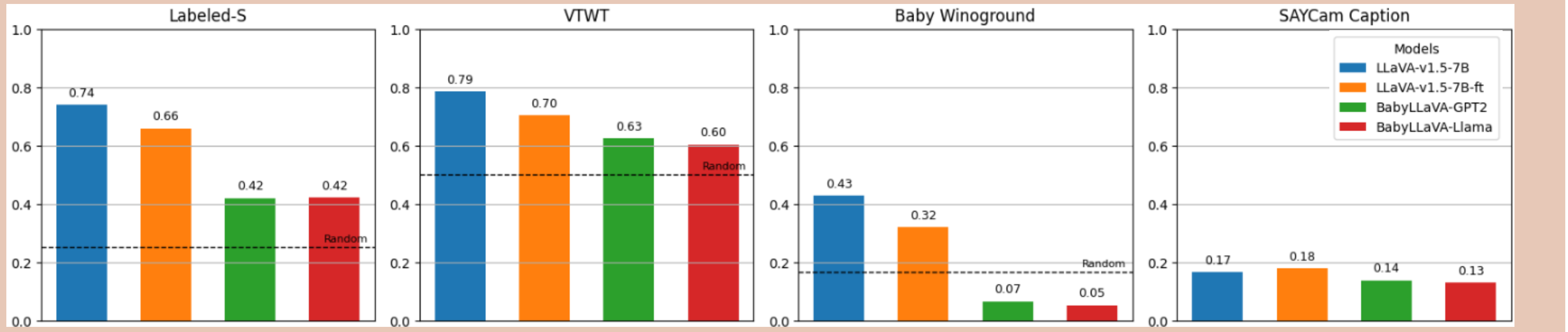


Baseline Model (BabyLLaVA)

- Trained from scratch purely on SAYCam / the transferred dataset
- Inspired by LLaVA, but compact and developmentally constrained.


BabyLLaVA Phase 0 - Language
Let's play! can you put sand in your cup? and sand in the bucket?
GPT-2 (7M)
Let's play!
Autoregressive Loss
BabyLLaVA Phase 1
Freeze language & vision backbones
Train MLP connector
Let's play! can you put sand in your cup? and sand in the bucket?
GPT-2 (7M)
MLP Connector
ResNeXt-50 (23M)
Let's play!
BabyLLaVA Phase 0 - Vision
match!
ResNeXt-50 (23M)
ResNeXt-50 (23M)
view 1
view 2
DINO Loss
BabyLLaVA Phase 2
Freeze vision backbones
Train MLP connector & vision backbone
Let's play! can you put sand in your cup? and sand in the bucket?
GPT-2 (7M)
MLP Connector
ResNeXt-50 (23M)
Let's play!

Experiments



Observation:

- The infant-directed SAYCam data contains rich information for models to perform various tasks.
- LLaVA fine-tuned on SAYCam data degrades, likely due to the dataset's limited image-text alignment.
- When dataset is limited, simply upscaling model doesn't help.

[1] J. Sullivan, M. Mei, A. Perfors, E. Wojcik, and M. C. Frank, "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective," Open Mind, vol. 5, pp. 20–29, 2021.

[2] W. K. Vong, W. Wang, A. E. Orhan, and B. M. Lake, "Grounded language acquisition through the eyes and ears of a single child," Science, vol. 383, no. 6682, pp. 504–511, Feb. 2024.

[3] A. E. Orhan and B. M. Lake, "Learning high-level visual representations from a child's perspective without strong inductive biases," *arXiv preprint arXiv:2305.15372*, 2023.