AI-based analysis of action and attention of child-parent interaction

Iba Baig 1,2 , Yanbin Xu 2, Kevin Li 3, Seiji Cattelain 3,4 , Hayato Ono 4, Sho Tsuji 3,4 , Mingbo Cai 2,4.

1 Northeastern University, 2 University of Miami, 3 Ecole normale supérieure, 4 University of Tokyo

Background

- Parent-child interactions--toy handling, spatial proximity, pose, gaze, language--are key in development.
- Manual annotation of observational videos is labor-intensive, time-consuming, and difficult to scale. Automation aids behavioral study (Weng et al. 2025).
- Vision Language Models (VLMs) offer opportunity to automate high-resolution video understanding and behavioral coding.

Research Questions

1. Can a multimodal Video-LM reliably annotate parent-child behavior at a fine-temporal scale?
2. What behavioral patterns, transitions, and temporal structures can be extracted automatically?

Methods

Variables: Controlled vocab & prompts

Our Architecture:

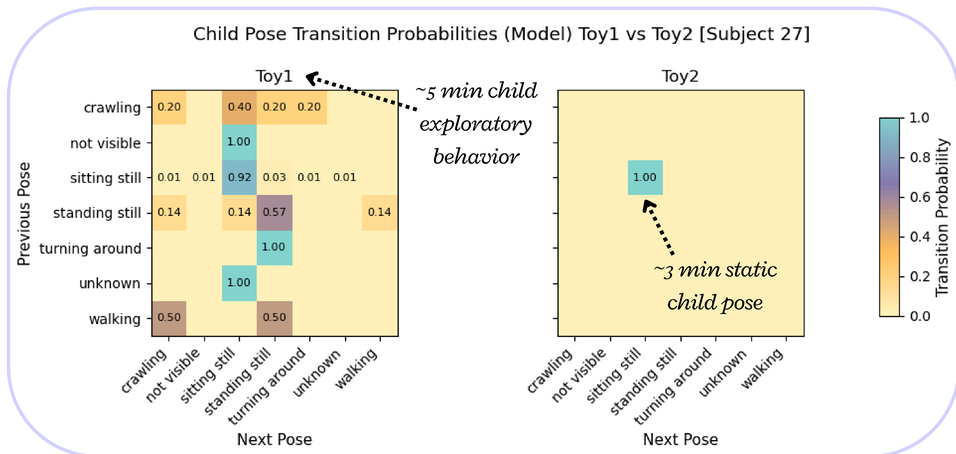
- Automated multi-turn conversational protocol to transform Tarsier outputs into a behavioral annotator

Tool: *Open-Source Pre-trained Video-Language Model* Tarsier2-7B

*Tarsier Model weights frozen

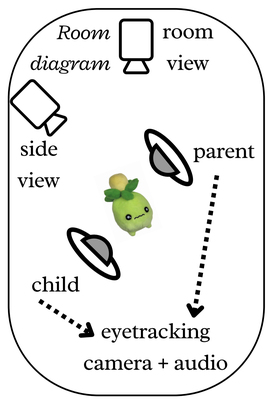
(Yuan et al. 2025)

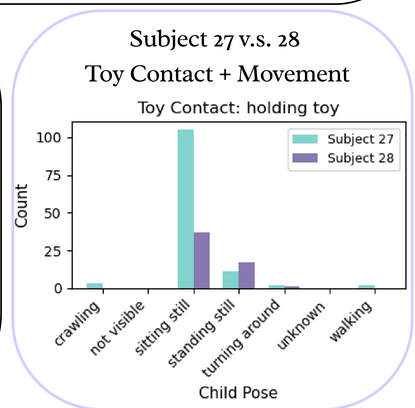
Q&A Prompt Choices Bank → Video Clip Preprocessing (3-sec clips) → Automated Q&A framework → Query Tarsier Response → Post-processing



Dataset

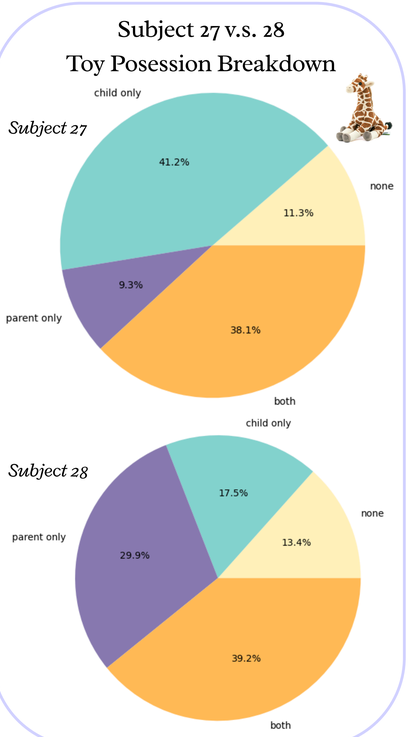
- Parent-child pairs recorded with four-camera videos
- Sample: 31 caregiver-child pairs (mean age=3.4 yrs)
- Parents instructed to teach child toy's name, hobby, food preference, and personality per toy

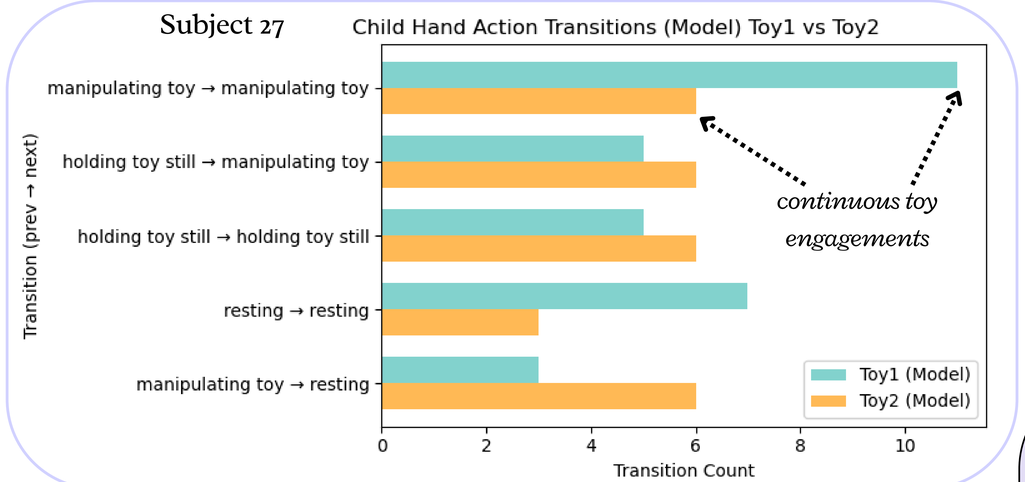




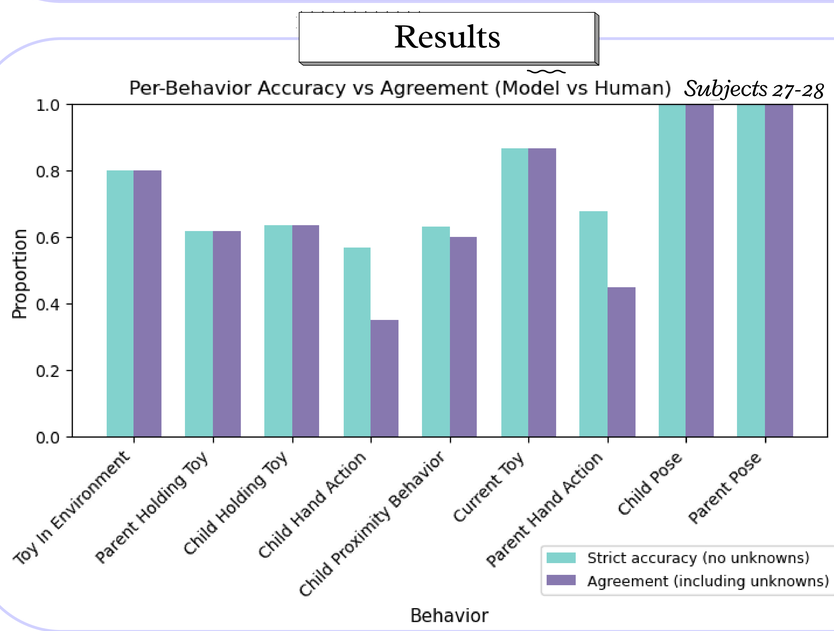
IRCN BABYLAB

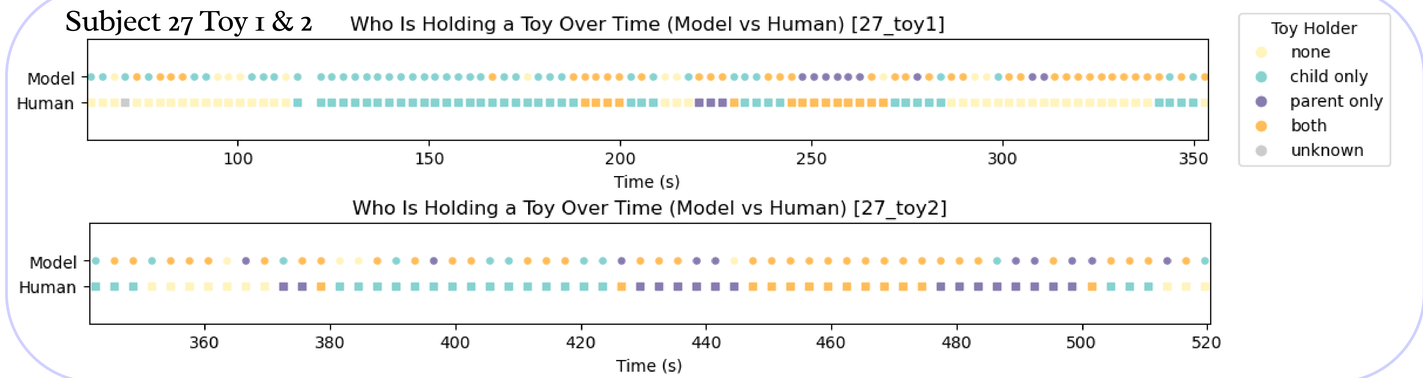
Eye-trackers





Stimuli





Conclusion

- Multimodal video-LLMs can automatically annotate parent-child interaction videos with reasonable agreement to human annotations
- Best performance for pose estimation

Limitations:

- Lower accuracy in hand actions/spatial proximity
- Pre-trained video-LM performance

Future Direction

1. Open source annotation tool!
2. Integrate multimodal analysis with collaborators (side/top + egocentric + eye-tracking + audio-visual)
3. Predictive modeling for individualized dynamics

SCAN ME

References

Weng et al. (2025). Artificial intelligence-powered 3D analysis of video-based caregiver-child interactions. *Sci Adv*

Yuan et al. (2025). Tarsier 2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. *arXiv:2501.07888*