Human Collaboration Dataset for Collaborative Multi-agents in the Real World

Dingxi Zhang, Peiyu Liu, Jingyuan Li, Zhao Huang, Alexey Gavryushin, Xi Wang
ETH Zürich
{zhangdi,peiliu,lijingy,zhahuang}@ethz.ch

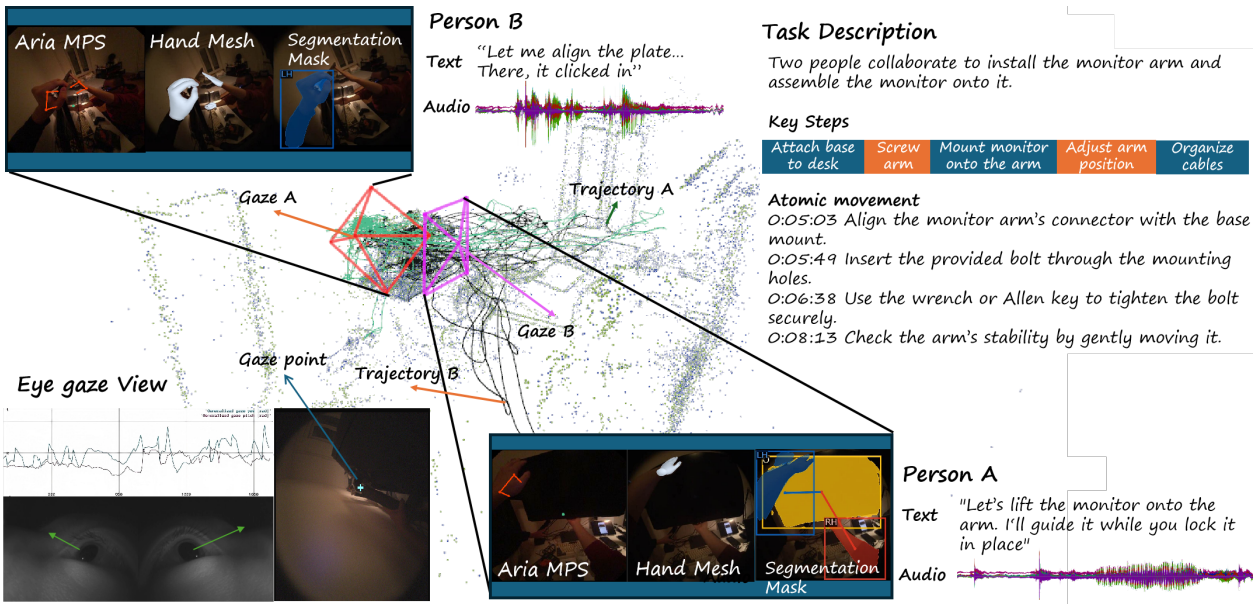

Figure 1. We introduce a novel egocentric dataset specifically tailored for studying multi-agent collaboration in everyday tasks. This dataset uniquely integrates time-synchronized multimodal sensory data—gaze, hand poses, audio, and spatial contexts—from the perspectives of multiple participants, facilitating an in-depth analysis of collaborative dynamics and many upstream tasks.

Abstract

Multi-agent collaboration is critical for achieving shared objectives in dynamic and complex environments. Existing datasets often fall short of capturing the naturalistic and unscripted interactions essential for studying real-world collaborative behaviors. To address this limitation, we present a novel egocentric dataset for analyzing daily multi-agent collaboration tasks, such as meal preparation and furniture assembly. Our dataset involves multimodal sensory data from the perspectives of multiple participants, enabling a comprehensive study of key dynamics under long-range collaboration. By bridging gaps in data diversity and realism, our work provides a robust foundation for advancing computational models of multi-agent collaboration, paving the way for deeper insights into human teamwork and its applications in robotics and artificial intelligence.

1. Introduction

Multi-agent collaboration [15] is a fundamental aspect of human behavior, underpinning activities ranging from daily tasks to complex industrial operations. Understanding how multiple agents—whether humans or robots—coordinate to achieve shared objectives is essential for advancing fields such as robotics, human-computer interaction, and artificial intelligence. Despite its significance, the computational modeling of multi-agent collaboration remains challenging due to the limited availability of high-quality datasets that capture the nuances of such interactions in real-world scenarios. Existing datasets either focus solely on egocentric views without collaboration [10,12,23], simulate collaboration in virtual environments [2,4,15] or only covers scripted short-range collaboration [3,8,20], limiting their applicability for studying naturalistic human interactions.

To address this gap, we propose a novel egocentric dataset designed specifically to study multi-agent collaboration in everyday tasks. Our dataset covers scenarios where

multiple humans collaborate together toward common goals that require multiple steps, such as preparing a meal, assembling furniture, or solving collaborative problems. Unlike traditional datasets that often emphasize individual actions or scripted interactions, our dataset captures unscripted and naturalistic collaboration scenarios, providing a rich and realistic foundation for studying multi-agent dynamics.

In summary, the key contributions made by our work are: (i) A dataset capturing naturalistic multi-agent collaboration in the real world, integrating multimodal sensory inputs from different perspectives. (ii) To the best of our knowledge, our dataset is the first to capture human collaboration interactions under unscripted and naturalistic goals in real world. (iii) Our dataset supports many novel and potential downstream tasks like gesture recognition, milestone recognition, interacted object identification, and so on.

2. Related Works

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	+	✓	✗	✓
EgoBody	Social interactions	+	✓	✓	✓
VirtualHome	Household task	✗	✓	✗	✗
ALFRED	Daily-life task	✗	✓	✗	✗
WAH	Cooperative task	✓	✗	✗	✗
PARTNR	Cooperative task	✓	✓	✗	✗
CoELA	Cooperative task	✓	✓	✗	✗
Ours	Cooperative task	✓	✓	✓	✓

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

Multi-agent collaboration, particularly in the context of human-robot and human-agent interactions, has been a significant research focus, with numerous studies exploring ways to enhance collaboration through various technologies, datasets, and methodologies [4, 9, 15, 16, 22]. These studies often investigate how agents can coordinate and collaborate toward shared goals, whether through verbal communication, physical interaction, or intelligent decision-making mechanisms.

2.1. Ego View Datasets

The study of human activity from an egocentric perspective has yielded valuable insights into individual behaviors [1, 5, 10, 12, 18, 21, 23], leveraging head-mounted devices to capture rich multimodal data, including gaze, hand poses, and environmental context. This wealth of information has significantly advanced our understanding of single-agent tasks and interactions. However, many of these works primarily focus on individual activities and fail to address the complexities of collaborative interactions.

While some studies have documented basic interaction motions, such as object passing or verbal instructions [11,21,23], these datasets often lack task-level context, limiting their applicability to understanding collaboration in complex scenarios. This gap highlights the need for datasets that capture more complex, real-world collaboration tasks, with multimodal data that reflects the intricacies of interaction and context.

2.2. Simulated Human-AI Collaboration

Moving beyond the individual perspective, numerous studies on simulation platforms have explored human-AI collaboration [4, 14, 15, 19, 22], examining how multiple agents—both human and robotic—coordinate and interact within shared environments. PARTNR [4] explores large-scale tasks and evaluation function generation using large language models (LLMs) in the Habitat 3.0 simulator [16]. CoELA [22] is a memory-augmented agent system that efficiently manages communication flows, promoting the exchange of essential information while minimizing unnecessary communication. In Watch-And-Help [15], the AI agent learns the underlying goal of the task by watching a single demonstration of the human-like agent performing the same task. However, collaborations in these works lack multi-modality cues from the real world and exclude cases in which agents help each other. Besides, these works focus on short-range collaboration with clear steps while collaborations in the real world are always long-range and without clear steps.

These collaboration-focused datasets offer invaluable insights into how agents coordinate and perform tasks, which is at the core of our work to improve multi-agent collaboration in real-world. Noticing the gap between current datasets and real-life collaborations, we aim to build models that capture the complex dynamics of collaboration, helping bridge the gap between theoretical research and practical implementations in multi-agent systems. In Table 1, we provide a comparison of these related works alongside our own, highlighting the differences and similarities between them to better contextualize our contributions within the broader landscape of multi-agent collaboration research.

3. Method

In this section, we introduce our dataset collection pipeline aimed at training and evaluating AI agents to collaborate daily life tasks in natural language context with humans (See Fig. 2). Specifically, we record over **110 min** of human collaboration data involving 3 scenarios and 5 participants. Our dataset covers three types of tasks: (1) Assembly Tasks. (2) Cooking Tasks. (3) Packing Tasks. Using sensors provided by Meta Aria glasses [7], our dataset offers multiple modality labels such as audio, transcripts,

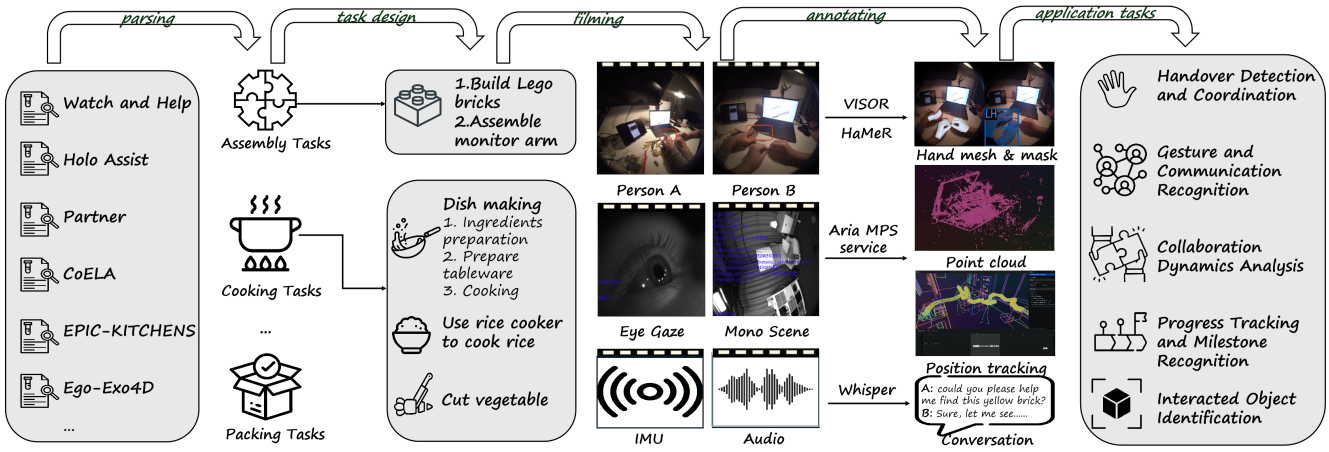

Figure 2. **Dataset Collection Pipeline**. We start by parsing existing collaboration action datasets and designing our own task scenarios. Each scenario is subsequently divided into distinct steps. After filming, the data undergoes synchronization and post-processing to yield synchronized multimodal data. This dataset is highly adaptable and can support a wide range of downstream analyses and tasks related to collaboration.

point clouds, trajectory recordings, etc., enabling downstream tasks.

3.1. Task Design

To distinguish our dataset with other human collaboration dataset [4, 15, 21, 22], our recording happens in the real word with human-human body contact. Aside from the short and clear instructions provided by other datasets, we provide long-term goals without specific instructions. We declare that this setting resembles collaborations in our daily life where the robot learns to reason and help given long-term goals.

Motivated by [4], we divide long-term collaboration tasks into two types: (1) Expert and Helper, where some subtasks can only be completed by one agent. (2) Cooperative Pairing, where subtasks can be completed in any manner by either agent. Based on the division, we design three long-term tasks containing human body contact to cover different senarios. Specifically, we design cooking tasks to represent expert and helper collaboration, in which two participants tried to finish one dish with most of the process finished by chef. And we also design assembly tasks for two collaborators tasks, in which participants need to find suitable pieces from a chaotic pile of Lego bricks and assemble them. We also record packing task where two participants need to collect items according to a provided checklist and pack these items into a suitcase. In assembly and packing task, both participants share the same pre-knowledge and could perform any subtasks.

3.2. Recording Hardware

The data was collected using Meta's Aria glasses [7], equipped with multimodal sensors. Specifically, we utilized

Figure 3. The Aria device used for egocentric recordings [11].

the following: (1) a single fisheye RGB camera with a horizontal field of view (HFOV) of 110° and a vertical field of view (VFOV) of 110°, capturing images at a downsampled resolution of 1408×1408 pixels with a nominal frame rate of 10 FPS; (2) a grayscale mono scene camera with a 150° HFOV and 120° VFOV, operating at a resolution of 640×480 pixels and a nominal frame rate of 10 FPS, which is primarily used for scene and SLAM data; (3) a seven-channel spatial microphone array, sampled at 48 kHz and used in stereo mode for directional audio capture; and (4) an eye-tracking camera recording at a resolution of 320×240 pixels with a frame rate of 10 FPS for gaze tracking. The device integrates a rich sensor suite that is tightly calibrated, capturing a broad range of modal ities as shown in Figure 4.

Since we have at least two participants for each scenarios, all sensor measurements are synchronized on a common nanosecond-resolution clock, ensuring temporal alignment across modalities. This setup enables precise multimodal data collection for tasks such as SLAM-based localization, gaze interaction analysis, and audio-visual correlation studies.

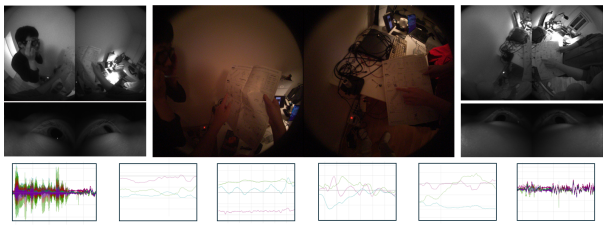

Figure 4. Sensor streams in two perspective. Center Top: RGB camera. Right and left side: IMU camera and eyetrack camera. Bottom: microphones, magnetometer, accelerometer respectively from side to center

3.3. Annotations

To enable comprehensive downstream analyses, we meticulously annotated the multimodal data collected using Meta Aria glasses. This section details the modalities captured, the methods applied for post-processing, and the derived annotations, ensuring a robust dataset for diverse research applications.

Captured Modalities. Our study leverages the multimodal capabilities of Meta Aria glasses to capture a diverse range of sensor data for each participant. The recorded modalities include high-resolution RGB video from the primary camera, stereo audio captured by two spatial microphones, gaze estimation streams from two eye-tracking cameras, motion data from two IMUs, and scene data from two SLAM cameras.

Hand Tracking and Reconstruction. The wrist and palm poses were computed using images captured by the SLAM cameras, with positions expressed in meters relative to the device's frame of reference. This computation leveraged the Machine Perception Services (MPS) provided by Project Aria [7]. To enhance the quality and detail of the hand tracking data, we employed the HaMeR [13] and the VISOR [6]. These tools facilitated the generation of reconstructed hand meshes and precise hand-object segmentation, providing a robust foundation for downstream tasks involving hand motion analysis and interaction modeling.

Gaze Estimation. Gaze estimation was also carried out using the MPS. Data from the Eye Tracking (ET) cameras enabled the computation of precise gaze directions for both the left and right eyes, as well as the depth at which these gaze vectors intersected. This multimodal gaze information serves as a critical component for understanding user attention and facilitating advanced interaction techniques.

Spatial Localization and Reconstruction. Leveraging data from the SLAM cameras and IMU sensors, we utilized MPS to compute a 6-DoF trajectory and a semi-dense point cloud for individual recordings. For multi-session alignment, the Multi-SLAM MPS was employed, enabling the generation of SLAM outputs in a shared coordinate frame across multiple recordings. This integration facilitates precise spatial localization and seamless scene reconstruction, supporting advanced downstream applications.

Conversation Annotations. To capture the conversational dynamics during collaboration tasks, we leveraged Whisper [17], an advanced speech recognition framework. Whisper was employed to transcribe spoken interactions, enabling detailed linguistic annotations that can support tasks such as dialogue analysis, context inference, and collaborative behavior modeling.

Action Annotations. To enable detailed downstream analyses of each participant's activities, we created both coarse-grained and fine-grained action annotations for the recorded sessions. Coarse-grained actions correspond to high-level task steps (e.g., "Use the rice cooker to cook rice"), which can be decomposed into a series of fine-grained actions. These fine-grained actions represent low-level, atomic movements (e.g. pressing, screwing, cutting) that are integral to completing a task step and typically last 1-2 seconds. This dual-layer annotation approach allows for a comprehensive understanding of task execution at both the broader and more granular levels.

This multimodal dataset forms the foundation for downstream analysis, integrating visual, auditory, and motion data streams to facilitate comprehensive studies in various domains.

3.4. Data Collection Protocol

Before initiating the recording, participants must ensure that the environment is well-lit, with natural lighting preferred to minimize glare and shadows, and that it is free from significant background noise to maintain audio quality. The recording setup should be carefully arranged to avoid obstructions in the field of view of the cameras. Each recording session shall include at least two participants, both of whom must wear Aria glasses continuously throughout the session. Participants should engage in natural interactions and movements relevant to the study objectives, ensuring a diverse range of activities is captured. The duration of each recording session must be sufficient to gather comprehensive data, with a recommended minimum of 10 minutes. Additionally, all equipment must be checked for proper functioning prior to the session, and calibration procedures, if required, should be completed to ensure data accuracy.

4. Result

Given the time constraints of the entire project, we have not conducted large-scale recordings up to this point. Instead, we have primarily focused on sample recordings for the three types of tasks selected earlier. As shown in Table 2, the diversity of these three types of tasks is already sufficiently rich in terms of both object types and activity

Object Scales	Object Categories
Small	LEGO Bricks
Medium	Cookers, Grocery Items
Big	Monitor, Monitor Arm
Activity Scope	**Activity Categories**
Seated Tabletop	LEGO Assembling
Standing Tabletop	Cooking
Indoor Movement-based	Monitor Arm Assembling, Baggage Packing

Table 2. Object Scales and Activity Scope. We classified objects by scale and corresponded activity scopes with categories.

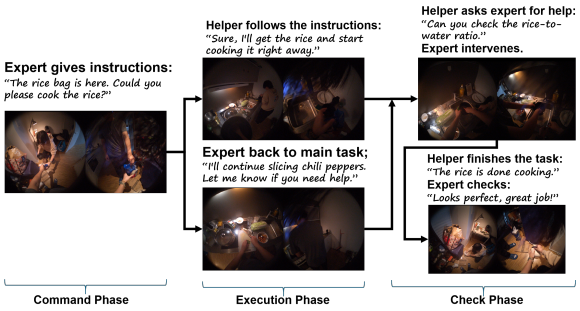

Figure 5. The Expert and Helper collaborate patterns.

scope. Based on these samples, we have identified key time points of collaboration, which we have used for discussion and analysis. This approach serves to demonstrate the value and necessity of expanding and completing the dataset in the next phase of the project.

4.1. Collaboration Analysis

1. Person A:	"I'll start by assembling the base."
2. Person B:	"Got it! I'll organize pieces by color. Do you need the red pieces too?"
3. Person A:	"Yes, thank you. I'll need them in a minute."
4. Person B:	"Alright. Now, I'll build the walls. You can get red pieces here."
5. Person A:	"Got it. Then I will start on the roof and attach the tiles."
6. Person B:	"Okay. After finishing your part we can start to build up the house."
7. Person A:	"Good. Just make sure all pieces fit together snugly. You're doing great."
8. Person B:	"I see. Where do the small pieces go?"
9. Person A:	"Here, now I am done. So we can attach them to the roof."

Table 3. Lego Building Dialogue (Person A & Person B). Dialogue illustrates the collaborative interaction between two people assembling a Lego set. Underline texts are atomic tasks for each person.

Expert and Helper Task. Due to the unscripted design, the cooking task—compared to the assembling task with higher ability threshold—places a stronger emphasis on the expert-and-helper task setting. In this scenario, the helper is assumed to possess limited pre-knowledge related to the task goal and is expected to respond promptly to the instructions provided by the expert. This setup leads to frequent interactions like in Fig. 5, where the expert issues specific commands during the task, the helper executes the instructions by applying their limited pre-knowledge, and upon completion, notifies the expert, who then verifies the execution of the task. Throughout this process, extensive dialogue occurs between the expert and the helper to synchronize task progress and adjust task objectives as needed. This continuous communication helps improve the efficiency and accuracy of achieving the final task goal.

Cooperative Pairing Task. In the Cooperative Pairing Task, where subtasks can be completed in any manner by either agent, collaboration patterns are characterized by a dynamic interplay of synchronization and task delegation. Using Lego building as a representation of these tasks, the collaboration typically begins with both agents synchronizing their actions and agreeing on initial task assignments (See Tab. 3). As the task progresses, one person completes their assigned subtask and notifies the other, ensuring that both agents are aware of each other's progress. This continuous exchange of status updates enables the agents to re-synchronize and adjust their actions accordingly, maintaining a seamless flow of collaboration. Such coordination ensures that both individuals are always aligned and working toward the same goal, enhancing efficiency and minimizing idle time. The key collaboration patterns thus revolve around synchronization, task updates, and adaptive re-coordination as each person contributes to the overall task completion.

4.2. Downstream Task

Handover Detection and Coordination. As shown in Fig. 6, our recorded tasks contain a substantial number of handover motions involving object passing. Handover detection and coordination are crucial for understanding the flow of collaboration, as they highlight moments of interaction and coordination between participants. Accurately identifying handovers provides insights into communication, role allocation, and joint action timing, which are key for modeling effective collaborative behaviors in both human-human and human-robot interactions.

Gesture and Communication Recognition. Our dataset, which includes multiview recordings of interactions

between multiple participants and multimodal data such as eye gaze and hand poses, provides a strong foundation for detecting and identifying gestures used in communication. Leveraging this rich data, we can capture a wide range of gestures, enabling a deeper understanding of non-verbal communication and how gestures contribute to collaborative tasks and interactions.

Progress Tracking and Milestone Recognition. Progress tracking monitors the sequence of actions during a task, while milestone recognition identifies key moments, such as task completion or critical transitions. This is important for assessing collaboration efficiency. Using both coarse and fine-grained action annotations, we track the task's progress by linking actions to their corresponding steps. We combine this with multimodal data, such as hand movements, gaze, and dialogue, to detect and recognize milestones in real-time, offering valuable insights into the collaborative process.

Interacted Object Identification. This task involves detecting when and which objects are interacted with during a collaborative task. Given the complexity of the interactions in our dataset, including frequent object-passing motions, identifying the specific objects involved is crucial for understanding task dynamics. To achieve this, we combine multimodal data—such as hand movements, gaze, and spatial context—with the global task framework. By analyzing these data points, we can accurately track which objects are being passed, manipulated, or used by participants, providing valuable insights into collaborative behaviors and object interactions.

4.3. User Study

Given the research-oriented focus of our project, we conducted a user study to gather feedback and insights from academia community. The overall response was largely positive, with many participants expressing strong confidence in the value of our dataset, especially for applications in both robotics and computer vision area.

Many participants recognized the value of our dataset, particularly its multimodal data and diverse annotations, for analyzing collaboration dynamics and supporting a wide range of downstream tasks. However, suggestions for improvement included using Aria glasses' microphone arrays to measure participant distance and enhancing spatial synchronization for more accurate interaction tracking. Additionally, incorporating more task-related environmental context was recommended. These insights will help refine and expand the dataset in future work, improving its accuracy and applicability.

5. Discussion

Our semi-dense point cloud, generated using Aria Machine Perception Services, has shown promising results but also presents a few challenges. While multi-view point cloud reconstruction boosts robustness, the dynamic data employed in constructing the semi-dense cloud can inadvertently introduce aliasing effects. To address this, we plan to integrate static radar scanning techniquesn. By capturing an initial, stable representation of the task environment, we aim to enhance the quality of scene scanning and minimize the impact of dynamic interference. Additionally, we recognize the importance of optimal lighting conditions for improving data accuracy and plan to adjust the lighting setup during recording to further refine the quality of the captured data. These refinements are expected to significantly elevate the reliability and precision of our task execution models.

6. Conclusion

In this work, we presented a novel dataset designed to explore the complexities of multi-agent collaboration in everyday tasks. By capturing synchronized multimodal data from various perspectives, including hand poses, gaze, audio, action annotations and spatial context, our dataset offers a comprehensive foundation for many upsteam tasks. Unlike previous datasets that focus on individual behaviors or scripted/simulated scenarios, our dataset captures dynamic, unscripted collaborations across various tasks in the real world. Moving forward, we aim to refine and expand the dataset based on feedback, improving its robustness and applicability for a broader range of research challenges. By bridging the gap between virtual simulations and real-world collaboration, our dataset paves the way for improved human-robot interaction and more sophisticated collaborative AI systems.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. *arXiv preprint arXiv:2411.19167*, 2024. 2

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 1

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[4] Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, et al. Partnr: A benchmark for planning and

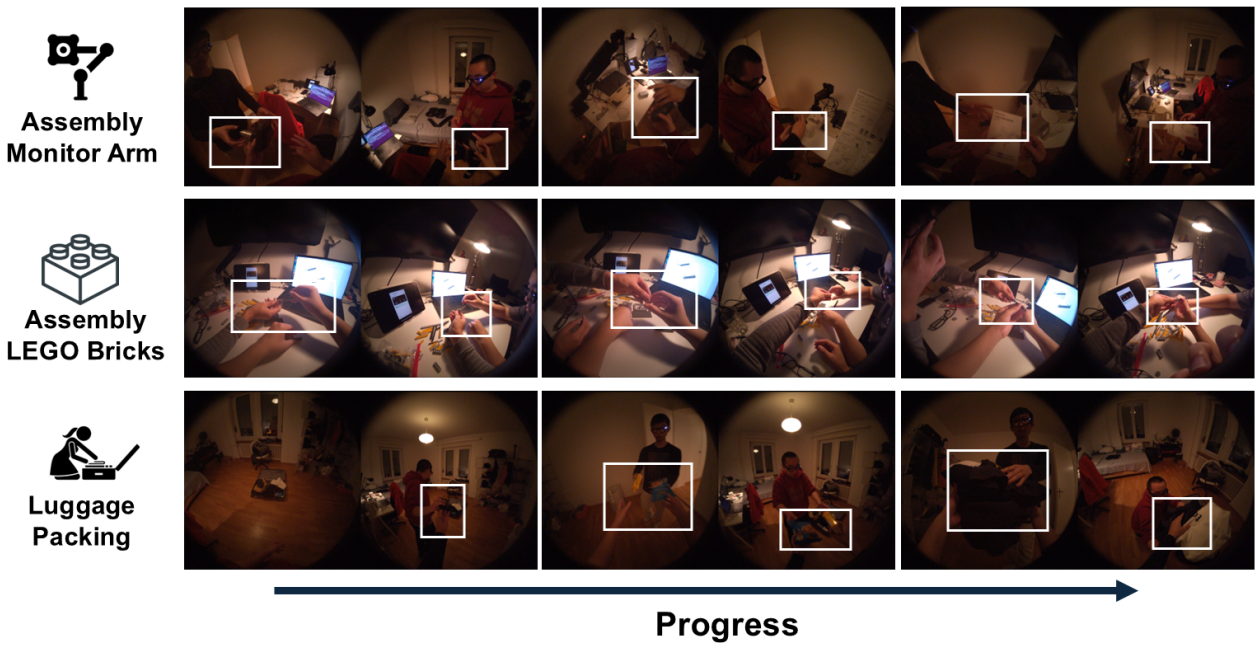

Figure 6. Our dataset captures various daily life collaboration tasks. From left to right, it sequentially presents the process of two individuals collaboratively completing different tasks from two synchronized egocentric perspectives. The white frames highlight the handover moments.

reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*, 2024. 1, 2, 3

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[6] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 4

[7] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 3, 4

[8] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 1

[9] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020. 2

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2

[11] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 3

[12] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2

[13] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 4

[14] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs, 2018. 2

[15] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020. 1, 2, 3

[16] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min,

et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023. 2

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4

[18] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2

[19] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 2

[20] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 1

[21] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023. 2, 3

[22] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023. 2, 3

[23] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 1, 2