This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Kaiwu: A Multimodal Manipulation Dataset and Framework for Robot Learning and Human-Robot Interaction

Shuo Jiang, Haonan Li, Ruochen Ren, Yanmin Zhou, Zhipeng Wang*, Bin He*

***Abstract*—Cutting-edge robot learning techniques including foundation models and imitation learning from humans all pose huge demands on large-scale and high-quality datasets which constitute one of the bottleneck in the general intelligent robot fields. This paper presents the Kaiwu multimodal dataset to address the missing real-world synchronized multimodal data problems in the sophisticated assembling scenario, especially with dynamics information and its fine-grained labelling. The dataset first provides an integration of human, environment and robot data collection framework with 20 subjects and 30 interaction objects resulting in totally 11,664 instances of integrated actions. For each of the demonstration, hand motions, operation pressures, sounds of the assembling process, multi-view videos, high-precision motion capture information, eye gaze with first-person videos, electromyography signals are all recorded. Fine-grained multi-level annotation based on absolute timestamp, and semantic segmentation labelling are performed. Kaiwu dataset aims to facilitate robot learning, dexterous manipulation, human intention investigation and human-robot collaboration research. The dataset is accessible via https://doi.org/10.57760/sciencedb.14937.**

***Index Terms*—Robot learning, Embodied AI, Human-Robot Interaction, Multimodal Fusion**

I. Introduction

Robots represent a key platform for artificial intelligence (AI) and are often considered the primary carrier of Embodied AI [1]. With the shift in demand from cyberspace to physical environments, traditional industrial robots are inadequate to adapt to changing environments. There is a growing need for intelligent robots capable of interacting with humans in uncluttered environment [2] and master human-level skills. However, the path to attaining human-level skills remains challenging and is still an ongoing research problem.

Datasets are the fundamental part of robots and embodied AI, and cutting-edge methods, including imitation learning or foundation models, all pose significant and growing demands on the high-quality, large-scale and multi-modal datasets [3]. Imitation learning is primarily achieved through human demonstration data, and most of the methods are based on vision data [4] or combined with depth data. New challenges on the data arise from changes in target skills, shifting from simple, fixed motions required for specific tasks, such as picking up an object, to more complex motions. In addition, demonstration via teleoperation represents another alternative approach. For example, Finn et al. proposed mobile ALOHA [5] which used whole-body teleoperation data collection methods. Tesla proposed to use real-human motion capture data to train the Optimus humanoid robots [6]. To be noteworthy, with the emergence of foundation models [7] including ChatGPT, the importance of data volume and quality in determining the algorithm's upper limits has become more widely recognized and there are a few pioneer works exploring the real-world datasets towards the foundation models for robotics.

RT Series [8], [9], [10] focus on long-term task samples that include multiple activity instances, emphasizing autonomous environmental perception and logical reasoning ability. The emerging challenges for foundational models now entail a shift from managing relatively short-term tasks to attaining long-term autonomy. Philosophically, long-term autonomy requires interaction with the environment and its context, because the brain-inspired algorithms alone are insufficient. Consequently, embodied AI addresses the interplay between humans, robots, and their environment.

Nevertheless, current datasets all suffer from the following critical limitations. First, most of the data mainly relies on the videos and thus lacks dynamics information. Computer vision methods in the form of images or videos, although advancing and widely deployed, can only represent kinematics information including trajectory and velocity. However, robot learning is a process of dynamics with the surroundings. Missing dynamics information including force will deteriorates the learning performance, resulting in the superficial learning. Second, there is a lack of a universal, sophisticated, intuitive human-level perception framework. The working environment for intelligent robots changes from fixed and clustered settings including manufacturing scenarios to open and complex settings including in situ homes. The environment not only involves static and dynamic objects but also includes human beings. Previous datasets have only incorporated certain sensing modalities, such as video, Inertial Measurement Unit (IMU), etc. which are inadequate to address the demands of increasingly complex tasks. Furthermore, one reason that the

Manuscript received: May 22, 2025; Revised: July 9, 2025; Accepted: September 2, 2025. This paper was recommended for publication by Editor Tetsuya Ogata upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Key R&D Program of China (Grant No. 2024YFB4707600); in part by National Natural Science Foundation of China, under grants W2441018, 62088101; in part by Shanghai Municipal Science and Technology 24511104400 and Major Project (2021SHZDZX0100).(*Corresponding Author: Zhipeng Wang and Bin He.)

All authors are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, State Key Lab of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai 200120, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (Email: hebin@tongji.edu.cn).

Digital Object Identifier (DOI): see top of this page.

©2026 IEEE

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

manipulation skills of robots cannot surpass those of humans is the lack of fundamental understanding of neural mechanisms. Additionally, relying solely on video as a modality is insufficient to fully reveal how humans accomplish such complex tasks.

To address the above challenges, this paper presents a multimodal framework and dataset for robot learning and evaluation. The key differences from previous datasets include the integration of multimodal data encompassing humans, robots, and the environment. This dataset not only supports robot learning from human demonstrations but also facilitates the prediction of human intentions to enable more effective policy development. We aim to develop skills sets for robot learning and the name Kaiwu is inspired by the ancient Chinese literature *Tiangong Kaiwu* (also known as *The Exploitation of the Works of Nature*), a comprehensive encyclopedia encompassing a wide range of fields, including agriculture and craftsmanship. The contribution of this paper is as follows:

(1) A multimodal data collection framework is proposed, featuring full situational awareness including manipulation dynamics information, human manipulation neural signals and attention information and multi-view manipulation vision information. This framework aims at complex scenarios including human assembly, towards the universal manipulation ability for embodied AI.

(2) High-quality, large-scale multimodal data for long-horizon autonomy are collected, utilizing state-of-the-art ground truth techniques for fine-grained and complex manipulation process, holding potential for future benchmark.

(3) The dataset is accompanied by an annotation of spatio-temporal relationships. Rich and fine-grained cross-modal synchronization data annotation are performed, including 298 annotations of regions of interest as personal attention data, 536,467 annotations of closed area elements for image segmentation, 7,197 motion segmentation events for left and right hand dexterity manipulation, 4,467 gesture event annotations, and 4,959 annotations for gesture classification, which significantly enhanced the cross-modal learning and multimodal fusion capabilities and its interpretability.

The rest of the paper is summarized as follows. In Section 2-4, research on related datasets, the experimental setup, process design, and the formulation of the target problems are first introduced. In Section 5, a detail introduction to the annotation of Kaiwu dataset is presented. Then, the descriptive statistics and directory structure of the dataset are presented, discussed, and introduced by category in Section 6. API, project official website and repository, known issues, proposed improvements, and future research directions are discussed in 7-8 sections. Finally, the conclusion is drawn in Section 9.

II. Related work

The main focus of this paper is on how to build a multimodal dataset for robot learning via human demonstration for human-level skills and future human-robot collaboration, and thus the related key technology are first reviewed.

A. Robot learning dataset

Multi-modal input can serve as a crucial foundation for robots to learn human behavior, enhancing the accuracy and efficiency of the learning process. The dataset focuses on teaching robots to perform daily interactions in changing environments through human demonstrations [7].

JD ManiData, ManiWAV [11], and RH20T [12] help robots acquire diverse and generalizable skills. The DROID dataset [13] uses a distributed approach to enhance robot performance and generalization. GraspNet-1Billion [14] improves object grasping in cluttered scenes with multi-modal data. The Open-X implementation [15] enables large-scale data integration and validates data transfer between robots to improve multi-robot capabilities.

However, the above datasets suffer from data homogenization. The ability to efficiently and purposefully train embodied intelligence models in complex scenarios is limited. Building on the excellent datasets above, more research has been done to further extend on a particular aspect.

The RT series datasets evolve to enhance robot capabilities. RT-1 [8] uses large-scale data and high-capacity models for multitask control in real-world environments, incorporating text, vision, and action data. However, it has limitations in scene and target object selection. RT-2 [9] builds on RT-1 using Web VQA to link images with text tokens, addressing these limitations. The Open X-Embodiment dataset [15] integrates RT-2 with more diverse data to train the versatile RT-X model. The latest RT-H adds human-robot interaction data to further improve learning capabilities.

The ARIO dataset [16], based on Open X-Embodiment, includes simulation and real-world data with diverse robot hardware and a uniform format. However, it lacks dynamic kinematic information and authenticity in raw data, and faces labeling generality issues.

B. Human activity recognition dataset

Enabling robots to recognize and understand human activity is crucial for embodied intelligence. High-quality datasets are essential. The industry-oriented dataset [17] captures human movement in industrial settings for classification and prediction. ActionSense [18] focuses on kitchen scenes with multi-modal data. HUMBI [19] provides a multiview camera dataset for high-resolution human body modeling, enhancing reconstruction and recognition capabilities. The Toyota Smarthome Untrimmed (TSU) dataset [20] collects untrimmed home videos to help robots understand the causality of human activity. Another data set [21] aims to help robots understand the intention of the team to better assist humans.

The current research lacks an extension of the application scenarios for the dataset and a focus on assembly task-related scenarios. In addition, there is a lack of balance in data modality homogeneity, cross-modality temporal consistency, and task causality analysis.

C. Human-Robot Interaction dataset

Embodied intelligence focuses on how intelligent systems interact with their physical bodies and environments. Human-

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

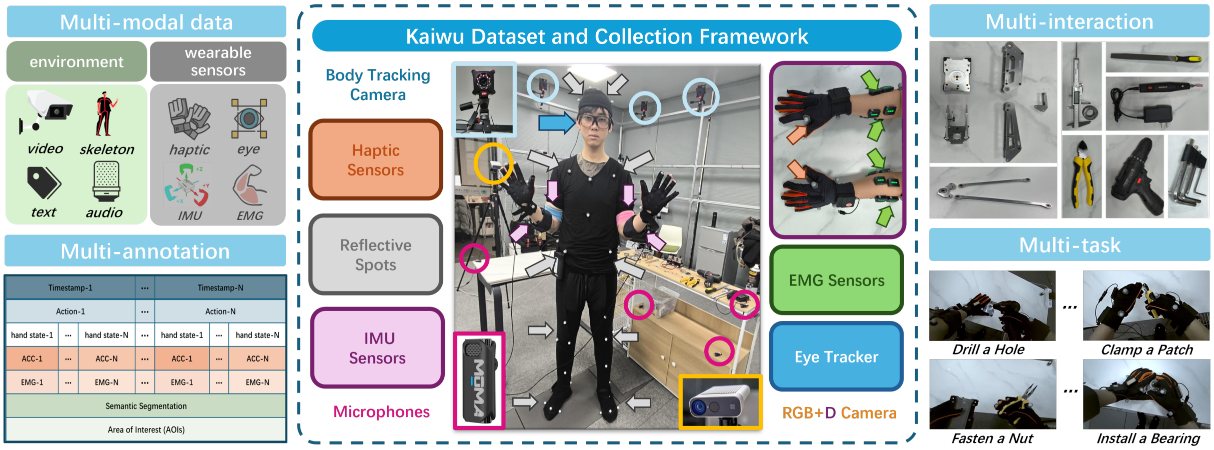

Fig. 1. A framework of wearable and environment-mounted sensors records rich activity information in the assembly environment.

robot interaction helps agents adapt to surroundings by learning from human interactions, improving the generalizability and stability of their behavior. Multimodal datasets, which include text, images, sound, and video, provide rich information about user intentions and emotions. In human-robot interaction, these datasets enhance agents' understanding of user needs, improve accessibility and perception, and enable more natural and intelligent interactions.

The dataset [22] teaches robots daily interactions in changing environments via human demonstrations. The RBO dataset [23] records human interactions with articulated objects. The RT-H [10] method enhances RT-2 by adding human-robot interaction interventions, using fine-grained phrases to describe motions. This improves the robot's understanding of interaction actions and language, builds an action hierarchy, and enables learning from human language interventions.

Eye-tracking devices can map the region of interest along the viewing trajectory, helping robots understand human intentions during tasks like assembly. EMG and IMU signals capture muscle behavior patterns during actions, enabling robots to learn and predict hand gestures and respond more accurately to human behavior.

The HARMONIC dataset [24] captures human-robot interactions in an autonomous setting, using a joystick-controlled robot and wearable sensors to gather data rich in human intent. The HBOD dataset [25] employs more wearable sensors to capture detailed human movements and interactions with tools, enhancing robot understanding of human intentions and maneuvers. The dataset [26] explores learning challenges in hand-object interactions, including intention recognition and motion generation. The OAKINK2 dataset [27] provides multi-view images and accurate pose annotations of humans and objects, supporting applications like interaction reconstruction and motion synthesis.

Regarding some shortcomings of the above datasets, the Kaiwu dataset directly collects dynamic and static data using wearable sensors. It integrates assembly actions into a coherent process, enhancing human-robot interaction with a narrative and causal structure. Additionally, a coordinate framework is established to enrich spatial integrity. A detailed comparison with other similar datasets is provided in Table I.

III. Data Collection Platform Setups

The data collection platform enables synchronized streaming, storage, and visualization of information from wearable and environment-mounted sensors, accommodating different sampling rates and data formats. Its modular structure simplifies the integration of new sensors. Multi-threading and multi-processing ensure efficient CPU and RAM usage. The platform also supports post-processing, allowing integration with third-party applications. It aims to enhance data collection efficiency by reducing experimental artifacts, enabling easy metadata recording, and allowing researchers to focus more on the subject build process.

A. Application and use cases

The main purpose of this dataset is to provide rich information on how human can achieve dexterous operation and how these information can help robot acquire human-level intelligence. Therefore, we identify several critical processes of human dexterity including dynamics information during the manipulation, attention during the cognition, understanding of muscle group mechanism and assembly logic, ubiquitous environment information along with the operation process and also ground truth recording. The overview diagram is shown in Fig. 1.

B. Sensor setups

Targeted at the above requirements, cutting-edge sensing acquisition equipment are utilized. These devices include wearable sensors and environment-mounted sensors (Fig. 1).

The data acquisition devices used in our study are commercially available and detailed in Table II, including exact product models and manufacturers. This information facilitates the replication of our experimental setup by other researchers, promoting transparency and scientific rigor.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
A COMPARISON OF RELATED DATASETS. MOTION CAPTURE CONTAINS 3D SKELETON AND GROUND TRUTH.

Dataset	Modalities	Application Tasks	Traj.	Lang. Instruct	Cam. Calibration	Verbs.	Collection
TSU	RGB,Depth,3D Skeleton	Daily Actions	√	√	×	√	Teleoperation-based
Harmonic	Gaze,EMG,RGB,Depth	Meal	√	×	√	×	Teleoperation-based
Hbod	3D Skeleton,Tactile,Hand Pose,IMUs	Tool Operation	×	√	×	√	Direct
OXE	Mainly RGB,Depth	Multiple Scenarios	√	√	×	√	Dataset Aggregation
Action sense Dataset	IMUs,3D Skeleton,Hand Pose,Gaze EMG,Tactile,RGB,Depth,Audio	Kitchen Activities	√	√	√	√	Direct
Kaiwu Dataset	IMUs,Motion Capture,Hand Pose,Arm,Gaze EMG,Tactile,RGB,Depth,Audio	Industrial Assembly	√	√	√	√	Direct

TABLE II
DEVICE DETAILS.

Sensors	Product Name
Data Glove	WISEGLOVE 19FE+
RGB-D Sensors	Kinect
Eye Tracker	Tobii Pro Glasses 3
Microphones	MOMA Lark m1
Motion-capture Sensors	Nokov XINGYING
EMG and ACC Sensors	DELSYS Trigno Wireless Biofeedback System

1) Data glove: Data gloves are utilized to simultaneous capture the hand movement and palm tactile interaction information. Therefore, the glove (Fig. 2) is chosen which can drive 3D animated human hand movements in real time. 19 finger angle sensors and 19 finger pressure sensors are equipped with a sensing accuracy of 9g, which fulfills the spatial and accuracy resolution of the experimental requirements. Additionally, there is an arm sensor that recorded quaternion data from the palm, forearm, and upper arm sensors. The system can capture human hand motion in real-time and drives 3D animation simultaneously to depict hand motions. This system is used to study participants' hand movements and finger coordination during assembling tasks.

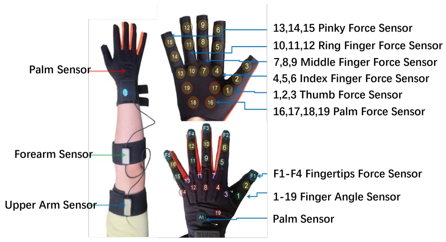

Fig. 2. Data glove with angle sensors and force sensors.

2) EMG and ACC: The muscle activity of the participants is accurately collected for studying the relevance of EMG data during assembly tasks. The Sensors monitor the electrical signals of the muscles through electrodes attached to the participant's skin and converts them into readable digital data. A total of 16 EMG sensors are used, with 8 sensors attached to the left side and 8 to the right side, each being attached to the basic muscle groups of the forearm.

The forearm muscles within the range of the EMG unit include the flexor carpi radialis, palmaris longus, flexor digitorum superficialis, and flexor carpi ulnaris. These muscles collectively participate in the adduction and abduction of the wrist joint as well as the flexion of the fingers. The actual deployment and the schematic diagram of the EMG acquisition unit on the human forearm are shown in Fig. 3.

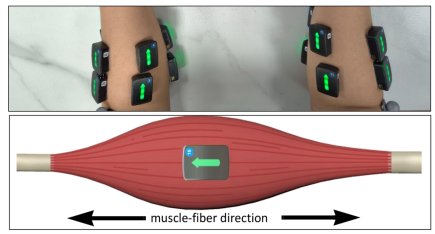

Fig. 3. EMG sensors distribution. The figure is sourced from the EMG device manual.

3) Environment depth and visual information: Camera is positioned directly in front of the participant to record the environmental information including RGB and depth information during the assembly process. This camera is used to capture participants' body postures and movements in real-time for investigating their motor performance and movement control in specific tasks.

4) Eye tracker: Eye-tracking data, including participants' gaze points mobility, are obtained to invesitgate visual attention and cognitive processes. The eye tracker is a binocular stereoscopic dark pupil tracking device that relies on corneal reflexes to record and analyze human eye tracking and first-person information in assembly processes. With an average accuracy of 0.6°, an average precision of 0.03°and an average data loss rate of 0.01%, the eye tracker device guarantees quality data collection.

5) Environment sound information: Four microphones are placed in different locations. Numbered 1-2, are placed on the operator's table to record the sound information of the operating environment. Others are placed in the accessory area to record ambient sound information for gripping tools and parts.

6) Ground truth: The 3D Motion Capture System is used to capture and record the movement and actions of a person or object in three-dimensional space. It utilizes high-precision optical sensors and cameras to track the movement trajectory of the object being measured and then transmitting the data to a computer for analysis and rendering. In the experiment, participants are fitted with reflective markers at 37 key points on the body to calibrate and build a mannequin. With this system, the ground truth information for analysis can be accurately captured.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

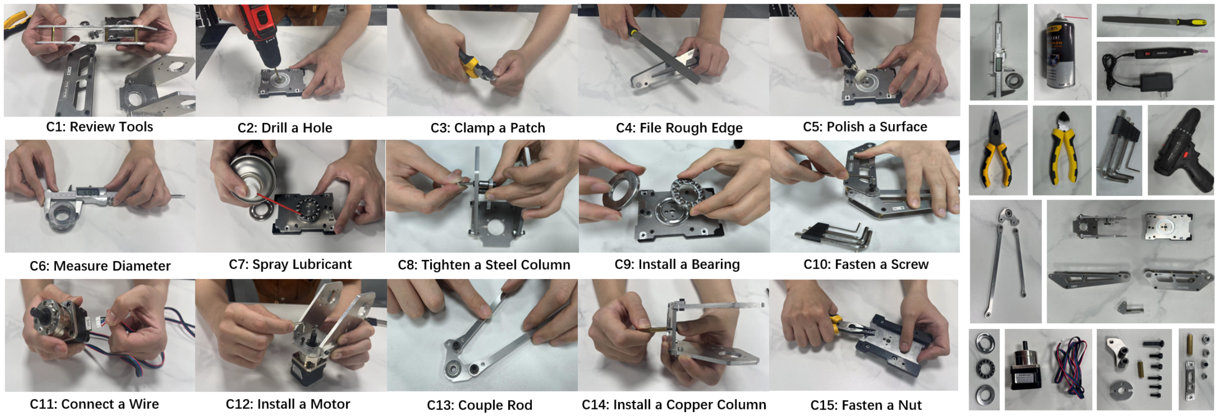

Fig. 4. Overview of assembling process. *C8* and *C14* have different focuses. The former collected experimental data on columnar parts that participants tightened by hand in a situation where there is ample mounting space. The latter captures experimental data by tools in a confined space with limited hand movement. The figure presents schematic illustrations of each action; therefore, the upper-limb sensors are not shown. During actual data collection, all volunteers wore the devices.

C. Calibration

Initial calibration of different participants can significantly improve the quality of metadata.

1) Equipment initialization calibration: The eye tracker uses a one-point calibration procedure to account for individual differences in eye shape and geometry, enhancing accuracy in predicting visual positions.

Physical stature varies across participants, so for each collection experiment, the motion capture system needs to be calibrated to the skeleton of the participants after the reflective markers are installed.

2) Process initialization calibration: After starting the experiments, the participants perform calibration of gesture movements to activate data synchronization and initialize the 19 sensors on the gloves. Each group of experiments can only be performed after completing the calibration process mentioned above for the corresponding assembly operation to ensure the synchronization of data acquisition.

3) Data synchronization: The data collection platform, designed with modular programming and multi-threading, ensures simultaneous start and synchronized acquisition across modules. It supports independent data collection and facilitates code adjustments. For devices with varying sampling rates, absolute timestamps are used to synchronize data streams during collection.

IV. Data Collection Protocol

A. Participants

Twenty participants with a mean height of 172.4±6.5 cm and a mean age of 24±4 yrs are recruited in this process, and all the consents are given. Each volunteer is recorded for 2 hours. The entire data collection process span 11 working days.

B. Assembling process

After the participant completed donning the device, calibration session, and data stream synchronization, data acquisition of the assembly actions began sequentially. During a single action acquisition session, the process would last 140 seconds, involving 0-10 seconds of data synchronization to receive calibration and 130-140 seconds of buffer time.

1) Assembly activity: In the assembly process of the robotic arm, links that closely reflect human dexterity and dynamics are selected for data collection. The entire assembly process is divided into 15 assembly links, with each link divided into left and right fine-grain motion elements labeled as action units. The segmentation facilitates the coding and integration of the entire assembly process into an action ensemble, and also facilitates subsequent labeling and recall. The assembly activities are shown in Fig. 4.

2) Assembly equipment and parts: For the action set, the relevant assembly objects include the robot front arm, end arm, middle arm, arm base, rear arm part, flange bearings, linkage, motor, steering bearings, brass post, steel post, bolt spacer, bolts, nuts, and power cable.

In our experiments, we do not subdivide the action units into finer degrees. Instead, we work with larger components because finger tactile sensors struggle to detect data changes in small, delicate parts. Before data collection, an initial pre-assembly of the small components is performed to meet acquisition requirements.

3) Tools: For all divided action sets, the following execution tools are required: drill, snips, file, polisher, vernier calipers, sharp-nosed pliers, and screwdriver.

C. Obtain data from humans

Imitation learning in robotics uses trajectory-based or bio-inspired methods to replicate human movements. Visual imitation learning faces kinematic challenges due to differences between human and robotic kinematics, requiring motion remapping and generalization algorithms [28], [29]. The Kaiwu dataset includes 3D spatial trajectories, multi-view RGB-D data, hand-related data (joint angles and tactile information), and EMG data. These data types support motion remapping and dynamic information capture. Glove-based solutions are used for synchronizing tactile and motion data [30], while eye-tracking and EMG data aid in designing optimization algo-

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

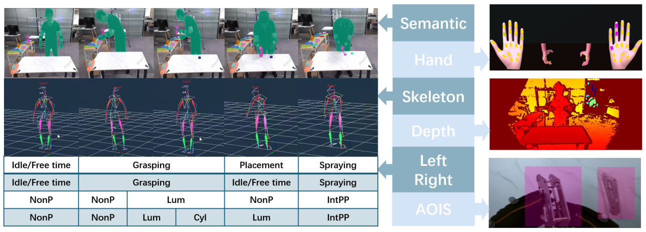

Fig. 5. Overview of annotation.

rithms [31]. Synchronization with other modalities is achieved through timestamp alignment during direct demonstrations.

Obtaining data directly from humans is crucial for human-robot interaction due to two key factors: First, the intentions of humans for actual actions are only observable through directly collected biosignals, which are not present in teleoperated trajectories, leading to intent uncertainty in real-world applications. Second, direct collection enables the construction of a first intention-to-motion mapping, while teleoperation lacks the necessary proprioceptive context for cross-embodiment transfer.

Although the proposed dataset construction method based on direct human demonstrations will increase costs compared to teleoperation or pure video methods, it better captures human motion intention information, essential for multimodal human-machine interaction. Teleoperation or pure video methods fail to provide sufficient interaction data, and additional efforts are required to fill data gaps and establish synchronization, which will also increase costs.

V. Data Annotation

An initial labeling of the dataset is performed to facilitate potential robot learning and human-robot interaction tasks. Various data are labeled to enhance the usability of the dataset. Subsequent users have the flexibility to annotate data according to their specific project requirements, while our initial annotations serve as a fundamental reference paradigm (Fig. 5). The annotation information, such as labels and instance, is shown in Table III.

TABLE III
Overview of Annotation.

Type of Mission	Annotation Element	Object Tags	Instance
Gesture Classification	Picture	10	4959
AOIs	Video	30	298
Semantic Segmentation	Closed Area	30	610778
Action Segmentation	Video Clip	26	7197
Gesture Segmentation	Video Clip	9	4467

A. Action&gesture segmentation and export

Fine-grained segmentation is conducted based on the timestamp series of actions and gestures to identify the start and end timestamps of the multimodal data collected from each assembly task action unit. The annotation includes:

- Action-level Annotation: Coarse segmentation of left and right hand actions.
- Fine-grain Gesture Annotation: Within the coarsely segmented time interval, a fine-grain segmentation of the left or right hand states is made

For action segmentation, the left and right-handed action segments of each participant are divided. The execution of actions in each segmentation interval is transitioned by the change of several gesture states, the actions of the left and right hands are further divided into corresponding hand state intervals, which are notated as Fine-grain Gesture Annotation. In the division of hand states, all hand states are divided into 8 gesture states [32]: Cylindrical grasp (Cyl), Oblique palmar grasp (Obl), Lumbrical grasp (Lum), Intermediate power-precision grasp (IntPP), Pinch grasp (Pinch), Lateral Pinch (LatP), Special pinch (SpP), Non-prehensile grasp (NonP).

The hand gestures are classified into two groups based on palm involvement: Group 1 (Cyl, Obl, IntPP) and Group 2 (Pinch, LatP, SpP, Lum). In Group 1, gestures are further divided by thumb and finger states: Cyl (thumb and fingers curved), Obl (thumb straight, fingers curved), and others as IntPP. Group 2 is divided by the number of fingers involved: Group 2a (2-3 fingers, including Pinch and LatP) and Group 2b (four fingers, including Lum and SpP). Finally, gestures are classified by lateral finger involvement: LatP and SpP involve lateral fingers, while Pinch and Lum do not.

B. Semantic segmentation

Semantic information annotation of 30 key objects in RGB video data is performed. By annotating the camera environment information, which records the third-person view of the participants, the data set is enriched with semantic information following the assembly logic. This enhancement aims to improve the robot's logical understanding of the assembly process and human action learning.

In semantic segmentation, semantic information is labeled by pumping the captured RGB video at one-second intervals to obtain frames with corresponding timestamps. The participants and all interaction tools are labeled. For objects with multiple entities, each object is labeled followed by the corresponding index for differentiation. For objects that are partially blocked in use, the remaining portion is locally labeled. All labeling errors are controlled within 8 pixels.

C. Gesture classification

According to the model playback, the state of the hand can be categorized [33] according to whether the thumb and four fingers are flexed or extended. Compared to the gesture segmentation in subsection 5.1, this part of the state segmentation simplifies the further division of the four-finger state in subsection 5.1.

D. Area of interest (AOIs)

The exported first-person view videos are labeled with a region of interest based on key object manipulation. This region contains the focus and intentional expression information within the participant's field of view, giving more weight to the data of objects in that specific area.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV
DESCRIPTIVE STATISTICS.

Data Type	Nums of Tasks	Objects	Duration(s)	File Size
Glove Data	300	51	22200	264 MB
Glove Export	300	51	22200	1124 MB
Eye Tracking	300	36	22200	14 GB
RGB-D Video	300	39	22200	3476 GB
Motion Capture Data	300	62	22200	4160 MB
Audio Data	300	15	22200	7955 MB
ACC Data	300	41	22200	354 MB
EMG Data	300	41	22200	362 MB

VI. DATA FORMAT

A. Descriptive statistics

A total of $6 \times 15 \times 20$ sets of experimental data on the assembling process are collected for Kaiwu dataset. The data comes from wearable sensors worn by 20 participants. Each participant is asked to complete 15 typical actions specified in the assembly process. Each participant spend approximately 18.5 minutes, resulting in the dataset representing about 6.16 hours of the assembling process. A summary of the related data is presented in Table IV.

B. Detailed information

This dataset contains experimental data from 20 participants in 20 folders with subject number. The folder corresponding to each participant contains the collected data from 6 recording devices.

1) EMG data: In EMGData folder, each sensor type outputs data in 17 TIME.csv files. Each file has 1 row for calibration values and 2-17 rows for signals from 16 sensors.

2) Glove data: For the assembly actions of C1-C15, the data gloves are transferred and stored as r/l_TIME.csv files with columns 1-12 representing the quaternion data of the palm, forearm, and upper arm, columns 13-31 representing the finger angle sensor data, and columns 32-50 representing the grip force sensor values. In addition, the data can be exported to present a visualization interface of the hand activity during data collection, and stored as an MP4 file.

3) RGB-D data: RGB-D data is stored in the C1-C15.mkv files in the RGB_video file and contains the timestamps used to generate the picture data, recorded as .txt files. The data generated based on the timestamp is stored in the IMG_picture file, including RGB images (.jpg file and .pcd file), depth information (.png file).

4) Ground truth data: After the initial body joint point calibration, the Motion Capture system captures and records the 3D spatial coordinates of each marker point as the assembly process proceeds, outputting the data as a (.cap .trb .xrb) File. Visualization is achieved by reimporting the data.

5) Eye tracker data: The Tobii-data folder includes first-person perspective data (tobii-data) and eye-tracking data (tobii-export). Tobii-data contains compressed .gz files and a visualization .mp4 file. Tobii-export includes .mp4 files showing gaze point movements for C1-C15 and .xlsx files.

6) Voice data: The microphone outputs environment sound information as .wav audio files and .txt timestamp data.

VII. INTENT RECOGNITION EXPERIMENT IN HUMAN-ROBOT INTERACTION

In this section, we analyzed and validated the intent recognition results on the Kaiwu dataset using existing models. The comparative analysis evaluated the strengths and weaknesses of these models. The results are shown in Table V.

TABLE V
THE RESULTS OF MULTIPLE METHODS.

Modality	Methods	Accuracy	Average Reliability
Unimodal	EMG-Based	54%	-%
	Gaze-Based	66%	-%
	Gesture-Based	72%	-%
Multimodal	Decision Matrix (DM)	71%	65.35%
	Bayesian Inference (BI)	75%	72.67%
	Dempster-Shafer (DS)	73%	83.33%
	Multimodal Intent Fusion (MIF)	84%	87.25%
	HIL-MIF	92%	89.94%

We employed the Human-in-the-Loop Multimodal Intent Fusion(HIL-MIF) [31] algorithm, which demonstrated the best multimodal performance, to evaluate the effectiveness of EMG data in enhancing multimodal data. Experiments and comparisons were conducted, with results shown in Table VI.

TABLE VI
THE RESULTS OF HIL-MIF METHODS WITH EMG DATA ENHANCING.

EMG-enhanced	Accuracy	Average Reliability	Standard Deviation
×	92%	89.94%	0.1459
√	**95%**	**90.55%**	**0.1247**

The results show that the HIL-MIF algorithm augmented with EMG data outperforms those without EMG data, indicating its potential as a robust and effective solution for human-robot interaction and other applications requiring accurate intent recognition.

VIII. ACCESSING THE DATA

The data has been uploaded to ScienceDB with DOI number: https://doi.org/10.57760/sciencedb.14937. The homepage of this dataset can be searched using the keyword ”Kaiwu” on ScienceDB.

The following files are available for download: 1) Kaiwu Data Annotation includes action segmentation with timestamped delsys_data and gloves_data. 2) AOIs includes annotation results and annotation process. 3) gesture segmentation includes delsys_data and gloves_data. 4) semantic segmentation includes RGB video annotation. 5) gesture classification includes hand state classification data. The dataset includes missing data and additions to the raw data. It contains raw voice, motion capture, RGB-D data, glove and eye tracker videos. The annotated data is saved as a .csv file, with timestamps used to index and call corresponding data from different sensors, and includes annotation information.

IX. CONCLUSION

This paper introduces a data collection framework and platform to build Kaiwu dataset, which enable to improve

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/LRA.2025.3609615

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

understanding of human assembly activity logic and robot learning. Multi-modal wearable sensors are equipped to capture various egocentric data which contains rich visual and dynamics info. This is also coupled with environmental and global ground truth info to build an absolute spatial coordinate frame. An initial labeling of the dataset is also annotated to facilitate the training of deep neural networks with multi-modal data.

Future work can utilize the Kaiwu dataset to develop knowledge pathways for embodied learning and analysis, exploring topics such as cross-modal prediction, assembly logic sequence prediction, assembly task planning, and robot self-assembly. In addition, we hope the collection platform of the Kaiwu dataset can be utilized as a medium for robot behavioral learning and human-robot skill transfer to explore new collection modes. Finally, the strategies and methods employed by imitation learning models in dealing with unobserved variables, as well as the kinematics issues in tasks involving contact or collision, will also be important directions for future research.

REFERENCES

[1] A. Zador, S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath *et al.*, “Catalyzing next-generation artificial intelligence through neuroai,” *Nature communications*, vol. 14, no. 1, p. 1597, 2023.

[2] H. Ding, X. Yang, N. Zheng, M. Li, Y. Lai, and H. Wu, “Tri-co robot: a chinese robotic research initiative for enhanced robot interaction capabilities,” *National Science Review*, vol. 5, no. 6, pp. 799–801, 2018.

[3] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman *et al.*, “Foundation models in robotics: Applications, challenges, and the future,” *The International Journal of Robotics Research*, p. 02783649241281508, 2023.

[4] X. Xu, M. You, H. Zhou, Z. Qian, and B. He, “Robot imitation learning from image-only observation without real-world interaction,” *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 3, pp. 1234–1244, 2022.

[5] Z. Fu, T. Z. Zhao, and C. Finn, “Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation,” *arXiv preprint arXiv:2401.02117*, 2024.

[6] Q. Zhang, P. Cui, D. Yan, J. Sun, Y. Duan, A. Zhang, and R. Xu, “Whole-body humanoid robot locomotion with human reference,” *arXiv preprint arXiv:2402.18294*, 2024.

[7] X. Xiao, J. Liu, Z. Wang, Y. Zhou, Y. Qi, Q. Cheng, B. He, and S. Jiang, “Robot learning in the era of foundation models: A survey,” *arXiv preprint arXiv:2311.14379*, 2023.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, “Rt-1: Robotics transformer for real-world control at scale,” *arXiv preprint arXiv:2212.06817*, 2022.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” *arXiv preprint arXiv:2307.15818*, 2023.

[10] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh, “Rt-h: Action hierarchies using language,” *arXiv preprint arXiv:2403.01823*, 2024.

[11] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, “Maniwav: Learning robot manipulation from in-the-wild audio-visual data,” *arXiv preprint arXiv:2406.19464*, 2024.

[12] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, “Rh20t: A robotic dataset for learning diverse skills in one-shot,” in *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[13] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, “Droid: A large-scale in-the-wild robot manipulation dataset,” *arXiv preprint arXiv:2403.12945*, 2024.

[14] H.-S. Fang, M. Gou, C. Wang, and C. Lu, “Robust grasping across diverse sensor qualities: The graspnet-1billion dataset,” *The International Journal of Robotics Research*, vol. 42, no. 12, pp. 1094–1103, 2023.

[15] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models,” *arXiv preprint arXiv:2310.08864*, 2023.

[16] Z. Wang, H. Zheng, Y. Nie, W. Xu, Q. Wang, H. Ye, Z. Li, K. Zhang, X. Cheng, W. Dong *et al.*, “All robots in one: A new standard and unified dataset for versatile, general-purpose embodied agents,” *arXiv preprint arXiv:2408.10899*, 2024.

[17] P. Maurice, A. Malaisé, C. Amiot, N. Paris, G.-J. Richard, O. Rochel, and S. Ivaldi, “Human movement and ergonomics: An industry-oriented dataset for collaborative robotics,” *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1529–1537, 2019.

[18] J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, and D. Rus, “Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 800–13 813, 2022.

[19] J. S. Yoon, Z. Yu, J. Park, and H. S. Park, “Humbi: A large multiview dataset of human body expressions and benchmark challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 623–640, 2021.

[20] R. Dai, S. Das, S. Sharma, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, “Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2533–2550, 2022.

[21] B. Reily, P. Gao, F. Han, H. Wang, and H. Zhang, “Real-time recognition of team behaviors by multisensory graph-embedded robot learning,” *The International Journal of Robotics Research*, vol. 41, no. 8, pp. 798–811, 2022.

[22] Y. Huang and Y. Sun, “A dataset of daily interactive manipulation,” *The International Journal of Robotics Research*, vol. 38, no. 8, pp. 879–886, 2019.

[23] R. Martín-Martín, C. Eppner, and O. Brock, “The rbo dataset of articulated objects and interactions,” *The International Journal of Robotics Research*, vol. 38, no. 9, pp. 1013–1019, 2019.

[24] B. A. Newman, R. M. Aronson, S. S. Srinivasa, K. Kitani, and H. Admoni, “Harmonic: A multimodal dataset of assistive human–robot collaboration,” *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 3–11, 2022.

[25] P. Kang, K. Zhu, S. Jiang, B. He, and P. B. Shull, “Hbod: A novel dataset with synchronized hand, body, and object manipulation data for human-robot interaction,” pp. 1–4, 2023.

[26] L. Lastrico, V. Belcamino, A. Carfì, A. Vignolo, A. Sciutti, F. Mastrogiovanni, and F. Rea, “The effects of selected object features on a pick-and-place task: A human multimodal dataset,” *The International Journal of Robotics Research*, vol. 43, no. 1, pp. 98–109, 2024.

[27] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. Lu, “Oakink2: A dataset of bimanual hands-object manipulation in complex task completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024.

[28] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation,” pp. 5628–5635, 2018.

[29] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, “Learning dexterous in-hand manipulation,” *The International Journal of Robotics Research*, pp. 3–20, 2020.

[30] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, and S.-C. Zhu, “A glove-based system for studying hand-object manipulation via joint pose and force sensing,” *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6617–6624, 2017.

[31] S. Jiang, W. Zhang, J. Liu, H. Li, Z. Wang, Y. Zhou, and B. He, “A novel human-in-the-loop multimodal intention fusion method for human-robot interaction,” *IEEE Transactions on Automation Science and Engineering*, pp. 16 888–16 898, 2025.

[32] M. Vergara, J. L. Sancho-Bru, V. Gracia-Ibáñez, and A. Pérez-González, “An introductory study of common grasps used by adults during performance of activities of daily living,” *Journal of Hand Therapy*, vol. 27, no. 3, pp. 225–234, 2014.

[33] R. Gopura and D. Bandara, “A hand prosthesis with an under-actuated and self-adaptive finger mechanism,” *Engineering*, vol. 10, no. 7, pp. 448–463, 2018.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 17,2025 at 09:11:54 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.