BOSTON UNIVERSITY

Benchmarking Small-Scale Vision-Foundation Models through Developmentally-Inspired Cognitive Assessments

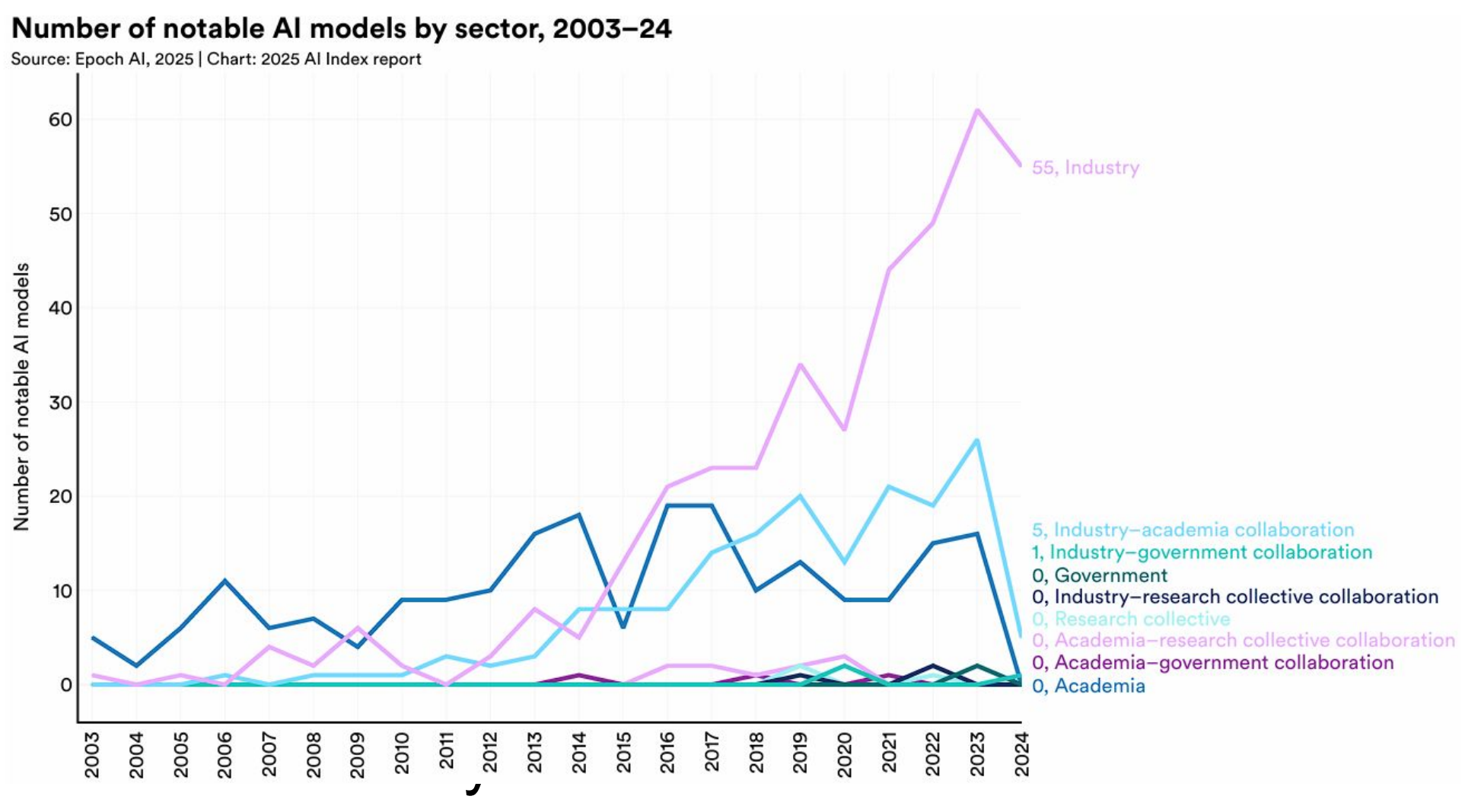
A New Ecosystem for Academic Pretraining Research

David Li[1, 2], Shawn Li[2], Jeffrey Li[2], Helen Chen[2], Andrew Zhu[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]

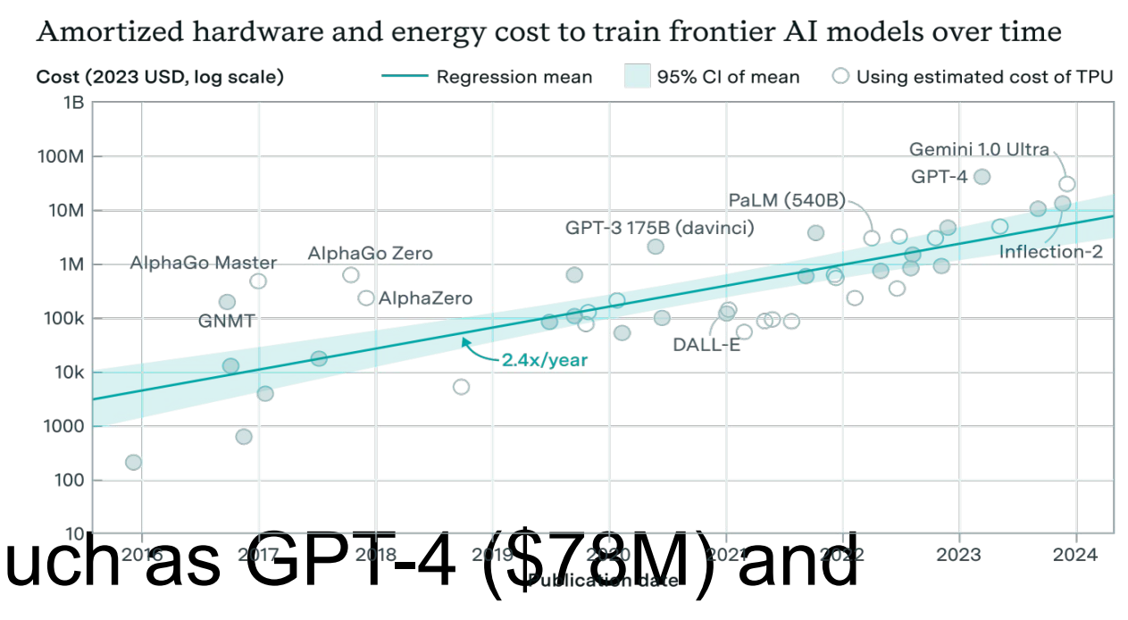
Mills High School, Millbrae, CA[1]; Boston University, Boston, MA[2]

INTRODUCTION

The massive resource and data demands of modern Vision-Foundation Models (VFMs) have created a structural divide, pushing fundamental pretraining research out of academia.


Number of notable AI models by sector, 2003–24

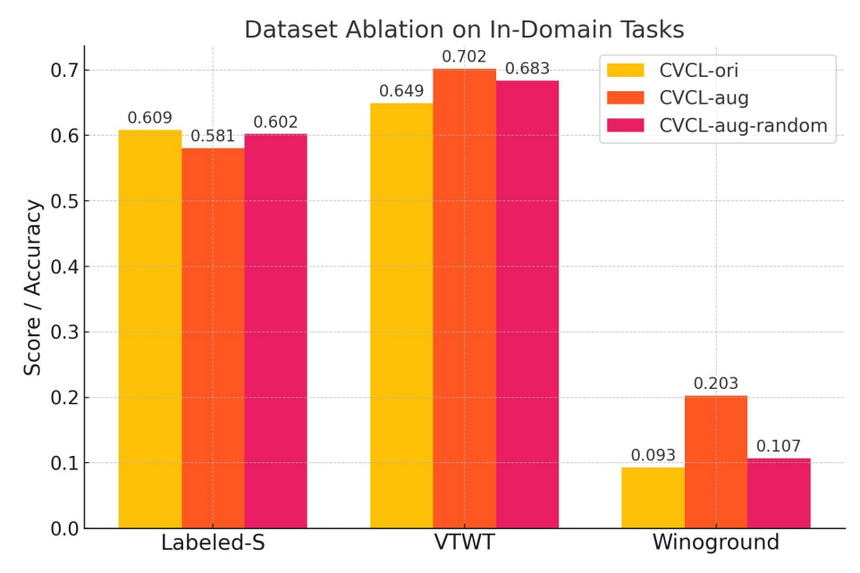
Source: Epoch AI, 2025 | Chart: 2025 AI Index report


Amortized hardware and energy cost to train frontier AI models over time

developed 55 notable ML models, academia developed 5.[1]

- Training cutting-edge models such as GPT-4 ($78M) and Gemini Ultra ($191M) vastly exceeds academic funding.[1] In 2024, global private AI investment reached $252.3 billion.[2]
- With academia is unable to match industry's scale, research has shifted to fine tuning corporate models, limiting progress.
- Rather than chasing scale, we aim to use in-domain data and benchmarks grounded in developmental psychology, lowering the barrier to entry for fundamental research.

BabyVLM shows Synthetic child-directed data (CVCL-aug) boosts VTWT and Winoground scores much more than random augmentation.[3]



METHODS

Data Foundation: SAYCam Dataset[4]

- 472 h of first-person infant video recordings
- Computationally tractable proxy for human development
- Practical academic alternative to petabyte-scale industrial datasets

Benchmark Foundation: NIH Baby Toolbox Tasks[5]

- Visual Delayed Response: attention, memory, executive control
- Picture Vocabulary Test: receptive language (adapted from MacArthur-Bates CDI)[6]
- Memory Task: processing speed and learning efficiency
- Covers developmental age range of 1–42 months for scientific validity

Infrastructure Foundation

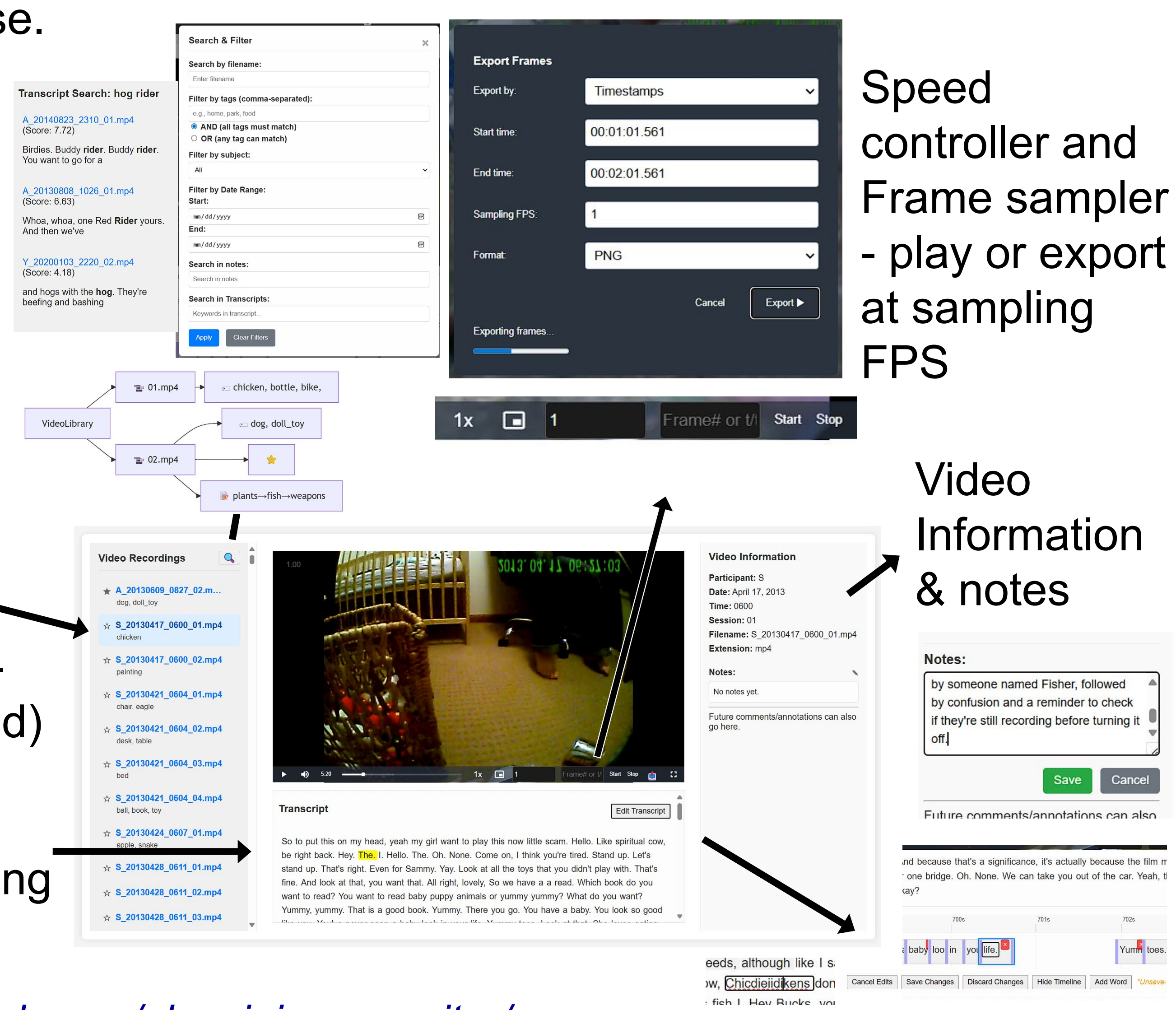
- Web-based GUI enabling complex SAYCam metadata queries
- Standardized model wrapper for consistent interfacing
- Automated execution on a private test set & Public leaderboard for transparent, reproducible comparisons

INFRASTRUCTURE

To bridge large-scale naturalistic data with standardized model evaluation, we developed a comprehensive infrastructure ecosystem. These tools enable researchers to efficiently curate data and benchmark models in a reproducible, scalable manner.

SAYCam Video & Transcript Navigator

- Manually navigating the 472-hour SAYCam dataset is prohibitively time-consuming.
- The provided Databrary software is severely lacking. No ability to speed up videos, view/edit their transcripts, categorize them, export them.
- Our browser-based navigator transforms this raw footage into a searchable, interactive database, enabling researchers to explore developmental data without needing data management expertise.


In-depth search: transcript, tags, etc. Metadata storage:

Speed controller and Frame sampler - play or export at sampling FPS

Star and Tags (comma-separated)

Clickable, auto-scrolling transcript

Video Information & notes

Timeline Transcript Editing: draggable / resizable words

Transcript Search: hog rider

A_20140823_2310_01.mp4 (Score: 7.72)

Birdies. Buddy rider. Buddy rider. You want to go for a

A_20130808_1026_01.mp4 (Score: 6.63)

Whoa, whoa, one Red Rider yours. And then we've

Y_20200103_2220_02.mp4 (Score: 4.18)

and hogs with the hog. They're beefing and bashing

Search & Filter

Search by filename: Enter filename

Filter by tags (comma-separated): e.g., home, park, food

● AND (all tags must match)

○ OR (any tag can match)

Filter by subject: All

Filter by Date Range:

Start: mm/dd/yyyy

End: mm/dd/yyyy

Search in notes: Search in notes

Search in Transcripts: Keywords in transcript...

Export Frames

Export by: Timestamps

Start time: 00:01:01.561

End time: 00:02:01.561

Sampling FPS: 1

Format: PNG

Exporting frames...

Video Recordings

★ A_20130609_0827_02.m... dog, doll_toy

☆ S_20130417_0600_01.mp4 chicken

☆ S_20130417_0600_02.mp4 painting

☆ S_20130421_0604_01.mp4 chair, eagle

☆ S_20130421_0604_02.mp4 desk, table

☆ S_20130421_0604_03.mp4 bed

☆ S_20130421_0604_04.mp4 ball, book, toy

☆ S_20130424_0607_01.mp4 apple, snake

☆ S_20130428_0611_01.mp4

☆ S_20130428_0611_02.mp4

☆ S_20130428_0611_03.mp4

Video Information

Participant: S

Date: April 17, 2013

Time: 0600

Session: 01

Filename: S_20130417_0600_01.mp4

Extension: mp4

Notes: No notes yet.

Future comments/annotations can also go here.

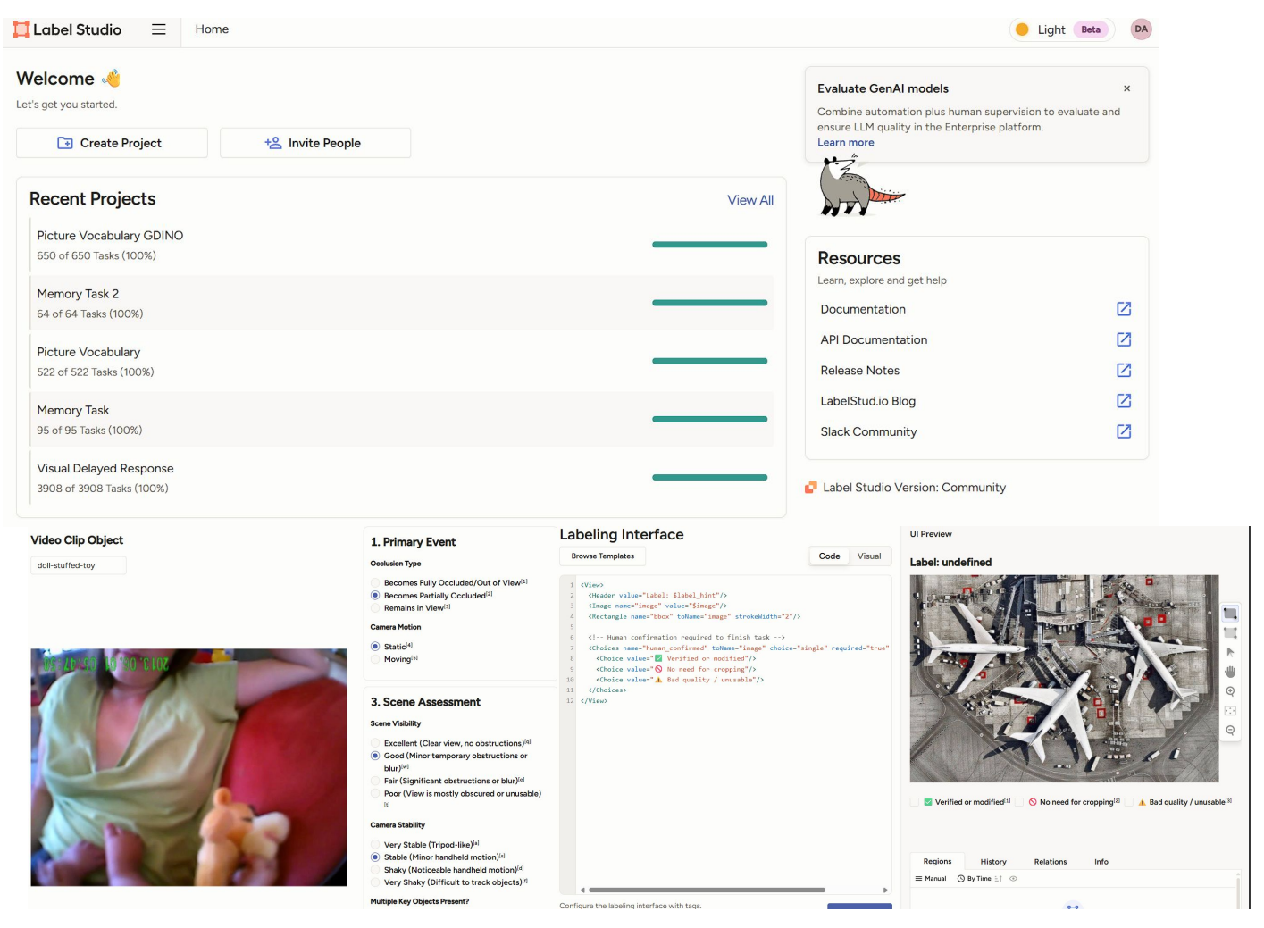
Notes: by someone named Fisher, followed by confusion and a reminder to check if they're still recording before turning it off.

Transcript

So to put this on my head, yeah my girl want to play this now little scam. Hello. Like spiritual cow, be right back. Hey. The. I. Hello. The. Oh. None. Come on, I think you're tired. Stand up. Let's stand up. That's right. Even for Sammy. Yay. Look at all the toys that you didn't play with. That's fine. And look at that, you want that. All right, lovely, So we have a a read. Which book do you want to read? You want to read baby puppy animals or yummy yummy? What do you want? Yummy, yummy. That is a good book. Yummy. There you go. You have a baby. You look so good

https://github.com/eleusinianexpositor/saycam_visualizer

To ensure benchmark validity, we built a scalable labeling pipeline with Label Studio, automating clip ingestion into verifiable annotation tasks. Custom interfaces for tasks like Visual Delayed Response were rapidly developed using Gemini, streamlining iteration.


Label Studio

Welcome

Let's get you started.

Recent Projects

Picture Vocabulary GDINO	650 of 650 Tasks (100%)
Memory Task 2	64 of 64 Tasks (100%)
Picture Vocabulary	522 of 522 Tasks (100%)
Memory Task	95 of 95 Tasks (100%)
Visual Delayed Response	3908 of 3908 Tasks (100%)

Evaluate GenAI models

Combine automation plus human supervision to evaluate and ensure LLM quality in the Enterprise platform.

Label Studio Version: Community

Video Clip Object: doll-stuffed-toy

1. Primary Event

Occlusion Type: Becomes Fully Occluded/Out of View / ● Becomes Partially Occluded / Remains in View

Camera Motion: ● Static / Moving

3. Scene Assessment

Scene Visibility: Excellent (Clear view, no obstructions) / ● Good (Minor temporary obstructions or blur) / Fair (Significant obstructions or blur) / Poor (View is mostly obscured or unusable)

Camera Stability: Very Stable (Tripod-like) / ● Stable (Minor handheld motion) / Shaky (Noticeable handheld motion) / Very Shaky (Difficult to track objects)

Multiple Key Objects Present?

Labeling Interface

```
<View>
  <Header value="Label: $label_hint"/>
  <Image name="image" value="$image"/>
  <Rectangle name="bbox" toName="image" strokeWidth="2"/>

  <!-- Human confirmation required to finish task -->
  <Choices name="human_confirmed" toName="image" choice="single" required="true">
    <Choice value="✅ Verified or modified"/>
    <Choice value="🚫 No need for cropping"/>
    <Choice value="⚠ Bad quality / unusable"/>
  </Choices>
</View>
```

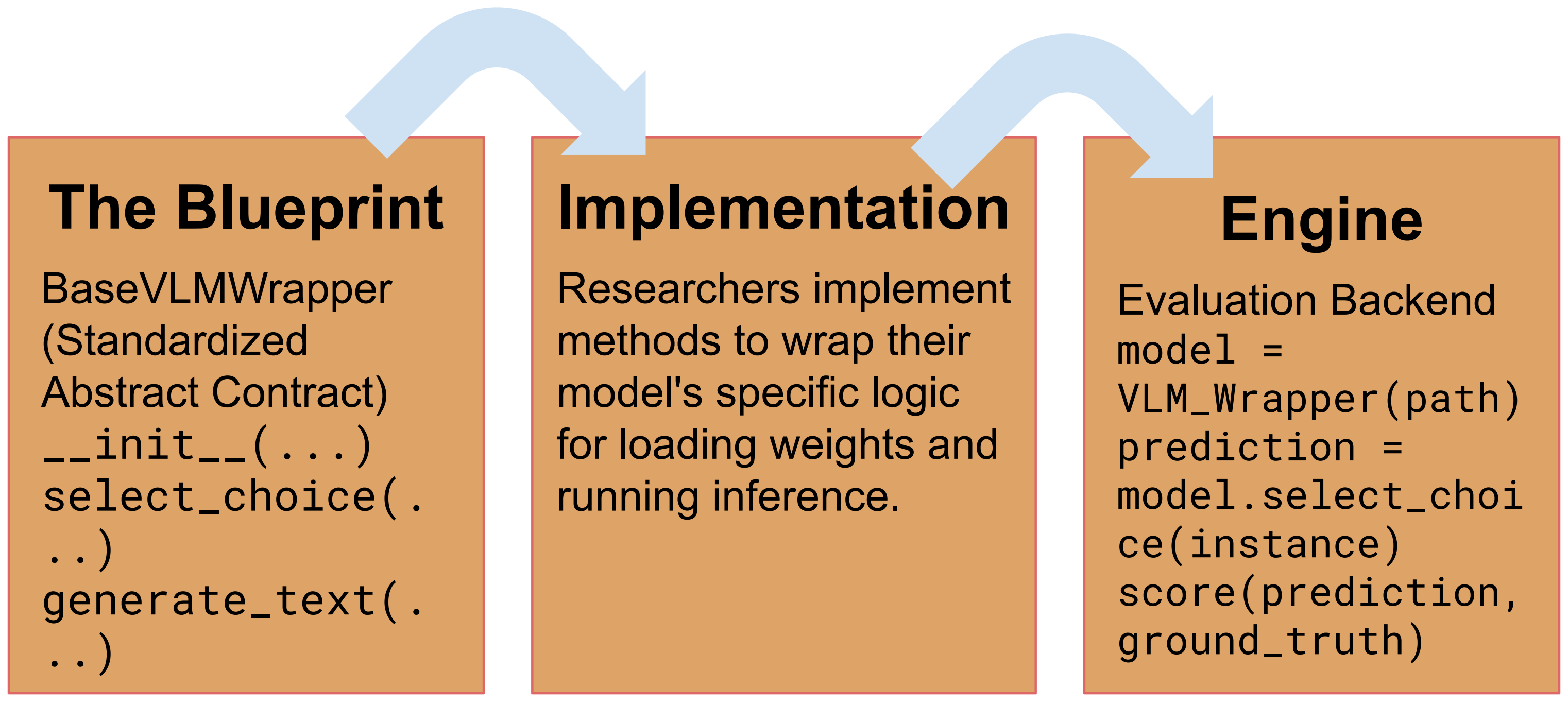
UI Preview

Label: undefined

Verified or modified / No need for cropping / Bad quality / unusable

We designed a **Model Evaluation Framework** that defines a standardized interface for model submission.

- Consistent, fair evaluation under identical conditions
- Streamlined integration accelerates computational research
- Automated contest foundation underpins a scalable model


The Blueprint

BaseVLMWrapper (Standardized Abstract Contract)

```
__init__(...)
select_choice(...)
generate_text(...)
```

Implementation

Researchers implement methods to wrap their model's specific logic for loading weights and running inference.

Engine

Evaluation Backend

```
model = VLM_Wrapper(path)
prediction = model.select_choice(instance)
score(prediction, ground_truth)
```

The final element is a **Public Contest Website**. This site will provide:

- open access to all benchmark documentation and starter code
- manage model submissions via our wrapper framework
- display results on a live, transparent leaderboard

BENCHMARKS

Visual Delayed Response (VDR)

An object appears in screen and then moves off-screen, left, right, up, or down. The VFM must identify direction.

Extracted 2200 usable clips from SAYCam, with diversity in direction and object type, yielding robust benchmark.

Memory Task (MT)

- *Learning: View overlapping image pairs.*
- *Testing: Given one old and one new image; pick the old.*

81 varying baby-level animal / object candidates found from SAYCam, can be combined to create adequate pairs.

Picture Vocabulary (PV)

The VFM sees an image with 2–4 labeled objects and must pick the correct bounding box based on a prompt.

Generated 409 questions and 1227 total semantic, phonological, and categorical distractors.

DISCUSSION

- Grounded in developmental science, our benchmarks level the playing field.
- Open-source tools make exploration and contribution to this new research accessible.
- Task suite will be broadened to include additional tasks, from NIH Baby Toolbox and other resources like Mullen Scales of Early Learning.
- We also plan to scale up our evaluation engine to support external model submissions and to refine our annotation tools with semi-automated methods.
- We release these resources publicly to catalyze a effort around data-efficient, developmentally-inspired pretraining.

REFERENCES

(1) AI Index Steering Committee. The 2025 AI Index Report; Stanford Institute for Human-Centered Artificial Intelligence: Stanford, CA, 2025.

(2) Cottier, B.; Rahman, R.; Fattorini, L.; Maslej, N.; Besiroglu, T.; Owen, D. The Rising Costs of Training Frontier AI Models; 2025.

(3) Wang, S.; Chandra, A.; Liu, A.; Saligrama, V.; Gong, B. BabyVLM: Data-Efficient Pretraining of VLMs Inspired by Infant Learning; 2025.

(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, 5, 20–29. https://doi.org/10.1162/opmi_a_00039.

(5) 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.

(6) 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, my PhD mentor Shengao Wang for his mentorship throughout this project, and my family for their encouragement during my RISE internship at Boston University.