Human Collaboration Dataset for Collaborative AI Assistants in the Real World

Dingxi Zhang[1], Peiyu Liu[1], Jingyuan Li[1] , Zhao Huang[1], Alexey Gavryushin[1*], Xi Wang[1*]

ETH zürich

CVG | Computer Vision and Geometry Lab

1 Introduction

Purpose: Create a large-scale dataset to capture daily collaborative interactions.

Device: Aria Glasses

Key Features:

- Simultaneous dual-viewpoint
- Multimodal data
- Fine-grained collaboration

Current Dataset Sample Statistics

- Task types**:** 6
- Task scenarios: 3
- Time length: ~110 min
- Total Participants: 5

2 Background

Current datasets like Epic-Kitchen-100, Assembly101, and Ego4D have advanced our understanding of task-oriented behaviors. These include cooking, toy assembly, and daily-life activities. However, they have limitations:

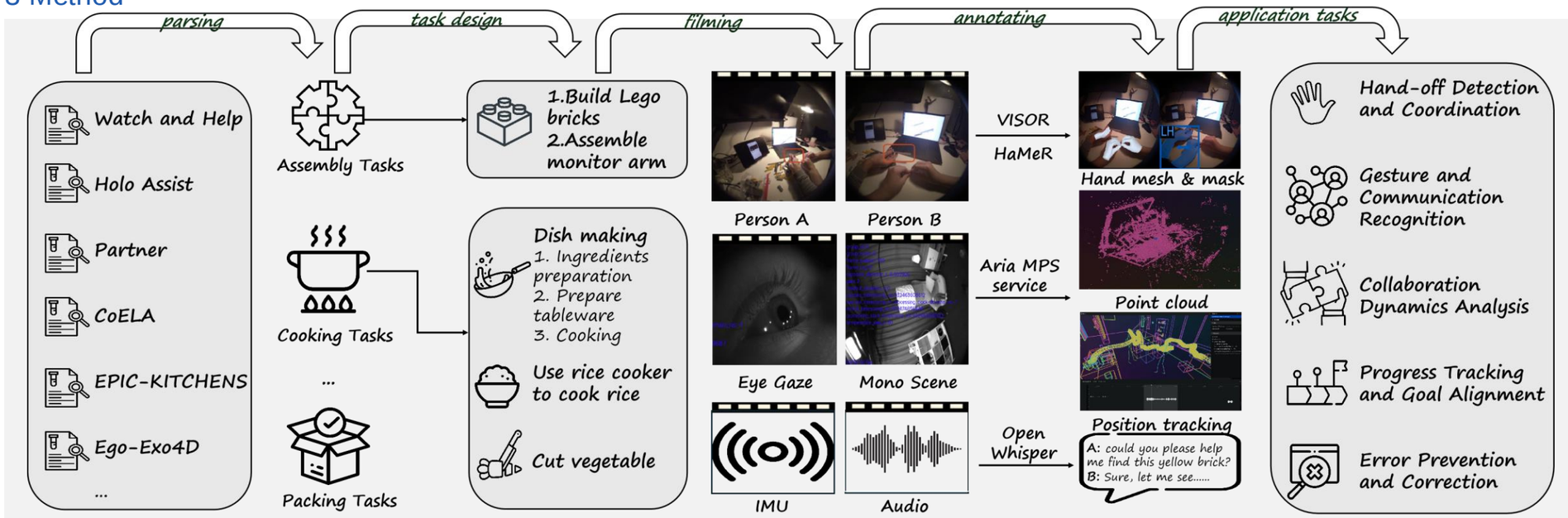
- No multi-agent collaboration contexts
- No real-time verbal communication
- Limited task diversity under shared visual feedback

Collaborative datasets like PARTNR and CoELA address some gaps but lack realistic scenarios.

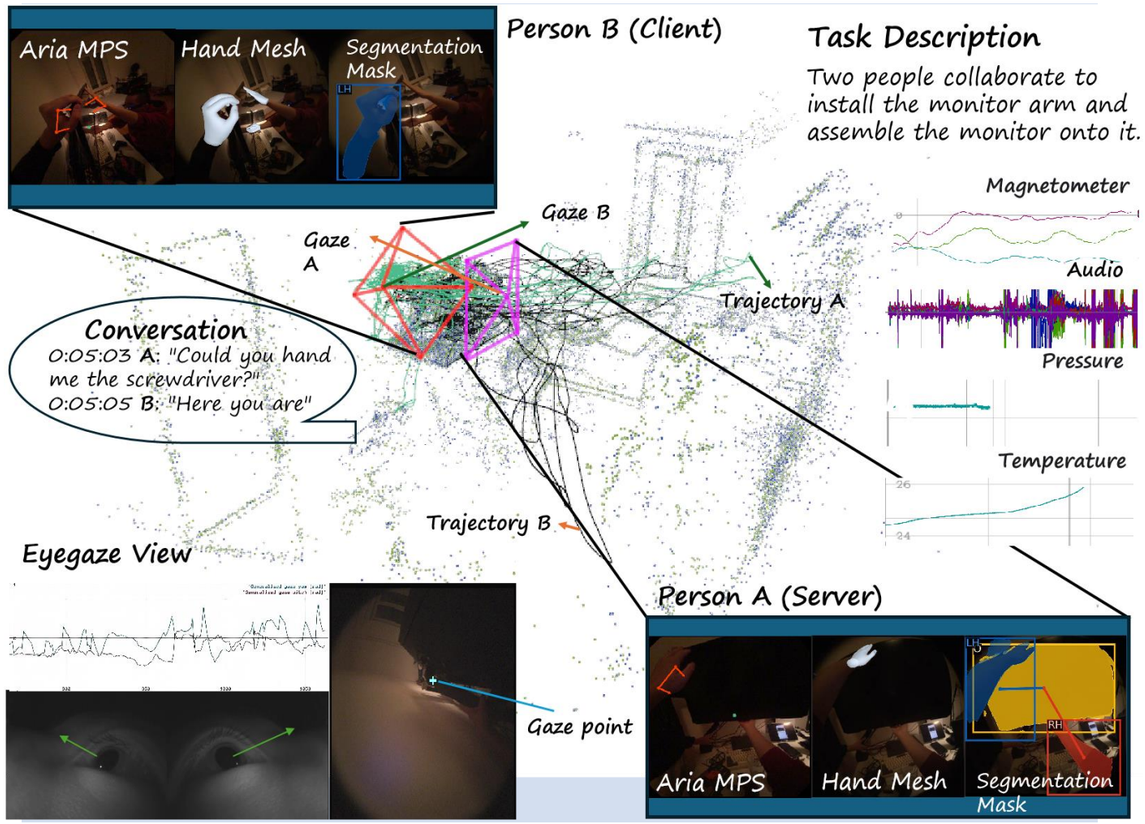
Dataset	Settings	Collaboration	Verbal Interaction	Physical Interaction	Realistic
Epic-Kitchen-100	Cooking	✗	✗	✗	✓
Assembly101	Toy assembly	✗	✓	✗	✓
Ego4D	Daily-life task	+	+	✗	✓
Ego-Exo4D	Daily-life task	+	+	✗	✓
HoloAssist	Assistive task	+	✓	✗	✓
VirtualHome	Household task	✗	✓	✗	✗
ALFRED	Daily-life task	✗	✓	✗	✗
WAH	Cooperative task	✓	✗	✗	✗
PARTNR	Cooperative task	✓	✓	✗	✗
CoELA	Cooperative task	✓	✓	✗	✗
Ours	Cooperative task	✓	✓	✓	✓

Table 1: Comparison of datasets for collaborative and instructional tasks.
Note: '+' indicates dataset partially support these attributes.

3 Method



4 Results



5 Discussion

Our dataset bridges gaps in collaboration studies, offering multi-modal recordings for diverse human-human interaction patterns, such as:

- Object handover
- Coordinated tool usage
- Error correction and recovery
- Real-time feedback exchanges
- Role-switching during collaboration
- Verbal negotiation

References

[1] Wang, Xin, et al. "Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[2] Chang, Matthew, et al. "PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks." *arXiv preprint arXiv:2411.00081* (2024).

[3] Pavlakos, Georgios, et al. "Reconstructing hands in 3d with transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[4] Zhang, Hongxin, et al. "Building cooperative embodied agents modularly with large language models." arXiv preprint arXiv:2307.02485 (2023).