arXiv:2607.28625v1 [cs.CV] 30 Jul 2026

ACE-Data-0: Human-Centric Ambient Capture as Embodied Data Engine

Yukang Cao[1*], Haozhe Xie[1*], Beichen Wen[2*], Runmao Yao[1], Yinghao Liu[2], Yue Huang[2], Zhichao Liao[2], Yunxiang Wang[1], Haiheng Liu[1], Xingshun Tian[2], Dawei Su[2], Long Zhuo[2], Dacheng Tao[2†], Xiaogang Wang[2†], Liang Pan[2‡], Ziwei Liu[1‡]

[1] S-Lab, Nanyang Technological University [2] ACE Robotics

[*] Equal Contributions [†] Project Advisor [‡] Project Lead

https://ace-data-engine.github.io/ACE-Data-0/

Figure 1 We introduce the Ambient Capture Engine (**ACE**), a capture system that transforms real home environments into recording studios. ACE observes household activity at two complementary scales, *i.e.*, table-scale and room-scale. Via ACE, we capture and release ACE-Data-0 with synchronized multiple modalities and rich data annotations.

Abstract

Embodied intelligence faces a fundamental data bottleneck. Learning to act in the physical world requires more than observing what an action looks like: *models must capture how first-person perception, whole-body motion, dexterous manipulation, object state, sound, and touch evolve together as humans pursue goals over time.* Existing datasets typically fragment this experience across viewpoints, modalities, or spatial scales, leaving the full perception-action loop only partially observed.

We introduce the **A**mbient **C**apture **E**ngine (ACE), a human-centric data engine that transforms real home environments into spatially calibrated, temporally synchronized recording studios. ACE operates at two complementary scales: a **table-scale** configuration resolves fine-grained hand-object manipulation, while a **room-scale** configuration captures whole-body motion, locomotion, and interactions distributed across a furnished home. Across both settings, ACE records egocentric and multi-view exocentric video, full-body and articulated hand motion, high-fidelity object geometry and 6-DoF trajectories, multi-channel audio, and tactile signals as a unified multisensory stream.

Using ACE, we build **ACE-Data-0**, comprising 150 hours and 17M video frames across 200 task categories, performed by 50 participants in 2 environments, for a total of 75,000 interaction episodes. The dataset spans atomic manipulation, long-horizon chains of household activities, and human-scene interaction, while preserving natural behavioral variation through goal-level rather than step-by-step instructions. We further introduce a hierarchical benchmark that progresses from **signals** to **scene components** and then to **interactions**. It encompasses tactile inference from visual observations, human motion recovery, and hand-motion reconstruction from both egocentric and exocentric perspectives. Evaluations of state-of-the-art methods expose substantial gaps under contact, occlusion, egomotion, and long temporal horizons.

Beyond benchmarking, ACE-Data-0 provides synchronized human demonstrations with aligned perceptual, kinematic, and contact supervision, offering a scalable foundation for imitation learning, world models, vision-language-action systems, and embodied AI.

Project Page: https://ace-data-engine.github.io/ACE-Data-0/
Hugging Face: https://huggingface.co/datasets/ACERobotics/ACE-Data-0

Contents

1 Introduction **4**

2 Related Work **5**
2.1 Multi-modal Datasets & Benchmarks . . . 5
2.2 Egocentric Datasets & Benchmarks . . . 7
2.3 Long-horizon Datasets & Benchmarks . . . 8

3 Ambient Capture Engine **9**
3.1 System Overview . . . 9
3.2 Capture System . . . 10
3.2.1 Environment Setup . . . 10
3.2.2 Hardware Setup . . . 11

4 ACE-Data-0 **12**
4.1 Data Acquisition Pipeline . . . 12
4.1.1 Sensor Synchronization . . . 12
4.1.2 Calibration . . . 14
4.1.3 Capture Workflow . . . 16
4.1.4 Multi-modal Recording . . . 16
4.2 Data Collection . . . 16
4.2.1 Task Design . . . 16
4.2.2 Capture Settings . . . 17
4.2.3 Dataset Statistics . . . 18
4.2.4 Modalities . . . 18
4.3 Annotation . . . 19
4.3.1 Annotation Types . . . 20
4.3.2 Annotation Pipeline . . . 22

5 Benchmark **22**
5.1 Low-level Signals . . . 23
5.1.1 Tactile from Vision . . . 23
5.2 Scene Components . . . 24
5.2.1 Human Motion Estimation . . . 25
5.3 Embodied Interaction . . . 26
5.3.1 HOI from Ego-View . . . 26
5.3.2 HOI from Exo-View . . . 27
5.3.3 Cross-View Analysis . . . 28

6 Conclusion **28**

1 Introduction

Humans spend the majority of their lives in built environments, continuously interacting with the objects around them, such as opening cabinets, pouring water, folding laundry, and assembling furniture. Effortless as they appear, these everyday interactions embody a form of intelligence that existing models have yet to attain: the seamless coordination of perception, whole-body movement, dexterous manipulation, and physical sensing [1, 2]. Reproducing this intelligence is a central pursuit of embodied AI, yet unlike the AI breakthroughs that preceded it, this pursuit inherits no data. Language and vision models [3–6] rose on archives that humanity had spent centuries accumulating. Physical skills, exercised without deliberation, have never been written down: how a hand closes around a cup, with what force a fragile glass is held, by what coordination of vision and balance an object is carried across a room. The data for embodied intelligence must therefore be built rather than found, by instrumenting everyday life itself and recording human-object interaction (HOI) as it naturally unfolds.

Constructing such a dataset, however, demands far more than pointing a camera at daily life. It requires a *holistic* recording of the interaction, spanning *what* the human is doing, *how* the body and hands move, *how* the object state evolves in response, *what* audio and tactile signals the interaction produces, and *how* the scene appears from both the human's first-person perspective and third-person viewports. More importantly, such a dataset for learning embodied AI demands all of these signals to be recorded synchronously and in registration with one another, a requirement that no existing dataset satisfies.

Specifically, existing datasets fall short in three aspects, as summarized in Table 1: **(1) Fragmented modalities.** Large-scale egocentric datasets (*e.g.*, Ego-Exo4D [7], EPIC-Kitchens [8], Xperience-10M [9]) offer naturalistic behavior at scale, yet provide neither ground-truth body and object motion nor synchronized third-person observation. Conversely, motion-captured HOI datasets, such as BEHAVE [10], GRAB [11], ARCTIC [12], OakInk2 [13], and HOT3D [14], supply accurate poses but omit the egocentric perspective. More importantly, audio or tactile sensing is absent from nearly all of them. **(2) Unnatural environments.** Physically annotated datasets [11, 12, 15–17] are captured almost exclusively in laboratories, whose sparse layouts eliminate precisely the occlusions, spatial constraints, and object diversity that render real homes challenging. **(3) Short horizons.** The vast majority of HOI clips span seconds and depict one simple movement [12, 15, 16, 18]. Genuine household activities, in contrast, are goal-directed. They may unfold over minutes or hours, chain together sub-tasks and multiple objects, and require the human to move across the scene rather than acting in a single place. Consequently, the complete perception-action loop of everyday interaction has remained beyond the reach of existing benchmarks.

To bridge this gap, we develop the *Ambient Capture Engine* (ACE), a capture system that turns real home environments into recording studios while preserving their lived-in realism. Fine-grained dexterous manipulation and room-scale activity impose conflicting requirements on sensor placement and coverage, so we build ACE as two complementary configurations. The first is table-scale: densely arranged close-range cameras and high-resolution tactile sensing resolve the local details of hand-object manipulation. The second is room-scale: sensors span an entire furnished environment to record whole-body motion, locomotion, and interactions distributed across the scene. While differing in sensor placement and spatial scope, both scales record synchronized egocentric video, multi-view exocentric video, optical full-body motion, and per-object 6-DoF trajectories, which are all registered into a common spatio-temporal frame.

With ACE, we collect **ACE-Data-0**: 150 hours of daily living interactions spanning 200 task categories (*e.g.*, cooking, tidying, and drinking), performed by 50 participants across 2 environments, amounting to 17M video frames and 75,000 interaction episodes. Rather than executing scripts, participants pursue goal-level instructions in their own manner: planning, hesitating, and improvising as they would at home. Crucially, this freedom sacrifices no measurement fidelity: every moment of activity is observed simultaneously from more than 8 viewpoints and grounded in metric body, hand, and object states. Upon these raw signals, we provide rich annotations for every sequence, including camera cal-

ibrations and synchronized timelines; full-body and hand poses and their projections onto every frame of every camera; per-object mesh models, 6-DoF poses, bounding boxes, and motion trails; and textual descriptions of the ongoing events. A large fraction of these annotations is derived automatically from the tracked physical states, instead of being estimated via existing pipelines.

Building upon ACE-Data-0, we further establish a perceptually hierarchical benchmark, comprising three tracks that advance from signals, to components, to interactions. The first track targets **low-level signals**: cross-modal prediction of tactile signals from visual observations. The second track focuses on **scene components**: methods reconstruct the pose of the human body and hands, evaluated against our tracked ground-truth. The third track evaluates the **interactions**: models estimate the hand motion during human-object interactions from egocentric and exocentric videos. Additionally, our collected paired viewpoints allow for a direct comparison between these two. Evaluations of more than 30 representative methods expose failure modes characteristic of long-horizon home-scene data, stemming from the diversity of interactions and tasks, their complexity and interleaving, heavy occlusion, extreme viewpoints, and continual movement through the scene. Notably, these three tracks mirror the perceptual capabilities an embodied agent must chain together: sensing contact, estimating scene state, and mastering hand-object interactions. The value of ACE-Data-0 to such agents, moreover, extends beyond evaluation to training itself: because our egocentric viewpoint, multi-view exocentric videos, motion trajectories, and contact-level supervision are synchronized rather than assembled from disparate sources, every training signal refers to the same physical moment, a property essential for robot learning, from imitation and policy learning to world modeling [19–22].

In general, our contributions can be summarized as follows:

- We present **ACE**, a "human-centric ambient capture as embodied data engine paradigm" realized as two complementary configurations, a table-scale setup for fine-grained local manipulation and a room-scale setup for whole-body motion in larger scenes. Within these two configurations, ACE records both temporally and spatially aligned egocentric video, multi-view exocentric video, human body and hand motion, object motion, audio, and tactile signals.
- We contribute **ACE-Data-0**, a large-scale, long-horizon home-scene HOI dataset comprising 17M frames and 75,000 interaction episodes, together with rich and high-quality annotations.
- We establish a three-level benchmark that advances from signals to components and ultimately to interactions, and provide evaluations of more than 30 state-of-the-art methods that present the open challenges for embodied perception and robot learning.

2 Related Work

2.1 Multi-modal Datasets & Benchmarks

Everyday interaction is inherently multi-modal. A single physical event is reflected simultaneously in visual appearance, body and hand motion, changes in object state, acoustic cues, and physical contact. These signals are complementary rather than interchangeable: vision captures what is externally observable, kinematics describes how the human and objects evolve, audio reveals impacts and state transitions, and touch provides direct evidence of contact and force. Accordingly, datasets for embodied perception have progressively moved beyond isolated RGB observations toward richer combinations of geometry, motion, contact, and semantic annotations.

Early physically grounded datasets primarily focused on the local geometry of grasping and manipulation. ContactPose [15] associates 3D hand pose with object-surface contact over approximately 2.9M multi-view RGB-D images from 50 participants and 25 objects, while DexYCB [16] provides 582K frames from eight synchronized RGB-D cameras with MANO hand [49] annotations and YCB object poses. H_2O [18] extends this setting to bimanual interaction through synchronized first- and third-person

Table 1 Comparison of ACE-Data-0 with related datasets, grouped by category. ✓: provided with measured (mocap-grade) ground truth; (✓): provided but estimated, pseudo-labeled, or device-tracked; ✓: partially provided (limited coverage or a subset); ✗: capability absent; –: number not applicable or not reported; a ✓in the #Exo column denotes that exocentric video is provided without a reported camera count; *Sync*: synchronization across modality families (ego video, exo video, motion, audio, tactile); a ✓requires at least two families, chiefly ego-exo or video-motion, recorded simultaneously and aligned, while ✓denotes that the synchronization is conducted only among one type of viewpoint, and '–' means that only one perspective is captured. *LH*: long-horizon (goal-directed activities of minutes or longer). *Setup*: capture environment. In the #Tasks column, dom., cat., scen., and skills denote domains, interaction categories, scenarios, and skills, following each dataset's own task organization.

Dataset	Year	Hours	#Frames	#Subj	#Obj	#Tasks	Ego	#Exo	Body	Hand	Obj. 6D	Tactile	Audio	Sync	Setup	LH
Egocentric video																
Ego4D [23]	2022	3670	–	931	–	Open	✓	✗	✗	✗	✗	✗	✓	–	In-the-wild	✓
EPIC-KITCHENS-100 [8]	2021	100	20M	37	–	Open	✓	✗	✗	✗	✗	✗	✓	–	Kitchens	✓
HoloAssist [24]	2023	166	–	222	16	20	✓	✗	✗	(✓)	✗	✗	✓	–	Desktop	✓
Ego-Exo4D [7]	2024	1286	–	740	–	8 Dom.	✓	4–5	(✓)	✗	✗	✗	✓	✓	In-the-wild	✓
EgoLife [25]	2025	266	–	6	–	Open	✓	✓	✗	✗	✗	✗	✓	✗	Shared house	✓
HD-EPIC [26]	2025	41	4.46M	9	–	69	✓	✗	✗	✗	(✓)	✗	✓	–	Home kitchens	✓
Hand–object interaction																
ContactPose [15]	2020	–	2.9M	50	25	2	✗	3	✗	✓	✓	✓	✗	✗	Table-top	✗
HO-3D [27]	2020	–	78K	10	10	–	✗	1–5	✗	✓	✓	✗	✗	✗	Table-top	✗
GRAB [11]	2020	–	1.6M	10	51	4	✗	✗	✓	✓	✓	(✓)	✗	–	Mocap lab	✗
DexYCB [16]	2021	–	582K	10	20	1	✗	8	✗	✓	✓	✗	✗	✗	Table-top	✗
H_2O [18]	2021	–	571K	4	8	36	✓	4	✗	✓	✓	✗	✗	✓	Table-top	✗
OakInk [28]	2022	–	230K	12	100	5	✗	4	✗	✓	✓	✗	✗	✓	Table-top	✗
HOI4D [29]	2022	22	2.4M	9	800	54	✓	✗	✗	✓	✓	✗	✗	–	Indoor rooms	✗
ARCTIC [12]	2023	–	2.1M	10	11	2	✓	8	✓	✓	✓	✗	✗	✓	Mocap lab	✗
TACO [30]	2024	–	5.2M	14	196	151	✓	12	✗	✓	✓	✗	✗	✓	Table-top	✗
HOT3D [14]	2024	13.9	3.7M	19	33	Open	✓	✗	✗	✓	✓	✗	✗	–	Lab rooms	✗
OakInk2 [13]	2024	–	4.0M	9	75	150	✓	3	(✓)	✓	✓	✗	✗	✓	Table-top	(✓)
GigaHands [17]	2025	34	183M	56	417	Open	✗	51	✗	(✓)	(✓)	✗	✗	✗	Table-top	✗
Full-body HOI, human-scene interaction, and daily motion																
BEHAVE [10]	2022	–	15K	8	20	–	✗	4	✓	✗	✓	✗	✗	✓	Lab rooms	✗
InterCap [31]	2022	–	67K	10	10	–	✗	6	✓	(✓)	✓	✗	✗	✓	Lab room	✗
CHAIRS [32]	2022	17.3	–	46	81	32	✗	4	✓	✓	✓	✗	✗	✓	Lab room	–
EgoBody [33]	2022	–	220K	36	–	5 Cat.	✓	3–5	(✓)	(✓)	✗	✗	✗	✓	Indoor rooms	✗
Aria Digital Twin [34]	2023	6.6	–	–	398	Open	✓	✗	(✓)	✗	✓	✗	✓	–	Apartment, office	✗
OMOMO [35]	2023	10	–	17	15	–	✗	✗	✓	✗	✓	✗	✗	–	Lab room	✗
HIMO [36]	2024	–	4.1M	34	53	–	✓	✗	✓	✓	✓	✗	✗	–	Mocap lab	✗
TRUMANS [37]	2024	15	1.6M	7	20	–	✗	✗	✓	✗	(✓)	✗	✗	–	Scene mockups	✗
ParaHome [38]	2024	8.1	–	38	22	–	✗	70	✓	✓	✓	✗	✗	✓	Home room	✓
Nymeria [39]	2024	300	–	264	–	20 Scen.	✓	✓	✓	✗	✗	✗	✓	✓	In-the-wild	✓
HuMoTo [40]	2025	2.2	236K	1	63	–	✗	✗	✓	✓	✓	✗	✗	–	Scene	✗
Robot data and human demonstrations for robots																
BridgeData V2 [41]	2023	–	60K Traj	–	100+	13 Skills	✗	✓	✗	✗	✗	✗	✗	–	Table-top	✗
RH20T [42]	2023	–	110K Traj	–	–	147	✓	✓	✗	✗	✗	(✓)	✓	✓	Table-top	✗
DexCap [43]	2024	–	–	–	–	–	✗	✓	✗	✓	✗	✗	✗	✓	Table-top	✗
DROID [44]	2024	350	76K Traj	–	–	86	✗	3	✗	✗	✗	✗	✗	✓	In-the-wild	✗
AgiBot World [45]	2025	2976.4	1M Traj	–	3000+	217	✗	✓	✗	✗	✗	✓	✗	✗	Staged scenes	✓
EgoDex [46]	2025	829	90M	–	–	194	✓	✗	✗	(✓)	✗	✗	✗	–	Table-top	✗
Galaxea Open-World [47]	2025	500	100K Traj	–	1600	150	✗	✓	✗	✗	✗	✗	✗	✓	Open-world	(✓)
EgoScale [48]	2026	20,854	–	–	43,237	6015	✓	✗	–	–	–	✗	✗	–	Table-top	✗
Xperience-10M [9]	2026	10,000	2.88B	–	–	–	✓	✗	(✓)	(✓)	✗	✗	✓	–	In-the-wild	✗
Ours																
ACE-Data-0	2026	**150**	**17M**	**50**	**50**	**200**	✓	**8**	✓	✓	✓	✓	✓	✓	**Home (Room + Table-top)**	✓

RGB-D observations, 3D hand keypoints, and object poses. H_2O-3D [50] further contributes more than 76K annotated images for challenging two-hand-object pose estimation under severe self-occlusion. At a larger semantic scale, HOI4D [29] contains 2.4M egocentric RGB-D frames involving roughly 800 object instances and supports category-level object pose tracking, action segmentation, and dynamic 4D interaction understanding. A complementary line of work expands the captured state from the hands to the full human-object system. GRAB [11] records detailed whole-body, hand, face, and object motion for interactions with 51 objects. BEHAVE [10] and InterCap [31] jointly recover humans and rigid objects from multi-view RGB-D observations, enabling the study of body-scale coordination, contact, and joint reconstruction.

More recent datasets increasingly emphasize dexterity, compositionality, and task structure. ARCTIC [12] captures approximately 2.1M images of bimanual interaction with articulated objects from one egocentric and eight allocentric views, grounded by optical motion capture. TACO [30] organizes 5.2M

frames around compositional tool-action-object relationships, whereas OakInk2 [13] represents complex bimanual activity through a hierarchy of affordances, motion primitives, and task-level structures. At a larger scale, GigaHands [17] collects roughly 183M frames from 56 participants and 417 objects using a 51-camera system, substantially increasing the diversity of bimanual activities and language annotations. HOT3D [14] provides approximately 3.7M egocentric multi-view images from 19 participants and 33 scanned objects, together with metric trajectories for hands, objects, cameras, and gaze. Collectively, these datasets have substantially advanced hand pose estimation [51–53], human-object reconstruction [31, 54], contact reasoning [12, 15, 55], and interaction understanding [14, 17, 30]. Their sensing configurations, however, are typically optimized for a particular spatial scale or source of supervision. Hand-centric datasets resolve detailed local geometry but usually confine activity to a compact workspace, whereas whole-body datasets preserve global coordination but may offer less detailed hand articulation or omit the actor's first-person view. Existing multi-modal resources also predominantly combine RGB, depth, pose, geometry, gaze, and language [56]. Audio is rarely aligned with metric human and object motion at the interaction level, and tactile sensing is scarcer still, despite directly revealing contact onset, pressure, and slip. Moreover, annotations may combine direct measurement with offline fitting or model-based reconstruction, limiting their uniformity under sustained occlusion, rapid motion, or long-duration activity. High-fidelity capture is therefore often achieved in controlled spaces that simplify the clutter, furniture occlusion, and spatial constraints of domestic interaction.

ACE-Data-0 complements these efforts by treating synchronized multisensory grounding as the central unit of data. The table-scale configuration resolves fine-grained hand-object manipulation, whereas the room-scale configuration captures full-body motion and interactions distributed across a furnished domestic environment. Both configurations provide egocentric video, multi-view exocentric video, full-body and articulated hand motion, object 6-DoF trajectories, audio, and tactile signals through a shared synchronization and calibration pipeline. The modalities therefore describe the same physical event on a common timeline and in a common spatial frame, rather than serving as independently produced annotations. This unified grounding supports a coherent benchmark progression from low-level signal inference, through scene component recovery, to interaction understanding.

2.2 Egocentric Datasets & Benchmarks

Egocentric video observes activity from the viewpoint of the acting subject rather than that of an external camera. This perspective naturally emphasizes action-relevant objects, hand proximity, gaze allocation, and the immediate visual consequences of self-motion, making it particularly relevant to embodied agents. At the same time, it introduces distinctive challenges: the camera moves continuously, motion blur is frequent, the hands often occlude manipulated objects, and much of the actor's body remains outside the field of view.

Large-scale datasets first established the breadth and naturalism of egocentric perception. Ego4D [23] collects 3,670 hours of unscripted first-person video from 931 camera wearers across 74 locations, with benchmarks spanning episodic memory, forecasting, social interaction, manipulation, and audio-visual understanding. HoloAssist [24] focuses on interactive task assistance, recording 166 hours of instructor-performer activity with synchronized RGB-D, gaze, head and hand pose, IMU, audio, and dialogue; these signals support mistake detection, intervention prediction, and hand-motion forecasting. Ego-Exo4D [7] further connects first- and third-person perspectives through 1,286 hours of synchronized skilled activity from 740 participants across 123 scenarios, together with audio, gaze, language, camera poses, and 3D scene information. Its paired-view design enables cross-view activity understanding, proficiency estimation, and pose recovery on the same underlying events.

Manipulation-oriented datasets trade some behavioral breadth for stronger physical supervision. H_2O [18] and HOI4D [29] associate first-person RGB-D observations with hand and object states, while HOT3D [14] provides calibrated egocentric multi-view recordings with metric trajectories for hands, objects, cameras, and gaze. EgoDex [46] scales human dexterous demonstrations to 829 hours across 194

task types using wearable hand tracking, creating a large source of action-relevant human manipulation data. PH2D [57] further connects human first-person demonstrations with humanoid policy learning through a robot-compatible action representation. A complementary source of first-person experience comes from robot demonstration data. Fourier ActionNet [58] records more than 30K teleoperated trajectories, totaling approximately 140 hours of dexterous bimanual manipulation across multiple humanoid platforms; it pairs observations from the robot's first-person cameras with language annotations and executable robot actions. Human egocentric datasets and robot demonstrations therefore provide complementary forms of supervision: the former preserve natural human strategies and embodiment-independent behavior, whereas the latter provide direct action labels in the target robotic embodiment.

Recent human-to-robot learning methods increasingly combine these complementary data sources. DexMV [59], EgoMimic [60], and EgoBridge [61] explicitly align or adapt human demonstrations to robotic embodiments. EgoVLA [20], UniVLA [62], and In-N-On [63] instead learn transferable action representations from mixtures of human videos and robot demonstrations. EgoScale [48] and Emergence of Human to Robot Transfer [64] further suggest that transfer benefits from increasing the scale and diversity of human experience, robot tasks, and embodiments. These developments reinforce the value of first-person data, but also expose a recurring trade-off between behavioral diversity and physical completeness. Large naturalistic collections capture broad activity distributions and extended temporal context, yet do not continuously measure full-body state, object trajectories, and physical contact throughout every sequence. More instrumented datasets provide stronger hand or object supervision, but often remain local in spatial scope and omit some combination of synchronized room-scale external views, audio, or tactile sensing.

ACE-Data-0 addresses this gap by embedding egocentric observation within a shared physical representation of the surrounding event. The wearable cameras preserve the close-range perspective available to an embodied agent, while synchronized exocentric views recover body and scene context that is frequently invisible from the first-person view. Both perspectives are registered with measured headset, human, and object motion, enabling direct comparison and fusion under the same physical ground truth. Audio and tactile streams further connect visual observation to the acoustic and contact consequences of the same interaction.

2.3 Long-horizon Datasets & Benchmarks

Long-horizon interaction [13, 38, 65, 66] is defined not merely by recording duration, but by persistent dependencies among actions, objects, and scene states. In household tasks, earlier decisions alter later possibilities, objects move in and out of relevance, and sub-tasks may be interrupted, reordered, or resumed elsewhere. Locomotion and manipulation are likewise interdependent: the actor must move through the environment while retaining object locations, intermediate states, and the remaining goal. These properties require models to maintain task and scene memory rather than treating an activity as a sequence of independent atomic actions.

Several human-centered datasets have begun to preserve this richer temporal structure. OakInk2 [13] organizes bimanual manipulation into hierarchical task components rather than isolated grasps, supporting the study of affordances, motion primitives, and complex task completion. OMOMO [35] models the temporal coupling between object motion and full-body human motion over extended interactions, demonstrating how object trajectories constrain human behavior beyond a single contact event. ParaHome [38] records 207 household activity sequences, totaling approximately 486 minutes, from 38 participants using 70 synchronized RGB cameras and wearable motion capture; it emphasizes continuous, concurrent, and multi-object interaction in domestic settings. HuMoTo [40] contains 735 motion-capture sequences involving 63 objects and 72 articulated parts, with task designs that emphasize purposeful progression and coherent multi-object activity. A parallel line of work synthesizes extended human motion under semantic, kinematic, collective, or musical conditioning [67–69], which likewise depends on motion data that remains coherent over long durations. Together, these datasets mark an important transition

from atomic HOI clips toward scene evolution and task-level temporal structure.

Robot-learning datasets address the same challenge from the control side. Mobile ALOHA [70] combines navigation and bimanual manipulation in 290 whole-body teleoperation demonstrations across seven tasks. AgiBot World [45] provides over one million trajectories across 217 tasks, while Galaxea Open-World [47] contains approximately 100K trajectories spanning 150 tasks. RoboCOIN [71] includes roughly 180K demonstrations across 421 tasks and 15 robotic embodiments. RoboMIND 2.0 [72] further scales to approximately 310K trajectories and 739 tasks across six embodiments, with RGB-D observations, mobile manipulation, digital-twin assets, and a subset of tactile episodes. Real-to-sim reconstruction offers a complementary route to scaling such data: HSImul3R [73] reconstructs physical scenes with physics in the loop, producing simulation-ready environments in which robot trajectories can be generated at low cost. Recent methods make the computational demands of long-horizon behavior explicit. SayCan [65] connects high-level task reasoning with executable skills, MEM [66] introduces multi-scale embodied memory, and WholeBodyVLA [74] integrates vision-language-action learning with whole-body loco-manipulation. DynamicVLA [75] instead targets manipulation of moving objects, where temporal anticipation and closed-loop adaptation become necessary. EMMA [76] and HoMMI [77] further use egocentric human demonstrations to reduce the cost of collecting mobile-robot trajectories.

Human-centered and robot-centered datasets nevertheless provide different forms of supervision. Human recordings preserve natural task decomposition, flexible sub-task ordering, hesitation, and recovery, while remaining largely independent of any particular robot embodiment. Robot datasets provide executable controls and precisely aligned observations, but their trajectories are tied to specific kinematics, sensors, and teleoperation interfaces. Few existing resources combine the natural behavioral structure of human demonstrations with continuous measurement of corresponding body, object, scene, and contact states. This separation also limits diagnosis: long-horizon robot benchmarks often emphasize final task success, whereas human activity datasets commonly focus on recognition, segmentation, or motion reconstruction. Final success alone cannot reveal whether failure originated from missed contact, inaccurate object-state estimation, lost task context, incorrect sub-task ordering, or poor motion execution.

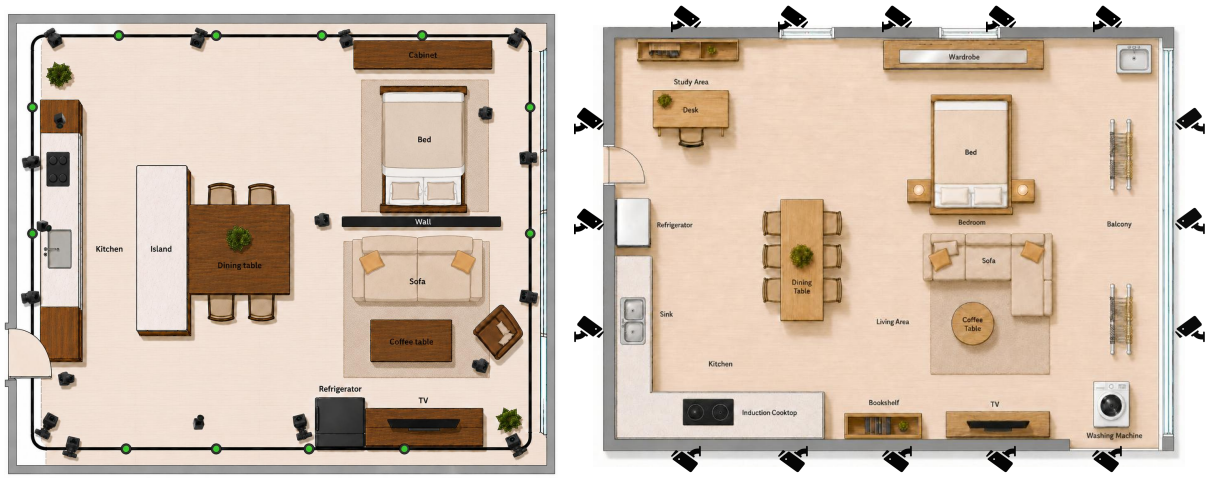
ACE-Data-0 is designed to expose these intermediate structures in goal-directed household activity. Participants receive goal-level instructions rather than fixed atomic scripts, allowing object choice, task ordering, movement paths, hesitation, and recovery to emerge naturally. Because visual observations, human and object motion, audio, and tactile signals remain synchronized throughout each task, the dataset bridges geometric HOI benchmarks, long-form egocentric video, and robot trajectory corpora, supporting analysis of not only what occurs, but also the signals, states, and interaction dynamics through which a long-horizon task is carried out.

3 Ambient Capture Engine

In this section, we present the Ambient Capture Engine (ACE) designed for collecting synchronized multi-modal data. Specifically, we first provide an overview of the system and its two scales (Section 3.1), and then describe the capture systems in detail, covering the home environments in which they are deployed and the sensors placed within them (Section 3.2).

3.1 System Overview

In Fig. 2, we provide the overall architecture of ACE. Recording everyday interaction holistically places three obligations on the capture system: it must capture every important signal an interaction may produce, keep those signals synchronized and registered in a common frame, and do all of these in environments that remain believably domestic. No single installation, however, can satisfy the first obligation at every spatial scale: resolving finger-object contact calls for close-range, densely placed sensors, whereas



(a) Home floor plan and camera setup of site I. **(b)** Home floor plan and camera setup of site II.

Figure 2 Overview of the Ambient Capture Engine. We design two complementary home scenes and capture at two spatial scales, with the equipment installed accordingly.

following locomotion across a scene calls for wide baselines and full-room coverage. Rather than compromise between the two, we build ACE at two spatial scales, table-scale and room-scale, and currently deploy it across two sites. The table-scale configuration concentrates on close-range cameras, optical motion capture, and tactile gloves, targeting fine-grained dexterous hand-object manipulation. The room-scale configuration spreads the same sensing suite over a fully furnished apartment, with wide-baseline cameras and optical motion capture covering the full activity area, targeting global motion and interactions distributed across the scene. Both systems share a common acquisition and annotation pipeline: all sensor streams are temporally synchronized, in hardware where sensors permit and in software otherwise (Section 4.1.1), registered into a shared world frame (Section 4.1.2), and enriched with annotations (Section 4.3). The result is a unified corpus in which every frame of every camera can be related to the ground-truth state of the human body, every tracked object, and the accompanying audio and tactile signals; this property underpins the benchmark in Section 5.

3.2 Capture System

Having sketched the architecture, we now zoom in on its physical form: first the environments that host the interactions, and then the sensors that observe them.

3.2.1 Environment Setup

A central design principle of ACE is to preserve the visual and physical realism of a lived-in home while permitting dense sensor coverage. The clutter, furniture, and spatial constraints that laboratory capture removes are precisely what make home-scene interaction difficult, so our environments deliberately keep them. The table-scale configuration is built around a work desk. It covers 30 square meters and is populated with over 25 interactable object instances from more than 8 categories, such as knives, bowls, and food containers. The workspace is instrumented with 8 exocentric RGB cameras rigidly mounted on stands at close range (0.3–0.5 m), providing full coverage of the manipulation area and yielding multiple synchronized views at sub-millimeter effective resolution. An optical motion capture system with 16 OptiTrack cameras, mounted on a shared truss, spans a tracking volume that covers the entire workspace. An overview of the table-scale configuration is presented in Fig. 2b. On the other hand, the room-scale configuration comprises a fully furnished apartment of approximately 200 square meters, including a kitchen with a kitchen island, a dining area, a living room, and a bedroom, populated with over 25 object instances. A truss suspended from the ceiling and spanning the apartment carries both the RGB cameras and the optical motion-capture system: 8 exocentric RGB cameras hang from the truss,

Table 2 Capture hardware of ACE for synchronized multi-modal recording. Quantities are the totals for both configurations combined; a dash marks an entry that does not apply to that device.

Device	Role	Qty	Resolution / rate	Notes
OptiTrack PrimeX 22	Optical motion capture	28	2048×1088, 60 Hz	IR tracking: 41 body markers, objects, ego rig
ZED One	Exocentric RGB capture	8	1920×1080, 30 FPS	GMSL2 to a single Jetson Orin host; shared frame trigger
GoPro	Exocentric RGB capture	8	1920×1080, 30 FPS	Rigidly mounted on stands; audio-triggered recording
ACE-Ego-Head-V02 Lite	Egocentric capture	4 cameras	$4 \times 1088 \times 1280$, 20 FPS	One headset: front/back fisheye pairs, IMU, 5 markers
Manus	Hand pose	2 gloves	60 Hz	Per-finger articulation, both hands
ACE-Sense-Glove Lite	Contact pressure	2 gloves	–	Full-palm pressure map, both hands
Jetson Orin	Recording host	1	–	Ingests all ZED One streams; NatNet-coordinated
Motive host (PC)	Mocap host, sync reference	2	–	Renders the optical clock for cross-system sync

each on an adjustable extension pole that fine-tunes its height and viewing angle, so that any point in the activity area remains visible from at least 4 views even under furniture occlusion; 12 OptiTrack cameras are mounted on the same truss, with their positions and orientations tuned so that the full apartment lies within a single tracking volume. An overview of the room-scale configuration is presented in Fig. 2a.

3.2.2 Hardware Setup

Within these environments, the sensor suite is assembled so that every facet of an interaction, from what the human sees, to how the body and objects move, to what the interaction sounds and feels like, has a dedicated sensor. Our two systems share this suite and differ only in sensor selection and placement, as summarized in Table 2.

Egocentric camera. Participants in both systems wear an ACE-Ego-Head-V02 Lite by ACE Robotics, a head-mounted egocentric capture device with four fisheye cameras facing front-left, front-right, rear-left, and rear-right, recording 1088×1280 at 20 FPS with an onboard IMU. Five OptiTrack markers are attached to the device to provide its initial placement and its real-time 6-DoF pose throughout the capture. This helps register the egocentric streams into the same world frame as every other sensor.

Exocentric camera. The table-scale configuration surrounds each workspace with 8 close-range GoPro RGB cameras (1920×1080 @ 30 FPS), rigidly mounted on stands, while the room-scale configuration covers the apartment with 8 wide-baseline ZED One RGB cameras (1920×1080 @ 30 FPS).

Human motion. Each participant wears a motion-capture suit with 41 markers attached at the joints across the whole body (including 3 markers around the head), tracked by the OptiTrack system (PrimeX 22; 12 cameras in the room-scale configuration, 16 in the table-scale configuration) at 60 Hz to yield 41-joint skeletons, complemented by articulated hand poses, acquired via Manus motion-capture gloves at 60 Hz in the room-scale configuration, and via RANSAC triangulation of 2D hand keypoints from the exocentric cameras followed by manual refinement in the table-scale configuration.

Object motion. Every interactable object is first digitized into a mesh by 3D scanning or 2D Gaussian Splatting reconstruction [78], and fitted with OptiTrack markers that are bound to this mesh in the

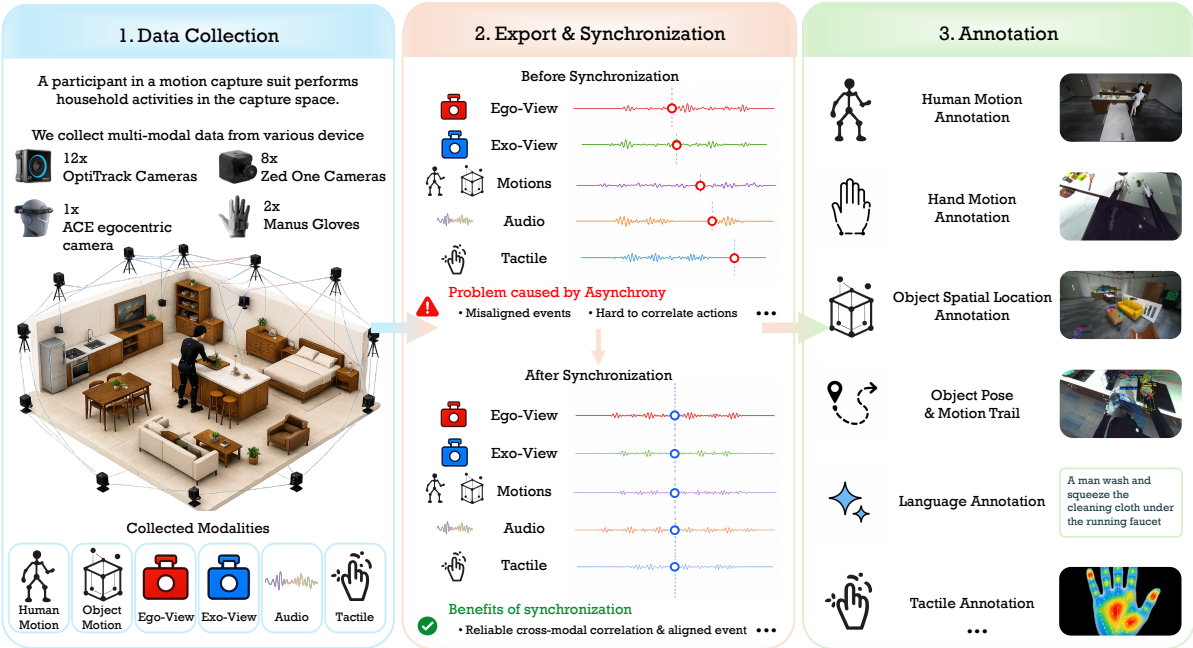


Figure 3 **Overall workflow of ACE.** Each participant wears the motion-capture suit before the multi-modal recording. Synchronization is conducted after data export to align the timeline of each modality. Finally, we annotate the collected data with rich and high-quality labels derived from the captured ground truth.

tracking system, so that the resulting 6-DoF trajectories at 60 Hz place the exact object geometry in the world frame.

Audio. We capture the audio via the GoPro exocentric cameras and the ACE-Ego-Head-V02 Lite to record the contact events, appliance operation, and ambient scene sound.

Tactile. Participants wear full-palm tactile gloves, which record contact pressure maps across the palm and fingers area.

4 ACE-Data-0

With the capture systems in place, this section presents ACE-Data-0. Specifically, Section 4.1 presents the data acquisition pipeline, covering the capture workflow, multi-modal recording, sensor synchronization, and calibration. In Section 4.2, we illustrate the data collection process, including task design, capture settings, dataset statistics, and modalities. Finally, Section 4.3 introduces the annotations provided with the dataset and the pipeline used to produce them.

4.1 Data Acquisition Pipeline

The sensors above produce a dozen heterogeneous streams, and these streams are useful only when they can be precisely related to one another in both time and space. This subsection first describes how the streams are aligned, temporally (Section 4.1.1) and spatially (Section 4.1.2), and then how capture sessions are operated in practice (Section 4.1.3 and Section 4.1.4). An overview of the workflow for ACE is presented in Fig. 3.

4.1.1 Sensor Synchronization

Temporal alignment is the foundation of multi-modal capture: cameras, motion capture, object trackers, and tactile sensors run at different rates on independent clocks, and even a few milliseconds of drift are enough to corrupt contact-level annotation, making a hand appear to close on a cup several frames before

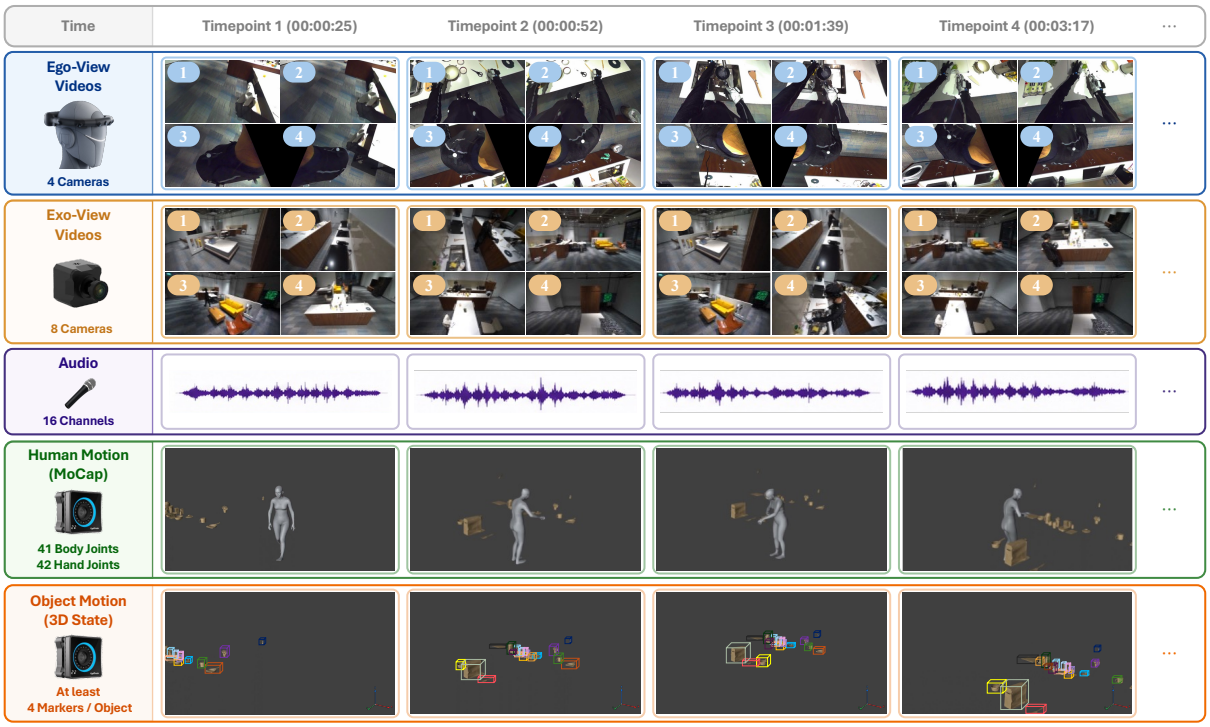


Figure 4 **Synchronized modalities captured by ACE.** ACE is capable of capturing synchronized multi-modal data, including egocentric and exocentric videos, audio, and human and object motions, etc.

the tactile stream registers the touch. We take the OptiTrack clock as the reference, since the tracking system internally synchronizes its own cameras and delivers marker positions as strictly simultaneous 60 Hz frames, and align each remaining device to it in turn.

Exocentric cameras. (1) For the room-scale configuration, the ZED One cameras are ingested by a single NVIDIA Jetson Orin host over GMSL2, whose capture cards drive all cameras from a common frame trigger, so their captures are aligned by construction. Alignment to the OptiTrack clock then proceeds in two steps. First, recording is coordinated over the local network: the camera host subscribes to the OptiTrack data stream via NatNet and starts and stops in lockstep with the motion-capture recording. This brackets the takes, but a network protocol without a shared clock cannot align individual frames. Unfortunately, the recorded timestamps cannot close this gap: they lag the true exposure moment by roughly 0.29s, and this lag also drifts slowly over time. To address this problem, we instead let the exocentric camera photograph a clock, in the form of QR codes. Specifically, the motion-capture host computer displays its own time on the monitor at nanosecond resolution. Reading this clock off a recorded frame gives the exact capture time of that frame, in motion-capture time. We sample such readings across a take and fit a line through them (a constant offset plus a slow drift). As a result, the final residuals after alignment are at the millisecond level, within a single motion-capture frame. (2) For the table-scale configuration, the GoPro cameras are synchronized with one another by aligning their audio streams.

Egocentric cameras. The four fisheye views of ACE-Ego-Head are driven by the device's onboard controller, with a measured mutual misalignment of under 2 ms. (1) For the room-scale configuration, the alignment to the OptiTrack clock reuses the optical clock above, with one complication: the egocentric cameras face the hands and the wearer's back, and never see the monitor during a take. Therefore, we start each take with a deliberate glance, in which the wearer points one camera of ACE-Ego-Head at the monitor for around 10 seconds. Reading the clock from these frames gives the offset between the clock of the egocentric camera and the OptiTrack clock. The device's shared clock then carries this offset to the other three views for overall synchronization. (2) For the table-scale configuration, the GoPro cameras and the egocentric headset read the same QR-code clock displayed on the monitor of

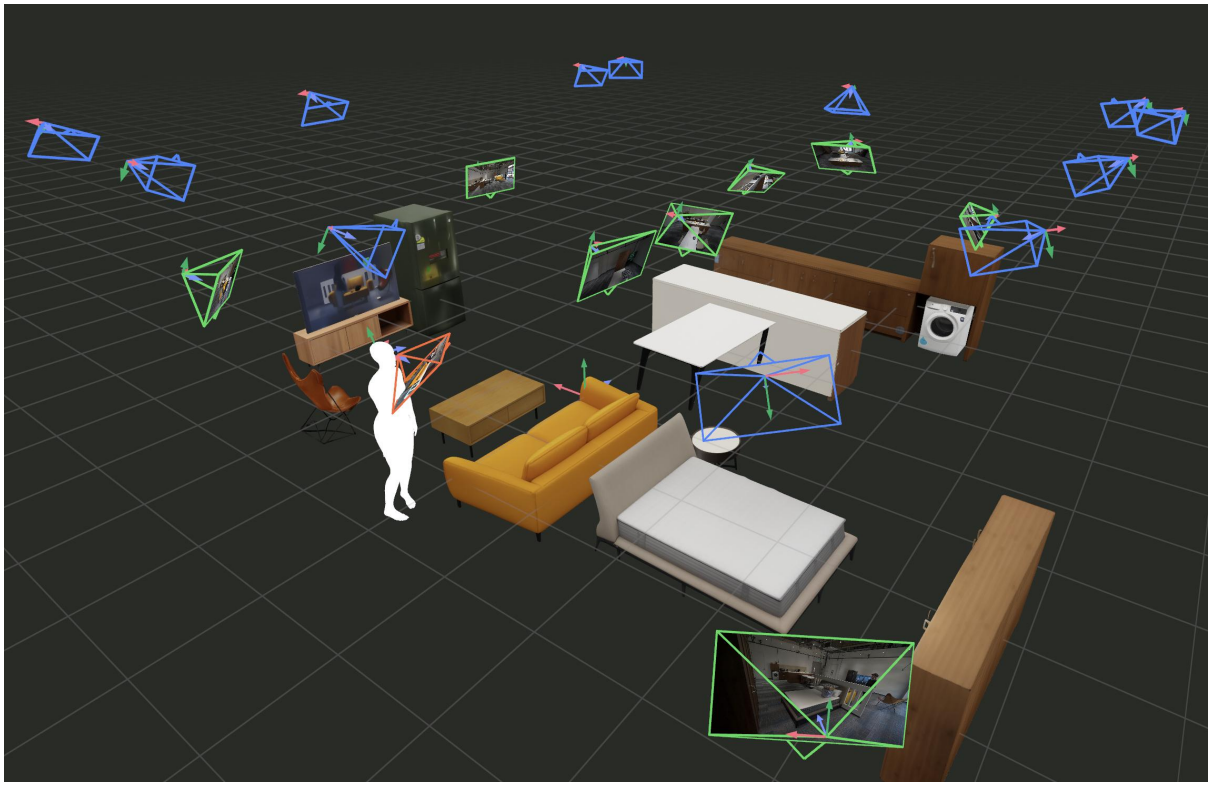


Figure 5 **Calibrated cameras in the room-scale configuration.** The blue cameras represent the OptiTrack PrimeX 22 for motion capture; the green cameras denote the ZED One for capturing exocentric RGB videos; the orange camera is one of the egocentric cameras in ACE-Ego-Head.

the motion-capture host, as in the room-scale setup; this aligns the GoPro streams, already mutually synchronized via audio, to the headset clock. The egocentric headset is then aligned to the OptiTrack system by registering its internally estimated poses (from AprilTag-based tracking) to the corresponding poses tracked via optical markers, completing the chain from the GoPro cameras, through the headset, to the OptiTrack timeline. Together, these procedures register all devices in both configurations to the OptiTrack timeline.

Other sensors. The tactile gloves are synchronized with ACE-Ego-Head using their onboard IMU signals. At the beginning of each take, the operator performs a short motion pattern while keeping the hand and head approximately rigid, producing correlated IMU signals between the glove and the egocentric headset. We align the two streams by matching this motion template, establishing temporal synchronization.

Verification and output. As an independent check, we find that the per-camera time offsets re-estimated during calibration (Section 4.1.2) are under 4 ms, confirming the alignment. All fitted offsets are folded into a per-take table that maps every camera frame, exocentric and egocentric alike, to its 60 Hz motion-capture frame. Downstream processing consumes only this table. See Fig. 4 for an example of synchronized capturing.

4.1.2 Calibration

Whereas the synchronization aligns the streams in time, calibration registers them in space, and the two camera systems pose opposite challenges: the exocentric cameras never move but barely share a view, while the egocentric cameras move every frame. The two calibration procedures are illustrated in Fig. 5 and Fig. 6.

Exocentric cameras. Standard multi-camera calibration assumes co-visibility: two cameras must observe the same target at the same time. Our sparse placement of the exocentric cameras breaks this

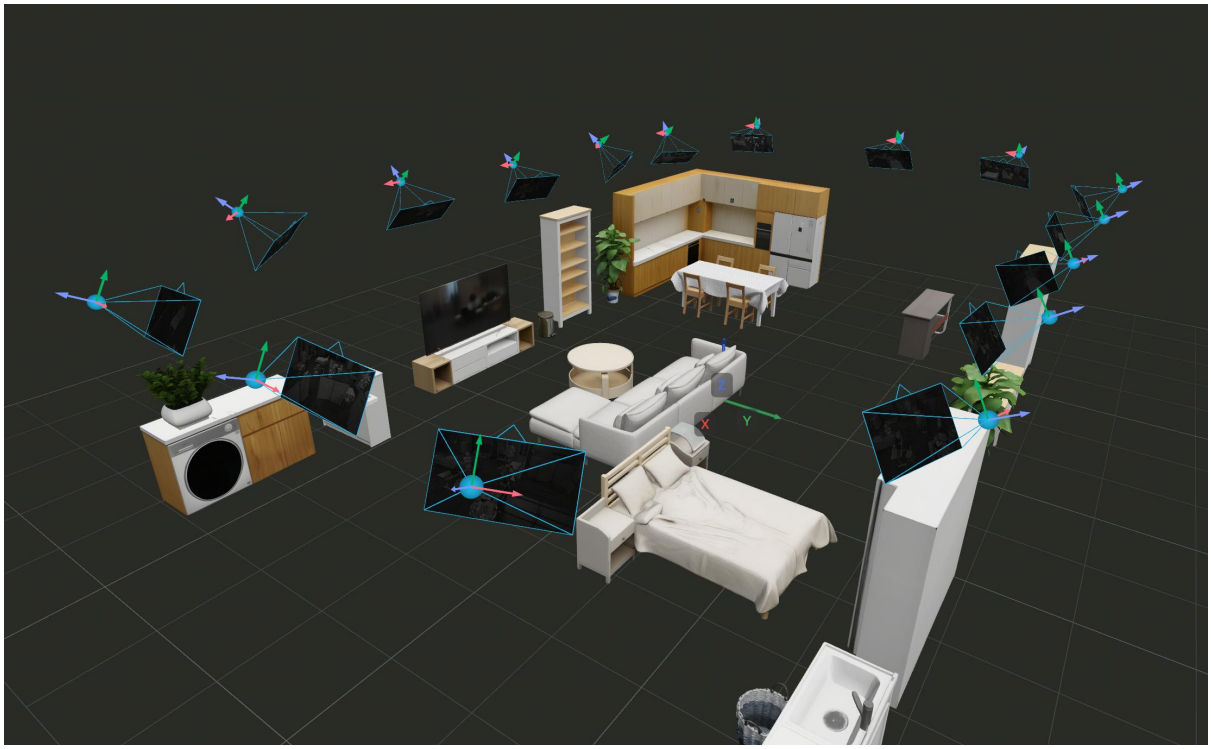


Figure 6 **Calibrated cameras in the table-scale configuration.** The blue cameras represent the OptiTrack PrimeX 22 for motion capture.

assumption, as most camera pairs share no common area. We therefore bridge the cameras through the motion-capture system instead. Specifically, we utilize an ArUco board [79] with a retroreflective marker at each corner as the calibration target. This one physical object is visible to every sensing system at once: the RGB cameras see the printed pattern of the ArUco board, while the OptiTrack infrared cameras see both that pattern and the retroreflective corner markers. The board therefore ties the RGB cameras and the motion-capture volume to a single reference without requiring any two cameras to share a view. Then, the calibration proceeds in three steps: **(1)** First, the infrared cameras estimate where the corner markers actually sit on the board. It is important to note that these markers are attached by hand, so their exact positions are not known in advance. Aggregating the estimates over thousands of board poses reduces the error in this step to the millimeter level. **(2)** Second, with their positions known, the markers alone give the board's pose in every frame, at sub-millimeter consistency. **(3)** Third, each RGB camera is fitted against this marker-anchored board trajectory. Not every observation contributes in this step: (i) frames in which the board was moving are discarded, because residual timing error grows with the motion, and (ii) the detections near the image edges are discarded, where the distortion model is least reliable. A final joint refinement then re-estimates the board pose from all cameras together and refits each camera in turn. Throughout, the tracked markers remain the dominant reference, keeping the solution anchored in the motion-capture world. On held-out frames, every camera reaches a median reprojection error below 3 px, roughly a centimeter of 3D error at typical distances.

Egocentric cameras. A moving rig is not described by one pose but by a pose for every frame. Considering our situation where the egocentric views are dominated by the wearer's own arms and torso and the front and back camera pairs share no field of view, recovering this trajectory from vision alone, as SLAM would, is unreliable. We therefore do not estimate where the cameras are; we measure them. The five markers on the ACE-Ego-Head chassis form a rigid body that OptiTrack tracks at 60 Hz. The only unknown left is the fixed transformation from each fisheye camera to this body, which is a classical hand-eye calibration problem [80]. We first solve it independently for each camera. A joint bundle adjustment [81] then refines the four transformations together, treating the board pose, shared by all cameras, and the per-camera time offsets as additional free variables. We also use only low-angular-

velocity frames for the fitting, as the timing error grows with the motion. The final median reprojection error is about 2 px. Once calibrated, producing camera poses for a take requires no images at all. OptiTrack tracks the egocentric camera's rig body at 60 Hz, while the cameras record at 20 FPS; for each egocentric frame, we interpolate the rig's tracked pose to the frame's timestamp and apply that camera's hand-eye transformation. Every pose is thus measured rather than estimated, and does not drift.

Other calibrations. The two procedures above answer where the cameras are. The remaining calibrations describe each sensor itself. For the cameras, this means intrinsics: the exocentric ones use their factory parameters, and the egocentric fisheyes are calibrated with Kalibr [82]. For the objects, this means geometry: each marker set is registered once to its scanned or 2DGS-reconstructed mesh, so that tracking the markers places the full object in the world frame. All of these are re-verified twice a day. Together with the synchronization above, we complete the unified spatio-temporal frame on which all subsequent annotation and benchmarking rest.

4.1.3 Capture Workflow

With alignment in place, every capture session follows the same five-step protocol at both sites. **(1)** Scene preparation: objects are placed at randomized yet plausible initial positions. **(2)** Participant setup: the participant puts on the motion-capture suit, the ACE-Ego-Head headset, and the gloves, then performs a short T-pose routine that registers their skeleton with the tracking system. **(3)** Task briefing: the participant receives a goal-level instruction verbally; how to achieve the goal is left entirely to them. **(4)** Recording: the take opens with the clock glance described in Section 4.1.1, after which all sensors record continuously while an operator monitors stream health on a live dashboard. **(5)** Post-checks: synchronization and tracking quality are verified after each take, and failed takes are flagged for re-capture.

4.1.4 Multi-modal Recording

During step (4) of the capture workflow, each system simultaneously acquires: (i) the exocentric RGB streams; (ii) 4 egocentric fisheye streams with IMU; (iii) full-body motion and hand poses at 60 Hz; (iv) 6-DoF poses of all tracked objects at 60 Hz; and (v) tactile signals. All streams are timestamped against the common clock established in Section 4.1.1. A one-hour session produces approximately 1 TB of raw data.

4.2 Data Collection

The preceding subsections established the recording machinery; the question that remains is what to record with it. We describe the task design (Section 4.2.1), the capture settings (Section 4.2.2), the resulting dataset statistics (Section 4.2.3), and the modalities included in each released sequence (Section 4.2.4).

4.2.1 Task Design

The long-horizon character of ACE-Data-0 originates in its task design. By a *long-horizon* task we mean an activity that unfolds over minutes rather than seconds and decomposes into a sequence of sub-actions whose order is constrained by the goal rather than fixed in advance. For example, consider the task of "prepare a cup of tea and serve it at the table". It involves a chain of human-object interactions: the participant walks to the cupboard, opens it, takes out a cup, fills the kettle, waits for the water to boil, retrieves a tea bag, pours, stirs, carries the cup across the room, and clears space before setting it down. This single take may thus contain planning, locomotion, and fine-grained manipulations.

However, most HOI datasets instead prescribe atomic actions [12, 15, 16, 18], capturing a single grasp or handover in isolation. To address this gap, we prescribe household goals and let the actions emerge: how to reach the goal is left to the participant. Different participants order the sub-tasks differently,

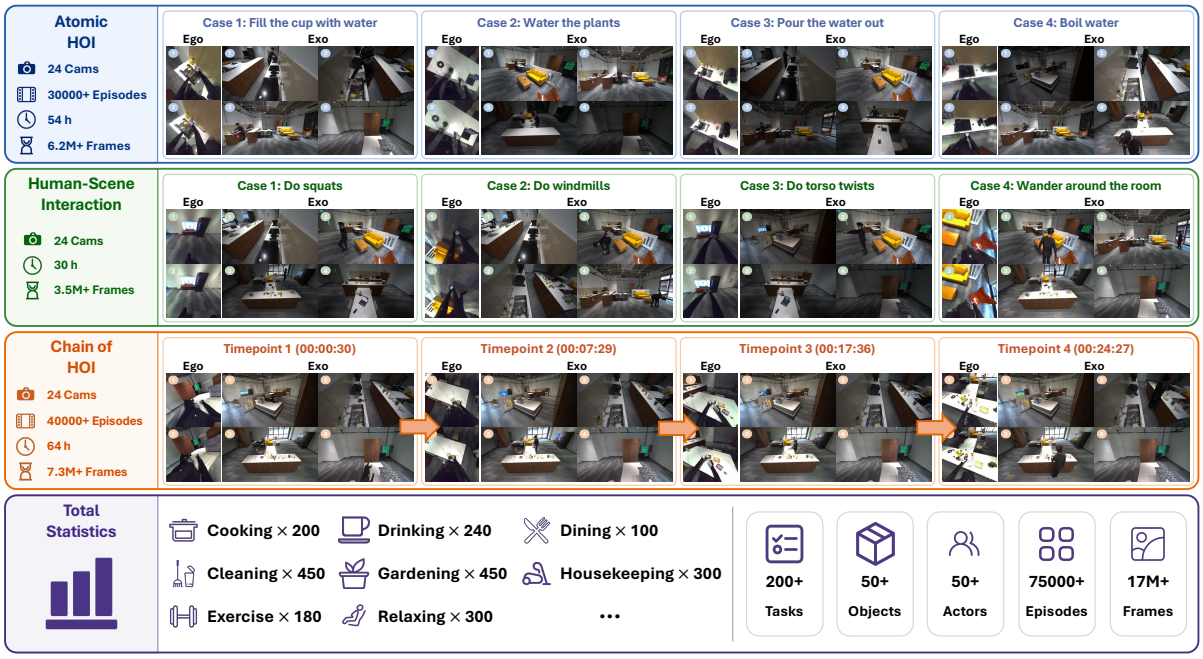


Figure 7 **Examples of different task categories captured in ACE-Data-0.**

grasp differently, and reach for different objects, so the variability of real behavior enters the data by itself.

Specifically, we record three types of takes:

- **Atomic HOI tasks** contain one to three household tasks each, drawn from more than 15 types of household activities, such as pouring water, drinking, making tea, watering plants, chopping vegetables, cooking, and tidying up; each take lasts roughly three minutes. Each of them provides clean, self-contained instances of everyday manipulation, the unit from which skills are most readily learned.
- **Chains of HOI tasks** combine the full range of short tasks into one continuous activity of roughly twenty to thirty minutes. Sub-tasks interleave freely, yet every take ends with the scene tidied back into order, tracing a full cycle of household activity. Compared to the atomic HOI tasks, they exercise long-horizon planning, state tracking, and memory at a further level of complexity.
- **HSI tasks** involve almost no objects, focusing instead on interactions with scene components such as tables, chairs, and sofas. Such takes record whole-body motion, such as walking and exercising, and human-scene contact, such as sitting, lying, and leaning, in about five minutes per take. These movements complete the range of behaviors a humanoid must master.

To decide what these recordings should contain, we surveyed the tasks that arise in everyday home life and designed the atomic HOI, chain-of-HOI, and HSI tasks accordingly. We present examples of different task categories in Fig. 7.

4.2.2 Capture Settings

To perform these tasks, we recruited 50 participants. Each participant completes all the designed tasks across a 2-day session. Diversity enters the data from two directions. On the human side, the goal-level instructions leave the behavior open: participants differ in where they start and end, the route they take in between, the interactions they choose, and how long each step takes. On the object side, takes vary in which categories appear and in what number, where the objects are placed, where they begin and end, and along what trajectories and in what manner they are moved.

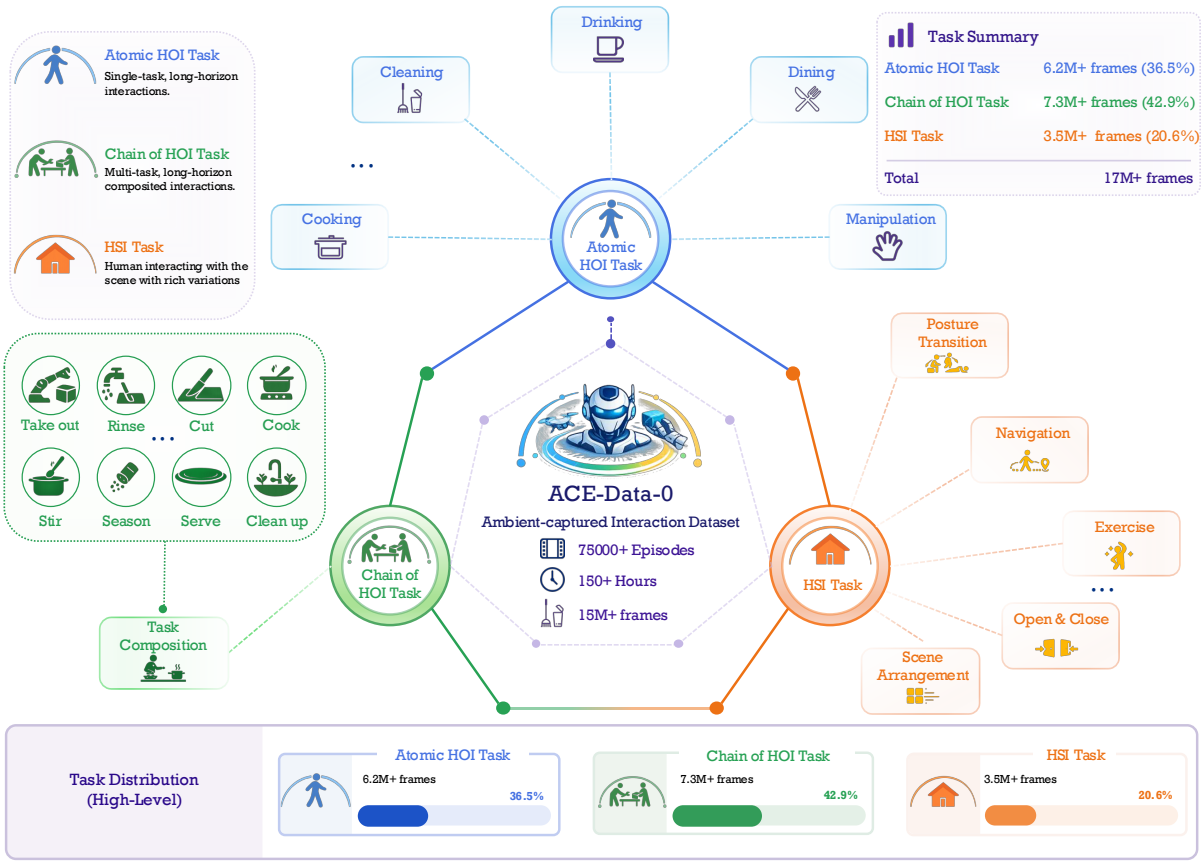


Figure 8 **Overview statistics of ACE-Data-0.** ACE-Data-0 covers three types of interaction tasks: atomic HOI tasks, chains of HOI tasks, and HSI tasks.

4.2.3 Dataset Statistics

This collection effort yields more than 150 hours of synchronized multi-modal capture, comprising over 17M frames across more than 75,000 episodes. We define an *episode* as a contiguous segment of interaction that realizes one meaningful sub-goal, the smallest unit that remains semantically self-contained when used as a training example. Episodes are counted within takes rather than recorded in isolation, so the surrounding context, namely the actions that precede and follow each segment, is preserved in the same stream. In ACE-Data-0, even the shortest takes run minutes rather than seconds, roughly an order of magnitude longer than typical HOI clips [12, 15, 16, 18]. Fig. 8 reports the distributions of task categories, take lengths, object categories, etc.

4.2.4 Modalities

A released take contains more than the raw streams of Section 4.1.4: it also carries everything the pipeline of Section 4.1 derives from them, in a common frame and on a common timeline. Concretely, each take bundles:

- **Egocentric video**: the four fisheye views, their IMU readings, and per-frame 6-DoF headset poses from the tracked rig;
- **Exocentric video**: all views, each with intrinsics and its pose in the world frame;
- **Human motion**: 41-joint body skeletons with articulated hand poses and converted SMPL-X motion parameters [83];
- **Object motion**: per-object 6-DoF trajectories, with scanned or 2DGS-reconstructed meshes for more than 50 instances;

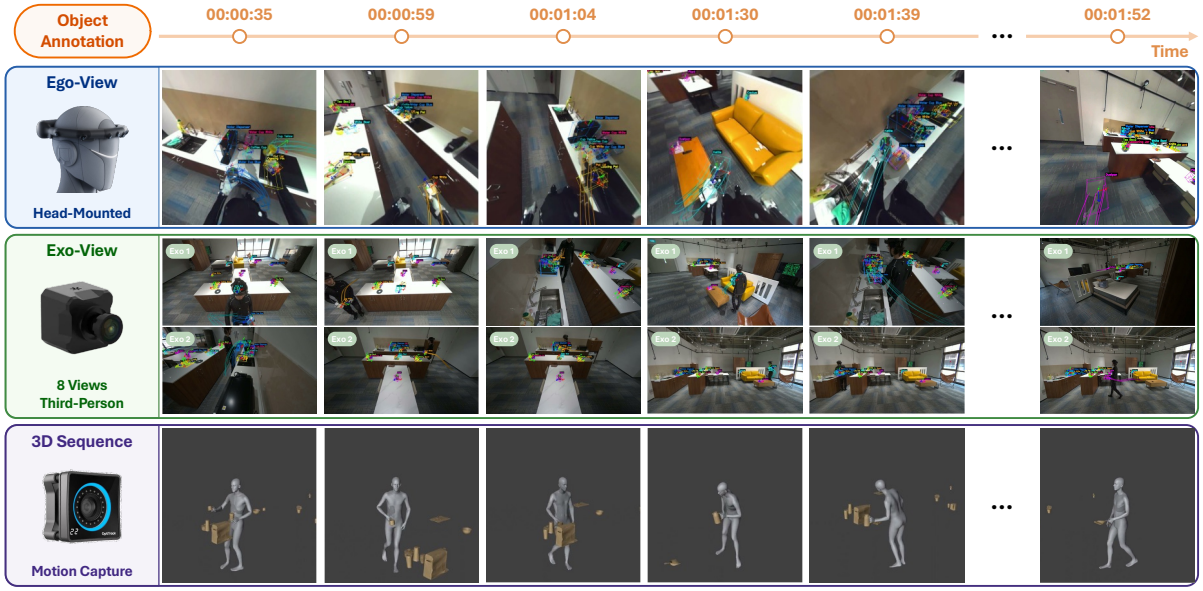


Figure 9 **Examples of data annotation for captured objects.** ACE-Data-0 provides rich annotations for the captured objects, including name tags, 6-DoF poses, bounding boxes, and motion trails.

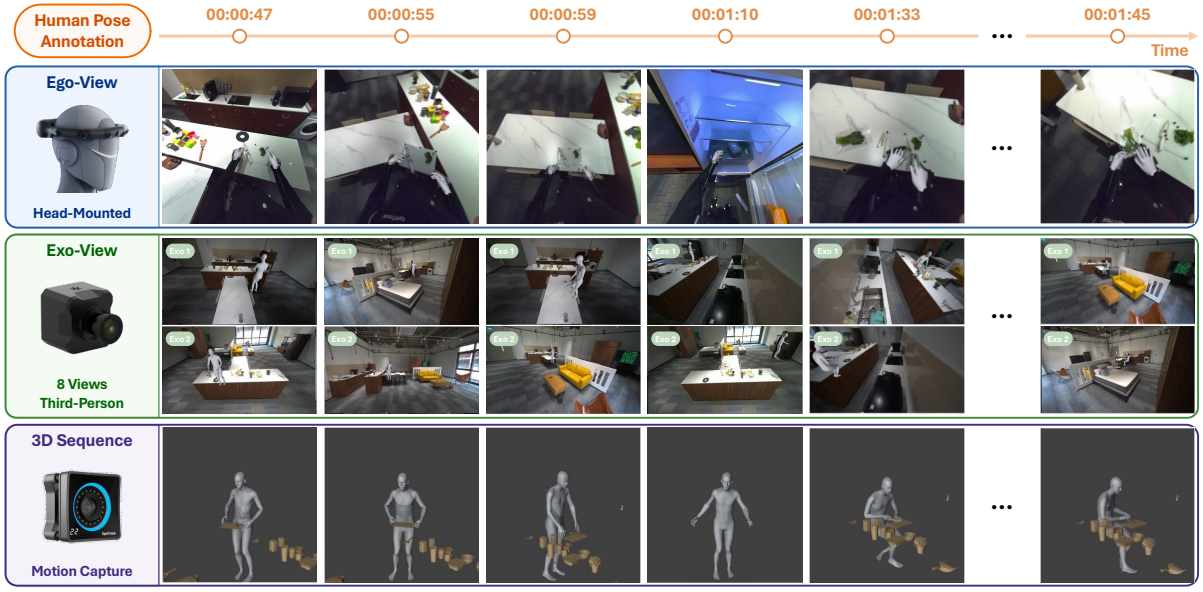


Figure 10 **Examples of data annotation for the captured human body and hands.** ACE-Data-0 provides ground-truth poses for both the human body and hands. We also project the poses to both egocentric and exocentric videos for better embodied AI learning with visual inputs.

- **Audio**: synchronized multi-source audio recorded from GoPro exocentric cameras and the egocentric headset;
- **Tactile**: hand-shaped pressure grids remapped from raw glove sensors, with calibrated normalization and baseline correction.

4.3 Annotation

The raw signals described so far become substantially more useful once enriched with annotations. We describe what annotations ACE-Data-0 provides (Section 4.3.1) and how they are produced at scale (Section 4.3.2).

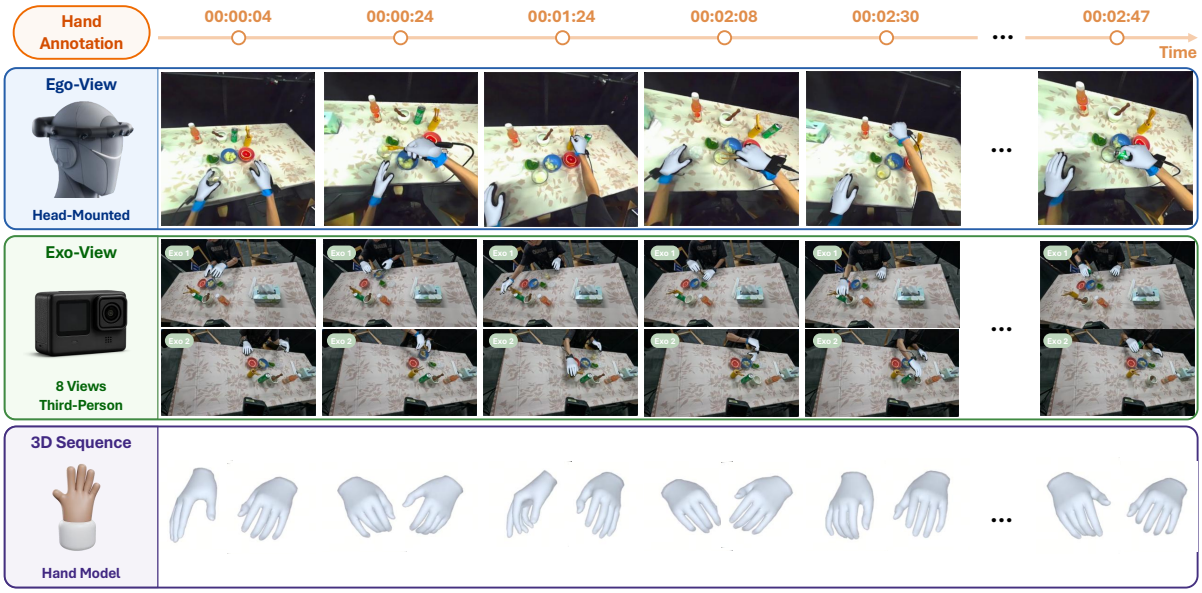


Figure 11 **Examples of data annotation for captured hands.** ACE-Data-0 provides ground-truth poses for human hands. We also project the poses to both egocentric and exocentric videos for better embodied AI learning with visual inputs.

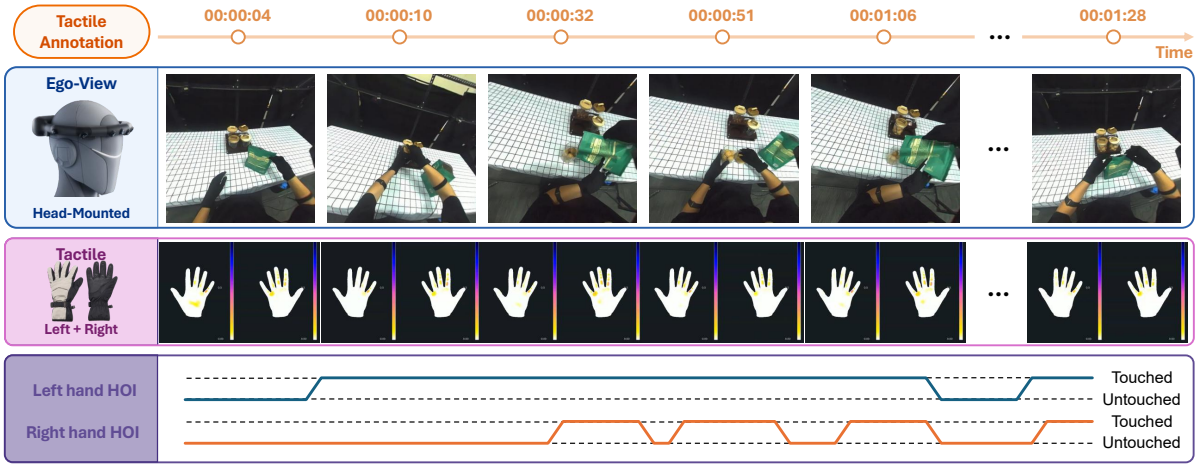


Figure 12 **Examples of data annotation for tactile signals.** ACE-Data-0 provides ground-truth full-hand grasp pressure values, resolving interaction events that remain ambiguous under visual occlusion.

4.3.1 Annotation Types

ACE-Data-0 provides five types of annotation, which jointly characterize every take: the objects involved in the interaction, the configuration of the actor's body, the articulation of the hands during manipulation, the physical contact established between hand and object, and the acoustic and semantic description of the activity itself. Specifically, the object annotations (Fig. 9) comprise a category label, a bounding box, and a per-frame 6-DoF pose for every tracked object, together with the motion trail traced by the object over the course of the take. Human annotations (Fig. 10) provide full body pose throughout each take, and hand annotations (Fig. 11) refine this to articulated finger configurations during periods of dexterous manipulation. Tactile annotations (Fig. 12) register the timing and spatial distribution of contact, resolving interaction events that remain ambiguous under visual occlusion. Audio and language annotations (Fig. 13) pair the recorded audio streams with natural-language descriptions that identify the sound events produced by the interaction, such as an object set down on a surface or a container being opened, and state the goal of each take alongside the sequence of sub-goals through which it is achieved.

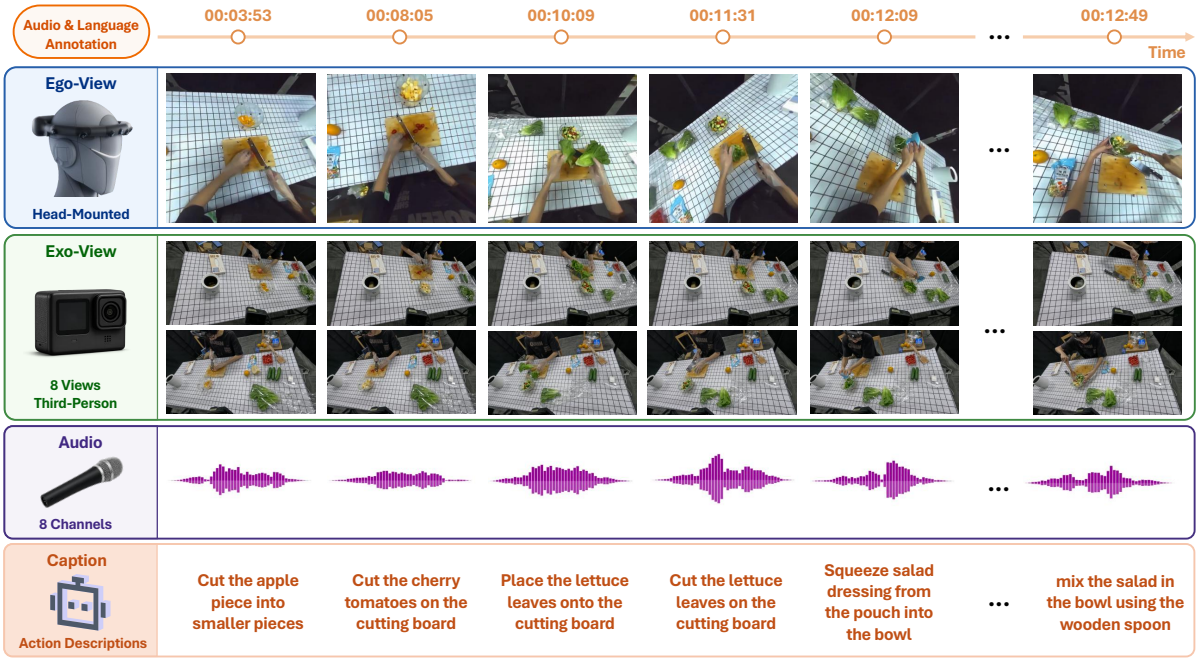


Figure 13 **Examples of data annotation for audio and language description.** ACE-Data-0 provides natural language descriptions aligned with the timeline of measured activities.

Calibration and synchronization. Every take ships with per-camera intrinsics, camera poses in the shared world frame, and the timeline that aligns all sensor streams. These are the outputs of Section 4.1, released as data: with them, any tracked 3D point can be projected onto any pixel of any view, and any two streams can be paired at any instant. Users can therefore combine the modalities freely for training, pairing any subset of them as inputs and supervision, without rerunning any part of our pipeline.

Human poses. The captured body and hand poses are reprojected onto every frame of every camera, giving per-view 2D mesh overlays consistent with the 3D capture. Because these projections come from measured 3D states rather than image-based detectors, they remain correct where detectors typically fail: under furniture occlusion, extreme viewpoints, and motion blur. Each frame thus carries pixel-aligned 2D poses in every view and metric 3D poses in the world frame.

Tactile labels. Each frame carries the tactile reading from the tactile glove, temporally aligned with the visual streams and the object label currently in use. Contact thus can be detected rather than inferred: the signal comes directly from the sensor surface, not from appearance or from proximity between reconstructed geometry. These readings mark when and with what an interaction happens, the anchor from which most HOI tasks start.

Object annotations. Every object instance carries its scanned or 2DGS-reconstructed mesh, per-frame 6-DoF poses, 2D/3D bounding boxes in all views, and its motion trail over the take. Mesh and pose together place the exact object geometry in the scene at every instant; the boxes and trails are their projections into each camera. The full history of an object, where it sat, when it moved, and where it ended, can therefore be queried at any point of a take.

Audio descriptions. Each take carries multi-channel audio recorded by the head-mounted egocentric capture rig and the GoPro cameras, sharing the synchronization clock of the visual streams and therefore aligned to the same timeline as the pose, object, and tactile annotations. Two classes of sound are present: those produced by the interaction itself, such as an object set down on a surface or liquid poured into a cup, and the ambient acoustics of the domestic environment, including appliance noise, footsteps, and room reverberation. Because the audio is captured from the actor's own viewpoint, the acoustic perspective moves with the participant, so a sound event varies in loudness and spatial character with proximity.

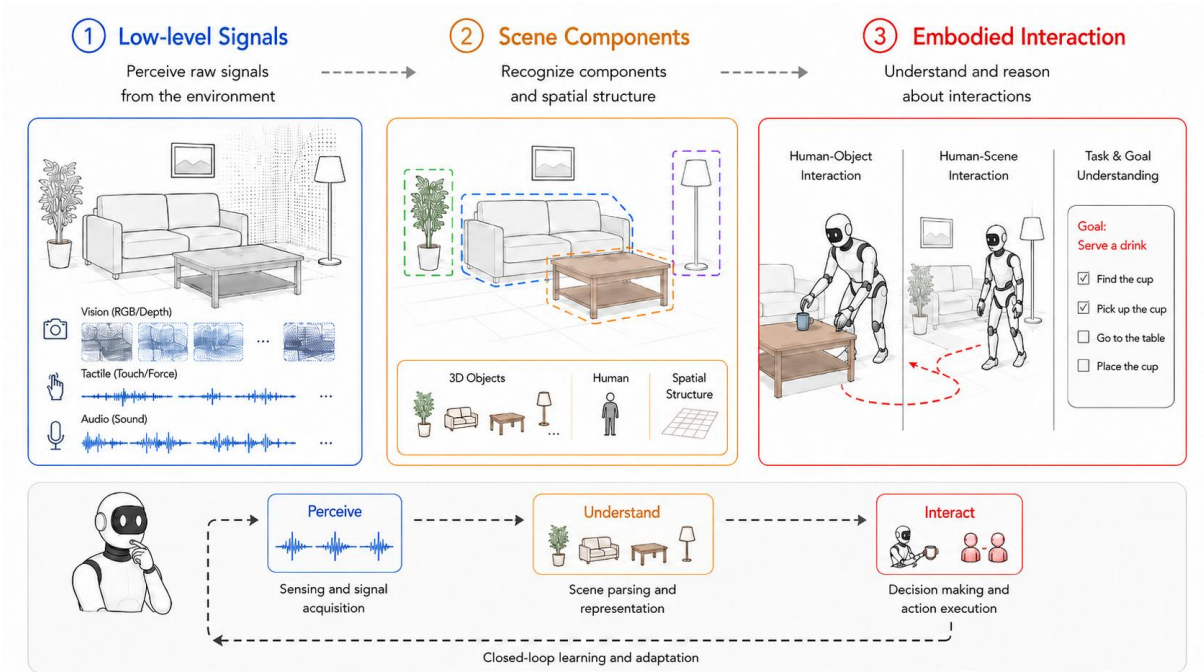


Figure 14 **Three hierarchical benchmark levels.** Our benchmarked levels, *i.e.*, low-level signal inference, scene component recovery, and interaction estimation, exactly mimic how human beings and embodied AI would perceive real-world environments.

Textual descriptions. Gemini-3.1-pro-preview [84] watches the ego-view video and describes each time span in natural language: what the person is doing, and what happens in the scene. The descriptions give each take a searchable storyline and connect the physical record to language, in the form that vision-language and vision-language-action models consume directly.

4.3.2 Annotation Pipeline

Producing annotations of this breadth would ordinarily demand prohibitive manual effort. ACE avoids most of it, because among our five annotation types, all but the textual descriptions are *measured* rather than estimated. Human and object states are metrically tracked, every camera is calibrated, and contact is directly sensed. The pose reprojections, bounding boxes, motion trails, and contact events therefore follow from the recorded states by projection and tactile sensing, with no estimation model in the loop. The textual descriptions are the one generated type: Gemini produces them from the ego-view videos, segment by segment. Finally, human annotators check the auto-generated labels and correct the descriptions by hand.

5 Benchmark

Having described how ACE-Data-0 is captured and annotated, we now use it to evaluate existing methods, with a diagnostic goal: to expose where current approaches break on long-horizon home-scene data, and to indicate where solutions might lie. The benchmark comprises three levels, advancing from *signals*, to *components*, to *interactions*, with each level building upon the one beneath it: *low-level signal inference* operates directly on the raw sensory streams, *i.e.*, predicting tactile signals from video (Section 5.1); *scene component recovery* assembles these signals into the 3D human components (Section 5.2); and *interaction estimation* recovers the hand motion through which the human engages objects, from egocentric and exocentric views (Section 5.3). This progression mirrors the perceptual capabilities an embodied agent must chain together when acting in a home: sensing contact, estimating scene state, and mastering hand-object coordination (see Fig. 14). We hold out 10 hours of capture as a test set

Table 3 **Tactile estimation from ego-view on ACE-Data-0**.

Method	Temp Acc. ↑	C-IoU ↑	V-IoU ↑	CoP ↓
PressureVision [85]	0.0093	0.0007	0.0000	10.9807
EgoPressureDiff [86]	0.2912	0.0197	0.0025	8.5152
TouchAnything [87]	**0.7095**	**0.1646**	**0.1357**	**6.5846**

for all benchmark evaluations in this report. Unless otherwise noted, baselines are evaluated with their officially released pre-trained checkpoints, for fair comparison.

5.1 Low-level Signals

We begin with the raw sensory signals that interactions produce, and among them, touch holds a special position. Vision tells an observer where things are; touch tells the actor whether a grasp has succeeded, how firmly to hold, and when an object begins to slip. Robots need this signal as much as humans do, yet contact sensing remains scarce in practice: tactile hardware is expensive and fragile, and absent from most deployed platforms, while cameras are everywhere. Learning to read touch out of video would therefore turn the most abundant sensor into a substitute for the scarcest one. ACE-Data-0 makes this mapping learnable: through the synchronization of Section 4.1.1, every video frame is paired with the tactile reading recorded at the same instant.

5.1.1 Tactile from Vision

Visual observations provide only indirect evidence of physical contact. This ambiguity becomes particularly severe in egocentric views, where the manipulating hand frequently occludes the fingertips and the contacted object regions precisely when pressure is applied.

Task. Given an egocentric video of a human-object interaction, the goal is to predict full-hand grasp pressure at each moment. We evaluate the predicted pressure distributions against synchronized measurements captured by the tactile glove.

Metrics. We report four complementary metrics. Temporal accuracy measures frame-wise contact-state prediction over the interaction sequence. Contact IoU (C-IoU) evaluates the spatial overlap between thresholded predicted and ground-truth contact regions, while volumetric IoU (V-IoU) additionally accounts for pressure magnitude through min-max aggregation. Center-of-Pressure (CoP) error measures the spatial distance between predicted and ground-truth pressure centroids over the fingertips and palm, with lower values indicating more accurate contact localization.

Baselines. We evaluate three baseline methods using their released pre-trained weights. PressureVision [85] directly regresses pixel-aligned pressure maps from RGB observations using a convolutional encoder-decoder. EgoPressureDiff [86] formulates pressure estimation as conditional diffusion and adapts a pre-trained video diffusion backbone. TouchAnything [87] learns a general vision-to-touch representation using cross-view fusion and view-dropout training.

Results. Table 3 reports the results on the close-range table-scale recordings. PressureVision provides little meaningful contact prediction, producing nearly zero overlap under both C-IoU and V-IoU and the largest CoP error. EgoPressureDiff substantially improves temporal contact recognition and pressure localization, but its spatial overlap remains limited, indicating that detecting when contact occurs is considerably easier than recovering where pressure is distributed across the hand. In contrast, TouchAnything consistently achieves the strongest performance across all four metrics. Its advantage is especially pronounced in temporal accuracy and spatial overlap, suggesting that broader visual-tactile training and cross-view modeling improve transfer to the diverse interactions in ACE-Data-0. Nevertheless, its absolute C-IoU and V-IoU remain modest, and the remaining CoP error indicates that accurate pressure localization is still challenging. In particular, a model may correctly identify the presence of contact while failing to recover its precise position and intensity under hand-object occlusion.

Table 4 **Human motion estimation on ACE-Data-0.** All metrics in mm.

Method	PA-MPJPE ↓	PA-PVE ↓	MPJPE ↓	PVE ↓	WA-MPJPE ↓
— Single-view exocentric, per-frame —					
Multi-HMR [89]	110.5	–	115.3	–	–
Multi-HMR2 [90]	85.5	–	92.8	–	–
SAM-3D-Body [91]	71.0	–	76.4	–	–
PyMAF-X [92]	94.7	177.5	99.9	177.0	–
PARE [93]	98.7	132.3	102.0	134.8	–
CameraHMR [94]	78.7	98.7	85.7	105.2	–
OSX [95]	59.2	73.7	62.5	76.6	–
— Single-view exocentric, temporal —					
SMPLer-X [96]	57.0	70.2	59.8	71.7	–
SMPLest-X [97]	**55.7**	**65.8**	**58.8**	**67.6**	–
Humans-in-4D [98]	77.1	94.4	79.6	96.4	–
GVHMR [99]	88.0	104.1	96.3	112.0	217.1
WHAM [100]	131.1	160.2	144.0	166.4	243.1
EasyMoCap [101, 102]	103.4	155.8	106.0	156.1	256.2
— Single-view exocentric, scene-aware —					
Phy-SIC [55]	64.3	78.9	71.3	84.6	–
UniSH [103]	80.0	111.2	82.8	114.0	196.3
JOSH [104]	64.0	89.2	69.9	95.0	245.1
Human3R [105]	60.1	75.4	63.5	80.6	**180.2**
— Multi-view exocentric —					
MAMMA [106]	70.1	86.8	74.8	89.6	230.8
U-HMR [107]	134.4	176.0	150.2	179.2	–
HSfM [108]	92.8	133.8	93.8	134.8	–
— Egocentric —					
EgoEgo [109]	159.6	245.3	163.4	263.9	306.2
EgoAllo [110]	131.7	196.8	147.9	220.3	252.2

Overall, the results reveal a substantial generalization gap for existing tactile-estimation models and establish ego-view pressure reconstruction as a challenging benchmark. Robust performance requires not only recognizing contact events, but also resolving fine-grained pressure distributions from visually ambiguous and frequently occluded interactions.

5.2 Scene Components

Where the previous track asks what an interaction feels like, this one asks where the components of the scene are and how they move over time. Between raw sensor signals and any understanding of an interaction lie the 3D states of the scene: the pose of the body, the articulation of the hands, and the 6-DoF pose of every object. These states are what most vision pipelines estimate first and what every downstream module consumes.

In home scenes, however, their reliability remains largely unquantified. In-the-wild footage offers no ground truth to measure against, datasets with metric ground truth are confined to laboratory settings, and datasets recorded in real environments are pseudo-labeled by the very methods under evaluation. ACE-Data-0 removes this obstacle by providing metric ground truth in furnished domestic scenes, allowing existing estimators to be evaluated under the conditions in which they are actually deployed. We therefore benchmark human motion estimation in this track. Hand articulation is treated separately in the following track, where dexterous manipulation provides the setting in which it matters most. Object pose estimation is left to future work: too few applicable methods [88] exist to support a meaningful comparison, though the released ground truth supports this evaluation directly.

5.2.1 Human Motion Estimation

Human motion estimation in home environments remains challenging for several reasons. Furniture often blocks the lower body for long parts of a sequence. Common household actions, such as crouching in front of a cabinet, reaching above the head, or lying on a sofa, also differ from the mostly upright poses found in many training datasets. In addition, each take can last several minutes, allowing small frame-level errors to build up and cause large drift in the estimated global trajectory [33, 55, 99, 100]. Our dataset captures these challenges together and provides metric ground-truth motion for the full sequence.

Task. Given the visual observations from a take, the goal is to estimate the articulated body pose over time. Per-frame methods process individual images, while temporal methods use the full video sequence. Methods with world-frame outputs are also expected to recover the person's global trajectory through the scene.

We evaluate all predictions against motion-captured ground truth. The evaluation covers three input settings: multi-view exocentric video, single-view exocentric video, and egocentric video. It also includes three method families: per-frame, temporal, and scene-aware methods. Since all settings use the same ground truth, the results allow us to compare the effects of viewpoint, temporal information, and scene context directly.

Metric. We report five metrics, ordered from local to global. PA-MPJPE and PA-PVE [111] evaluate the articulated pose and the recovered body surface after Procrustes alignment, independent of global position and orientation. MPJPE and PVE [83, 112] evaluate the same quantities with the global orientation retained. WA-MPJPE [99, 100] evaluates the global trajectory in the world frame, after a single alignment of the whole path.

Baselines. We evaluate 22 methods, covering every major family of human motion estimation. For single-view exocentric input, these include per-frame approaches (Multi-HMR [89], Multi-HMR2 [90], SAM-3D-Body [91], PyMAF-X [92], PARE [93], CameraHMR [94], and OSX [95]), video-based approaches (SMPLer-X [96], SMPLest-X [97], Humans-in-4D [98], GVHMR [99], WHAM [100], and EasyMoCap [102]), and their scene-aware counterparts, Phy-SIC [55] for the former and UniSH [103], JOSH [104], and Human3R [105] for the latter. For multi-view exocentric input, we evaluate MAMMA [106], U-HMR [107], and HSfM [108]. For egocentric input, we evaluate EgoEgo [109] and EgoAllo [110]. All methods run with their released pre-trained weights. To our knowledge, this is the broadest evaluation of human motion estimation conducted in real home scenes to date.

Results. We evaluate primarily on the room-scale recordings, which combine locomotion, furniture occlusion, and minutes of continuous motion. Quantitative results are reported in Table 4, and three findings stand out.

First, the results show a clear gap between local pose estimation and global trajectory recovery. Several methods achieve strong results on the Procrustes-aligned metrics, with accuracy close to that reported on standard benchmarks. This suggests that they can still recover articulated body poses under household occlusion and uncommon postures. However, their world-frame trajectory errors remain much higher. The ranking on local pose metrics also differs from that on global trajectory metrics. In other words, estimating the body pose correctly in each frame does not guarantee an accurate motion path over the full sequence.

Second, the comparison between scene-aware and temporal methods further shows where scene information is most useful. Scene-aware methods achieve lower trajectory errors, while their Procrustes-aligned results remain similar to those of temporal methods. This suggests that scene context mainly helps determine the body's position in the room, rather than improving the relative joint configuration. Scene geometry provides useful constraints on where a person can stand or move, but gives less direct information about the exact body pose.

Third, the results also show a strong effect of viewpoint. Egocentric methods perform worse across most metrics because much of the body is outside the camera's field of view and must be inferred from head motion. Multi-view methods, however, do not outperform the strongest single-view methods in

Table 5 **Hand motion estimation from ego-view on ACE-Data-0**.

Method	PA-MPJPE ↓	MPJPE ↓	F@5 ↑	F@15 ↑	AUC_J ↑	Traj. err. ↓
WildHands [113]	**11.2**	**12.6**	**0.175**	**0.774**	**0.776**	–
Dyn-HaMR [114]	18.9	21.1	0.042	0.413	0.624	**98.2**
HaWoR [52]	13.8	17.4	0.130	0.666	0.729	102.1

this evaluation. We do not view this as evidence that multiple views are less useful. Instead, it likely reflects the strong performance of recent single-view whole-body and scene-aware methods, as well as the limited range of available multi-view baselines.

5.3 Embodied Interaction

Knowing the position of the person is not enough to describe how an interaction is carried out. Much of this information is contained in the hand motion, including how the hand approaches an object, forms a grasp, adjusts its position, and releases it. Hand motion is also especially relevant to robot learning because a hand trajectory can be retargeted to a robotic gripper more directly than raw visual observations.

We therefore evaluate how accurately current methods [59–61] recover hand motion from video. In ACE-Data-0, each interaction is recorded at the same time from both egocentric and exocentric viewpoints. This allows us to compare the two settings under the same conditions, using the same takes and the same ground truth while changing only the viewpoint.

5.3.1 HOI from Ego-View

Egocentric video provides the viewpoint that is most similar to what a future robot may observe. The camera stays close to the hands and often captures fine details of the interaction. However, this viewpoint also introduces several practical challenges. The camera moves with the head, the wide-angle lens distorts the image near the boundary, and the hands often appear in these distorted regions. Large head or body movements may also move one or both hands outside the field of view.

Task. Given the egocentric video of a take, a method estimates the 3D motion of both hands as MANO [49] sequences. This includes the articulated finger pose in each frame and, when supported by the method, the hand trajectory in a world coordinate frame. We obtain the ground truth from the Manus motion-capture gloves for the room-scale recordings; for the table-scale recordings, we apply RANSAC triangulation to 2D hand keypoints from the eight synchronized exocentric cameras, followed by manual refinement. Joints that are visible in too few views are masked during evaluation. The method is therefore not penalized on frames for which reliable ground truth cannot be obtained.

Metrics. We report five metrics for hand pose and one metric for the global trajectory. The same definitions and units are used for both egocentric and exocentric evaluation. PA-MPJPE applies a separate Procrustes alignment to each predicted hand using rotation, translation, and scale. It therefore mainly measures errors in finger articulation. MPJPE uses only rotation and translation while keeping the scale fixed, and thus also reflects errors in metric hand size. F@5 and F@15 report the fractions of joints with errors below 5 and 15 mm, respectively. AUC_J is the normalized area under the PCK curve over thresholds from 0 to 50 mm. For methods that predict a persistent world coordinate frame, we also report the world-frame trajectory error. We align the predicted and ground-truth wrist trajectories using one similarity transform for the entire clip and compute the mean residual error. Unlike per-frame alignment, this metric captures drift and inconsistent scale across the sequence.

Baselines. We evaluate three methods using their released pre-trained weights. WildHands [113] is a per-frame regressor and is applied independently to each egocentric frame. Dyn-HaMR [114] and HaWoR [52] process videos and recover the hands in a world coordinate frame by combining hand estimation with camera-motion estimation.

Table 6 **Hand motion estimation from exo-view on ACE-Data-0.**

Method	PA-MPJPE ↓	MPJPE ↓	F@5 ↑	F@15 ↑	AUC_J ↑	Traj. err. ↓
HORT [54]	10.8	12.3	0.276	0.776	0.784	–
HaMeR [51]	9.6	10.4	0.281	0.848	0.812	–
HaPTIC [53]	10.0	10.7	0.247	0.838	0.804	**63.0**
WiLoR [115]	**9.1**	**9.9**	**0.313**	**0.854**	**0.819**	–
OmniHands [116]	10.7	11.5	0.211	0.814	0.791	–

Results. Table 5 shows a clear difference between frame-level hand pose estimation and world-space hand motion recovery. The per-frame method achieves the strongest articulation accuracy, while the two video-based methods perform less well on the local pose metrics. One possible reason is that the video-based methods estimate camera motion and hand pose together. Errors in camera motion can therefore affect the recovered hand pose, especially in our recordings where head motion is often large. The difference becomes more evident for global trajectory estimation. Both world-space methods produce trajectory errors that are much larger than their local joint errors. This suggests that the main challenge in egocentric hand reconstruction is not estimating finger articulation in individual frames, but maintaining a stable hand trajectory in the world coordinate system. Together with the exocentric results below, this points to egomotion estimation as a major source of error.

5.3.2 HOI from Exo-View

Exocentric video presents a different set of advantages and challenges. The camera remains fixed, and the person and surrounding scene usually stay within the image. This provides a stable reference frame for tracking motion over time. However, the hands occupy only a small part of the image and are often occluded by the manipulated object or by the person's body [12, 17]. Since the egocentric and exocentric cameras record the same interactions, we can compare the two viewpoints using the same takes and ground truth.

Task. Given either one exocentric stream or all synchronized exocentric streams, a method estimates the articulated pose of both hands in each frame. Methods with world-frame outputs additionally recover the hand trajectories across the sequence. The ground truth and evaluation procedure are the same as those used in Section 5.3.1, allowing direct comparison between the two viewpoints.

Metrics. We use the same five hand-pose metrics defined in Section 5.3.1. For video methods, we also report the world-frame trajectory error. Although the exocentric camera is static, a method must still recover hand positions consistently and at the correct metric scale throughout the clip.

Baselines. We evaluate five single-view methods using their released pre-trained weights. HaMeR [51], WiLoR [115], and OmniHands [116] estimate the hands independently in each frame. HaPTIC [53] processes video and produces hand motion in a world coordinate frame. HORT [54] jointly reconstructs the right hand and the manipulated object, and therefore produces predictions only for frames containing object manipulation.

Results. Table 6 summarizes the results. The five methods achieve similar articulation accuracy, with PA-MPJPE values ranging from 9.1 to 10.8 mm. WiLoR performs best, reaching 9.1 mm PA-MPJPE, an F@5 score of 0.313, and an AUC_J of 0.819. HaMeR follows closely with a PA-MPJPE of 9.6 mm. These results indicate that current crop-based hand regressors can recover detailed hand poses even when the hands occupy a small region of a 1080p image.

The fixed camera also makes global trajectory estimation more reliable. HaPTIC obtains a trajectory error of 63 mm while maintaining a PA-MPJPE of 10.0 mm, which is comparable to the per-frame methods. Its trajectory error is substantially lower than the 98–102 mm obtained by the egocentric world-space methods. This difference suggests that a stable camera coordinate system removes much of the uncertainty caused by egomotion.

HORT achieves a PA-MPJPE of 10.8 mm on the manipulation frames for which it produces predictions. This result is close to those of the other methods. Under this evaluation, jointly modeling the held object does not provide a clear improvement in articulated hand-pose accuracy, but it also does not noticeably reduce it.

5.3.3 Cross-View Analysis

Because the two viewpoints observe the same interactions, their results can be compared directly. Although the egocentric camera provides a closer view of the hands, the exocentric methods achieve better results for both articulation and global trajectory estimation. Among these results, the trajectory analysis shows the largest difference. The exocentric video method obtains an error of 63 mm, while the egocentric methods remain close to 100 mm. With a fixed exocentric camera, the camera coordinate system already provides a stable reference frame. Egocentric methods must estimate this reference frame from head motion, and errors in this step become a major part of the final trajectory error.

Overall, the two viewpoints still provide complementary observations. Egocentric video is more affected by truncation, distortion, and motion blur, while exocentric video is more affected by occlusion from the body and manipulated objects. Combining both streams could therefore reduce their individual failure cases. Providing measured headset motion to egocentric methods would also help separate errors caused by hand reconstruction from those caused by egomotion estimation. In addition, supplying object pose as an auxiliary input could test whether scene and object information improve the recovery of grasps and hand motion. We believe that ACE-Data-0 will support further research on combining egocentric and exocentric views for more accurate pose estimation.

6 Conclusion

We presented ACE, an ambient capture methodology for recording everyday household behavior in real homes. ACE uses two complementary configurations at different spatial scales: a table-scale setup for fine-grained hand-object manipulation and a room-scale setup for whole-body activity across a fully furnished apartment. Both capture synchronized egocentric and exocentric video, body, hand, and object motion, audio, and tactile signals. An optical-clock procedure aligns all streams to the motion-capture clock at millisecond precision, while marker-bridged calibration registers static and wearable cameras in a common world frame.

Using ACE, we collected ACE-Data-0, comprising over 150 hours of recording, 75,000 interaction episodes, 17M frames, and per-frame annotations. We benchmarked existing methods on touch prediction from video, full-body motion recovery, and hand-motion estimation from egocentric and exocentric views. Across these tasks, existing methods degrade under contact, occlusion, and long-duration activity—conditions that distinguish real homes from controlled laboratories. These results motivate models that fuse views and modalities, enforce physical constraints, and learn from measured contact and motion supervision.

By combining egocentric observations, demonstration trajectories, and contact-level supervision in a single time-aligned stream, ACE-Data-0 is designed to support research on manipulation policies, world models, and vision-language-action systems that connect perception, action, and physical state in real homes.

Limitations. ACE has several limitations. First, it covers only two sites and therefore captures limited variation in layouts, furnishings, and lighting. Second, ground truth is restricted to instrumented entities: tracked objects must be scanned and equipped with markers in advance, and the dataset does not annotate state changes of articulated mechanisms, fluids, or deformable materials. Third, the suit, gloves, headset, and markers remain visible in the recordings and may introduce dataset-specific visual cues.

Ethics statement. All participants volunteered and provided informed consent for both recording and public data release.

References

[1] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[2] Rolf Pfeifer and Josh Bongard. *How the body shapes the way we think: a new view of intelligence*. MIT press, 2006.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[7] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, et al. Ego-Exo4D: understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *IJCV*, 130 (1):33–55, 2022.

[9] Ropedia. Xperience-10M: A large-scale egocentric multimodal dataset with structured 3D/4D annotations. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/ropedia-ai/xperience-10m.

[10] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: dataset and method for tracking human object interactions. In *CVPR*, 2022.

[11] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

[13] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *CVPR*, 2024.

[14] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: hand and object tracking in 3D from egocentric multi-view videos. In *CVPR*, 2025.

[15] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020.

[16] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021.

[17] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. GigaHands: A massive annotated dataset of bimanual hand activities. In *CVPR*, 2025.

[18] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2O: two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021.

[19] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies. *arXiv 2602.15922*, 2026.

[20] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. EgoVLA: learning vision-language-action models from egocentric human videos. In *CoRL*, 2025.

[21] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *CoRL*, 2025.

[22] Yue Su, Sijin Chen, Haixin Shi, Mingyu Liu, Zhengshen Zhang, Ningyuan Huang, Weiheng Zhong, Zhengbang Zhu, Yuxiao Liu, and Xihui Liu. World Guidance: world modeling in condition space for action generation. *arXiv 2602.22010*, 2026.

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[24] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *ICCV*, 2023.

[25] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: towards egocentric life assistant. In *CVPR*, 2025.

[26] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *CVPR*, 2025.

[27] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020.

[28] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022.

[29] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*, 2022.

[30] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: benchmarking generalizable bimanual Tool-ACtion-Object understanding. In *CVPR*, 2024.

[31] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: joint markerless 3D tracking of humans and objects in interaction. *IJCV*, 132(7):2551–2566, 2024.

[32] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023.

[33] 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 *ECCV*, 2022.

[34] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar M. Parkhi, Richard A. Newcombe, and Carl Yuheng Ren. Aria Digital Twin: A new benchmark dataset for egocentric 3D machine perception. In *ICCV*, 2023.

[35] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM TOG*, 42(6):197:1–197:11, 2023.

[36] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, 2024.

[37] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024.

[38] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: parameterizing everyday home activities towards 3D generative modeling of human-object interactions. In *CVPR*, 2024.

[39] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob J. Engel, Renzo De Nardi, and Richard A. Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024.

[40] Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D dataset of mocap human object interactions. In *ICCV*, 2025.

[41] Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *CoRL*, 2023.

[42] Haoshu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *ICRA*, 2024.

[43] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. DexCap: scalable and portable mocap data collection system for dexterous manipulation. In *RSS*, 2024.

[44] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *RSS*, 2024.

[45] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *RSS*, 2025.

[46] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: learning dexterous manipulation from large-scale egocentric video. In *ICLR*, 2026.

[47] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and G0 dual-system VLA model. *arXiv 2509.00576*, 2025.

[48] Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: scaling dexterous manipulation with diverse egocentric human data. *arXiv 2602.16710*, 2026.

[49] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6):245:1–245:17, 2017.

[50] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint Transformer: solving joint identification in challenging hands and object interactions for accurate 3D pose estimation. In *CVPR*, 2022.

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

[52] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: world-space hand motion reconstruction from egocentric videos. In *CVPR*, 2025.

[53] Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J. Black. Predicting 4D hand trajectory from monocular videos. In *3DV*, 2026.

[54] Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: monocular hand-held objects reconstruction with transformers. In *ICCV*, 2025.

[55] Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. PhySIC: physically plausible 3D human-scene interaction and contact from a single image. In *SIGGRAPH Asia*, 2025.

[56] Yukang Cao, Jiahao Lu, Zhisheng Huang, Zhuowen Shen, Chengfeng Zhao, Fangzhou Hong, Zhaoxi Chen, Xin Li, Wenping Wang, Yuan Liu, and Ziwei Liu. Reconstructing 4D spatial intelligence: A survey. *arXiv 2507.21045*, 2025.

[57] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, Lars Paulsen, Ge Yang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy $\sim$ human policy. In *CoRL*, 2025.

[58] Fourier ActionNet Team and Yao Mu. ActionNet: A dataset for dexterous bimanual manipulation. Dataset website, 2025. URL https://action-net.org/.

[59] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: imitation learning for dexterous manipulation from human videos. In *ECCV*, 2022.

[60] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: scaling imitation learning via egocentric video. In *ICRA*, 2025.

[61] Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y. Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. EgoBridge: domain adaptation for generalizable imitation from egocentric human data. In *NeurIPS*, 2025.

[62] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: learning to act anywhere with task-centric latent actions. In *RSS*, 2025.

[63] Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-N-On: scaling egocentric manipulation with in-the-wild and on-task data. *arXiv 2511.15704*, 2025.

[64] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv 2512.22414*, 2025.

[65] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *CoRL*, 2022.

[66] Marcel Torne, Karl Pertsch, Homer Walke, Kyle Vedder, Suraj Nair, Brian Ichter, Allen Z. Ren, Haohuan Wang, Jiaming Tang, Kyle Stachowicz, Karan Dhabalia, Michael Equi, Quan Vuong, Jost Tobias Springenberg, Sergey Levine, Chelsea Finn, and Danny Driess. MEM: multi-scale embodied memory for vision language action models. *arXiv 2603.03596*, 2026.

[67] Chenyang Gu, Mingyuan Zhang, Haozhe Xie, Zhongang Cai, Lei Yang, and Ziwei Liu. Bridging semantic and kinematic conditions with diffusion-based discrete motion tokenizer. *arXiv 2603.19227*, 2026.

[68] Ronghui Li, Zhongyuan Hu, Li Siyao, Youliang Zhang, Haozhe Xie, Mingyuan Zhang, Jie Guo, Xiu Li, and Ziwei Liu. InfiniteDance: scalable 3D dance generation towards in-the-wild generalization. In *ECCV*, 2026.

[69] Yukang Cao, Liang Pan, Kai Han, Kwan-Yee K. Wong, and Ziwei Liu. AvatarGO: zero-shot 4D human-object interaction generation and animation. In *ICLR*, 2025.

[70] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *CoRL*, 2024.

[71] Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, et al. RoboCOIN: an open-sourced bimanual robotic data COllection for INtegrated manipulation. *arXiv 2511.17441*, 2025.

[72] Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, et al. RoboMIND 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv 2512.24653*, 2025.

[73] Yukang Cao, Haozhe Xie, Fangzhou Hong, Long Zhuo, Zhaoxi Chen, Liang Pan, and Ziwei Liu. HSImul3R: physics-in-the-loop reconstruction of simulation-ready human-scene interactions. In *ECCV*, 2026.

[74] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: towards unified latent VLA for whole-body loco-manipulation control. In *ICLR*, 2026.

[75] Haozhe Xie, Beichen Wen, Jiarui Zheng, Zhaoxi Chen, Fangzhou Hong, Haiwen Diao, and Ziwei Liu. DynamicVLA: A vision-language-action model for dynamic object manipulation. *arXiv 2601.22153*, 2026.

[76] Lawrence Y. Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. EMMA: scaling mobile manipulation via egocentric human data. *IEEE RA-L*, 11(3):3087–3094, 2025.

[77] Xiaomeng Xu, Jisang Park, Han Zhang, Eric Cousineau, Aditya Bhat, Jose Barreiros, Dian Wang, and Shuran Song. HoMMI: learning whole-body mobile manipulation from human demonstrations. *arXiv 2603.03243*, 2026.

[78] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024.

[79] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *PR*, 47(6):2280–2292, 2014.

[80] Roger Y. Tsai and Reimar Lenz. A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. *IEEE T-RA*, 5(3):345–358, 1989.

[81] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — A modern synthesis. In *Vision Algorithms: Theory and Practice*, 2000.

[82] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *IROS*, 2013.

[83] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019.

[84] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv 2312.11805*, 2025.

[85] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D. Twigg, Chengde Wan, James Hays, and Charles C. Kemp. PressureVision: estimating hand pressure from a single RGB image. In *ECCV*, 2022.

[86] Yuan Zeng, Yujia Shi, Tiao Tan, Xingting Li, Yaqi Qin, Zongqing Lu, Wenming Yang, Jing-Hao Xue, and Qingmin Liao. EgoTactile: learning grasp pressure for everyday objects from egocentric video. In *ICML*, 2026.

[87] Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, Xushi Wang, Yuxiang Jiang, Wenbo Ding, and Shuo Yang. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv 2605.13083*, 2026.

[88] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: unified 6D pose estimation and tracking of novel objects. In *CVPR*, 2024.

[89] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-HMR: multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024.

[90] Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-HMR 2: multi-person camera-centric human detection, mesh recovery and tracking. *arXiv 2606.14841*, 2026.

[91] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3D Body: robust full-body human mesh recovery. *arXiv 2602.15989*, 2026.

[92] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: towards well-aligned full-body model regression from monocular images. *IEEE TPAMI*, 45(10):12287–12303, 2023.

[93] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: part attention regressor for 3D human body estimation. In *ICCV*, 2021.

[94] Priyanka Patel and Michael J. Black. CameraHMR: aligning people with perspective. In *3DV*, 2025.

[95] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *CVPR*, 2023.

[96] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023.

[97] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. SMPLest-X: ultimate scaling for expressive human pose and shape estimation. *IEEE TPAMI*, 48(2):1778–1794, 2026.

[98] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

[99] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024.

[100] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024.

[101] EasyMoCap: make human motion capture easier. GitHub, 2021. URL https://github.com/zju3dv/EasyMocap.

[102] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH*, 2022.

[103] Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, Yuan Liu, and Yike Guo. UniSH: unifying scene and human reconstruction in a feed-forward pass. *arXiv 2601.01222*, 2026.

[104] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. *arXiv 2501.02158*, 2025.

[105] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: everyone everywhere all at once. In *ICLR*, 2026.

[106] Hanz Cuevas-Velasquez, Anastasios Yiannakidis, Soyong Shin, Giorgio Becherini, Markus Höschle, Joachim Tesch, Taylor Obersat, Tsvetelina Alexiadis, and Michael J. Black. MAMMA: markerless and automatic multi-person motion action capture. In *CVPR*, 2026.

[107] Xiaoben Li, Mancheng Meng, Ziyan Wu, Terrence Chen, Fan Yang, and Dinggang Shen. Human mesh recovery from arbitrary multi-view images. *arXiv 2403.12434*, 2024.

[108] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *CVPR*, 2025.

[109] Jiaman Li, C. Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023.

[110] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *CVPR*, 2025.

[111] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36 (7):1325–1339, 2014.

[112] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, 2015.

[113] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3D hand pose estimation in everyday egocentric images. In *ECCV*, 2024.

[114] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-HaMR: recovering 4D interacting hand motion from a dynamic camera. In *CVPR*, 2025.

[115] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: end-to-end 3D hand localization and reconstruction in-the-wild. In *CVPR*, 2025.

[116] Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Yebin Liu, Wei Jing, Qi Yan, Qianying Wang, and Hongwen Zhang. OmniHands: robust motion capture of interactive hands via a versatile transformer. *ACM TOG*, 42(6):197:1–197:11, 2026.