RoboMoCap: A Large-Scale Whole-Body Motion Capture Dataset with Egocentric Vision for Robot Learning

Mayur Deshmukh[1] Aaron Tung[2] Junyi Li[2] Nikash Bhardwaj[3]
Ademi Adeniji[3] Vincent Liu[3] Michael Cho[2] Jiafei Duan[1]

[1]MAGIC Lab, National University of Singapore [2]BitRobot [3]Grounded Superintelligence

https://bitrobot-foundation.github.io/robomocap/

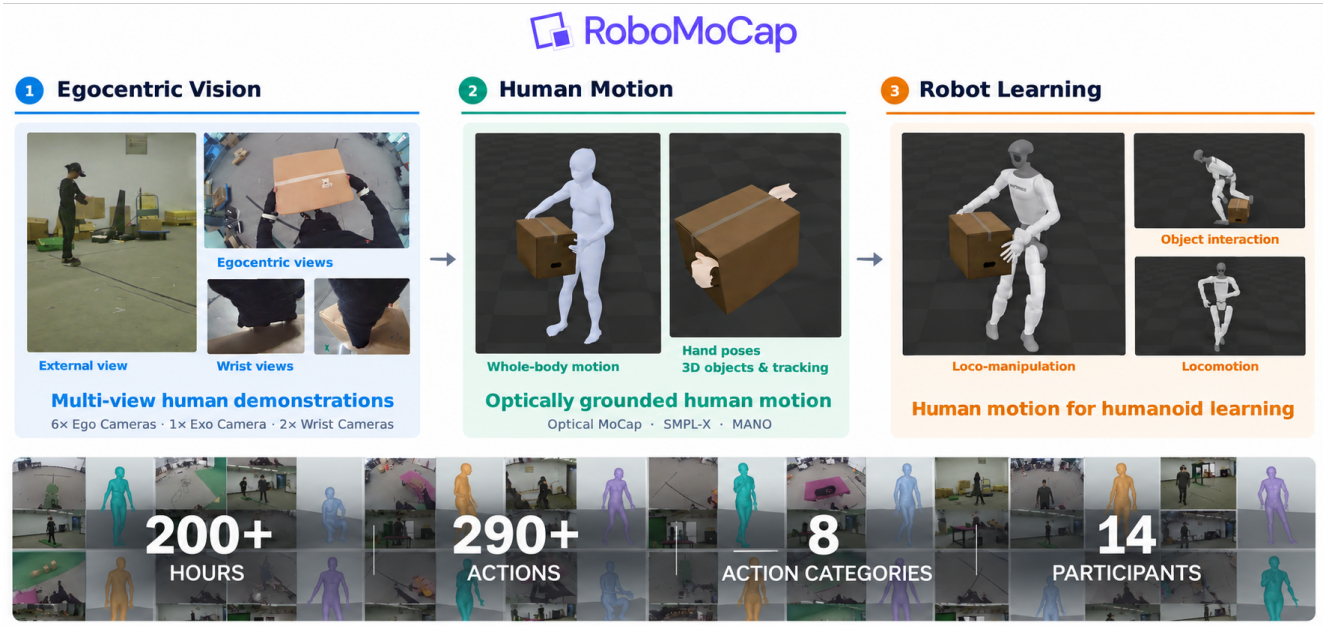


Figure 1. **RoboMoCap: From Human Demonstrations to Robot Learning.** RoboMoCap connects egocentric vision with whole-body motion for humanoid robot learning. (1) **Egocentric Vision** captures synchronised multi-view observations from head-mounted, wrist-mounted, and exocentric cameras. (2) **Human Motion** provides optically captured whole-body motion, articulated hand poses, and tracked 3D objects. (3) **Robot Learning** leverages diverse locomotion, manipulation, and loco-manipulation demonstrations to support humanoid robot learning. The dataset comprises over 200 hours of recordings from 14 participants, covering over 290+ actions across eight categories.

Abstract

Humanoid robots must coordinate locomotion and manipulation to navigate complex environments and perform everyday tasks. Learning such behaviours from human demonstrations requires accurate whole-body motion paired with egocentric vision, particularly for lower-body movements and foot placements that are frequently invisible from a first-person perspective. However, existing egocentric datasets often lack high-fidelity optical motion capture or systematic coverage of diverse loco-manipulation behaviours. To address this gap, we introduce RoboMoCap, a large-scale multimodal dataset combining egocentric vision with marker-based optical motion capture, specifically designed for locomotion-centric humanoid learning. RoboMoCap comprises over 200 hours of recordings from 14 participants, featuring six head-mounted cameras, wrist-mounted visual and inertial sensing, and optical full-body motion

capture at 120 Hz. The dataset systematically covers diverse locomotion primitives, dynamic whole-body movements, and loco-manipulation tasks, including variable-speed walking, turning, constrained navigation, terrain traversal, object transportation, and manipulation during locomotion. It provides temporally aligned multimodal observations and metric whole-body motion representations, enabling the study of human movement beyond directly observable body regions. We systematically evaluate existing methods on three benchmark tasks: egocentric full-body motion reconstruction, foot contact and placement estimation, and human-to-humanoid motion retargeting. By pairing egocentric vision with high-fidelity human motion across diverse locomotion and interaction scenarios, RoboMoCap provides a foundation for egocentric motion understanding and humanoid learning from human demonstrations.

1. Introduction

Humanoid robots are increasingly expected to operate in human environments, where successful task execution requires the seamless coordination of locomotion, balance, and manipulation [18, 23, 30, 35]. Everyday activities, such as carrying objects across a room, navigating narrow passages, stepping over obstacles, or approaching a table to place an object, demand continuous coordination between the upper and lower body. Rather than treating locomotion and manipulation as independent skills, these behaviours require adapting foot placement, body posture, and movement trajectories to the surrounding environment and ongoing interactions. Learning such coordinated behaviours from human demonstrations offers a promising direction for developing versatile humanoid robots. However, this approach fundamentally depends on datasets that capture not only what humans observe and interact with, but also how their entire bodies move throughout these activities.

Egocentric sensing provides a natural perspective for learning from human behaviour, capturing visual observations closely aligned with the information available to an embodied agent. Large-scale egocentric datasets have demonstrated the value of first-person recordings for activity understanding, human-object interaction, and robot learning. Nevertheless, recovering precise whole-body motion from egocentric observations remains challenging. In particular, the lower body is frequently outside the field of view of head-mounted cameras, making foot placement, support transitions, and lower-body articulation difficult to infer directly. Furthermore, similar head trajectories may correspond to substantially different locomotion patterns, while head orientation does not necessarily align with body heading or travel direction. These ambiguities become particularly pronounced during turning, lateral or backward walking, dynamic posture transitions, and object transportation. Consequently, connecting egocentric perception with accurate whole-body motion supervision is essential for understanding locomotion and its interaction with manipulation.

Existing datasets capture complementary aspects of human motion. AMASS [38] provides detailed human kinematics but lacks synchronised egocentric observations. Nymeria [36] pairs wearable first-person sensing with full-body motion, while EgoHTR [6] focuses on challenging terrain traversal. More recently, ACE-Data-0 [8] combines egocentric observations with optical motion capture, and Ego-Exo4D-HM [37] provides dense motion reconstructions from multi-view recordings. However, datasets combining synchronised egocentric observations and high-fidelity optical whole-body supervision across diverse locomotion and loco-manipulation behaviours remain limited. Table 1 situates RoboMoCap within this broader landscape, comparing sensing modalities, motion supervision, and behavioural coverage.

To bridge this gap, we introduce *RoboMoCap*, a large-scale multimodal human motion dataset designed for locomotion-centric humanoid learning. RoboMoCap combines a custom six-camera head-mounted sensing platform with wearable wrist sensors and a marker-based optical motion capture system. The wearable cameras record complementary scene-facing and body-facing perspectives, while optical motion capture measures full-body motion at 120 Hz. Through temporal synchronisation and spatial calibration, we associate egocentric observations with metric body motion, tracked camera trajectories, and fitted SMPL-X [43] representations. Crucially, marker-based optical motion capture provides whole-body supervision independently of egocentric visual inference, enabling direct evaluation of lower-body motion even when the corresponding body regions are not visible in the wearable cameras.

A central design objective of RoboMoCap is to capture human locomotion beyond conventional forward walking. To this end, our collection protocol systematically organises activities into locomotion primitives, posture and whole-body dynamics, and loco-manipulation. Locomotion tasks span variations in speed, movement direction, turning, obstacle negotiation, and terrain conditions, while dynamic activities introduce changes in posture, support configuration, and movement intensity. Loco-manipulation tasks further couple locomotion with object acquisition, transportation, and placement under varying loads and environmental constraints. Together, these activities capture both isolated motion primitives and coordinated whole-body behaviours, enabling systematic analysis of how locomotion adapts to environmental conditions and concurrent manipulation demands.

Beyond providing synchronised multimodal recordings,

RoboMoCap enables systematic evaluation of egocentric whole-body motion understanding and human-to-humanoid motion transfer. We evaluate existing methods across three benchmark tasks: egocentric full-body motion reconstruction, foot contact and placement estimation, and human-to-humanoid motion retargeting. These evaluations examine the reconstruction of whole-body kinematics, the accuracy and physical plausibility of lower-body motion, and the preservation of human motion characteristics when transferred to humanoid embodiments. Together, the dataset and benchmarks provide a resource for investigating egocentric motion understanding and its potential applications in humanoid learning.

Our main contributions are summarised as follows:

- **The largest locomotion-centric multimodal human motion dataset.** We introduce RoboMoCap, combining synchronised egocentric visual and inertial sensing with marker-based optical full-body motion capture at 120 Hz, with a focus on diverse locomotion and loco-manipulation behaviours.
- **A systematic locomotion and loco-manipulation collection protocol.** We design a structured action inventory spanning diverse gait patterns, movement directions, posture transitions, terrain conditions, and object interactions during locomotion, providing calibrated camera trajectories and metric whole-body motion representations.
- **Benchmarks for egocentric motion understanding and humanoid retargeting.** We evaluate existing methods on egocentric full-body motion reconstruction, foot contact and placement estimation, and human-to-humanoid motion retargeting, establishing reference results for studying lower-body motion accuracy and cross-embodiment motion transfer.

2. Related Work

2.1. Motion Capture Datasets

Motion capture provides high-quality references for human motion modelling and synthesis. AMASS [38] unifies marker-based datasets under a common body-model parameterisation, while LAFAN1 [17] and 100STYLE [40] offer diverse locomotion and stylised gait sequences. Interaction-oriented datasets such as GRAB [51], OMOMO [29], and HUMOTO [32] extend motion capture to whole-body object manipulation. Although these datasets provide detailed body motion, they generally lack corresponding real egocentric observations, limiting their direct use for learning vision-conditioned humanoid behaviours. Moreover, locomotion and object manipulation are typically studied separately rather than through systematic variations of their coordination.

2.2. Egocentric Human Motion Datasets

Early egocentric pose datasets employ body-facing cameras to observe the wearer. EgoCap [47] and Mo^{2}Cap^2 [56] explore this setting, while xR-EgoPose [52], UnrealEgo2 [1], and EgoWholeBody [54] provide synthetic training data. Real-world datasets such as UnrealEgo-RW [1], EMHI [13], Ego-Elec [63], and Eva-3M [10] further support egocentric body reconstruction.

Scene-facing datasets, including EgoBody [60], HPS [16], and ADT [42], capture human activities in their environmental context. EgoBody3M [61], Ego-Exo4D [15], Nymeria [36], and NymeriaPlus [11] extend real-world egocentric motion capture through larger-scale recordings and alternative body supervision. FRAME [7] further addresses real-world egocentric body reconstruction, while EgoSim [19] and EgoWholeView [21] explore multi-camera wearable configurations for whole-body motion capture. More recently, EgoHTR [6] targets terrain traversal, while ACE-Data-0 [8] combines egocentric observations with optical motion capture for long-horizon household activities. Despite these advances, large-scale optical supervision with systematic coverage of locomotion dynamics, foot placement, and loco-manipulation remains relatively limited.

2.3. Human Motion for Humanoid Learning

Human motion datasets increasingly support humanoid control through retargeting and imitation. Humanoid-X [39] derives 240 h of motion from internet videos, while PHUMA [27] curates 73 h of physically refined locomotion and BONES-SEED [34] provides 288 h of motion capture for humanoid training. OmniRetarget [57] further enables interaction-preserving retargeting involving objects and terrain. However, these motion-oriented datasets generally lack paired egocentric observations.

Conversely, egocentric datasets for robot learning, such as EgoDex [20] and EgoVerse [44], provide large-scale human manipulation demonstrations, while HumanVerse-500 [55] and EgoExplorer [62] extend first-person observations to whole-body loco-manipulation using estimated human motion. RoboMoCap bridges these directions by pairing egocentric observations with optical whole-body motion across diverse locomotion and loco-manipulation tasks, supporting egocentric motion understanding, humanoid retargeting, and vision-conditioned humanoid learning.

3. RoboMoCap Capture System

3.1. Hardware Setup

Egocentric Cameras. All participants wear a RoboCap [2], a lightweight head-mounted sensing platform equipped with six hardware-synchronised global-shutter RGB cameras, two inertial measurement units (IMUs), and a magne-

Table 1. **Comparison of RoboMoCap with related datasets.** ✓: provided; (✓): partial, estimated, limited coverage, or third-party release; ✗: absent; –: not applicable or not reported. Year: publication venue year, or arXiv/submission year otherwise. Body GT: Opt = optical marker mocap, IMU = inertial suit, MV = markerless multi-view fitting, Ps = pseudo-labels from video or wearable devices, Syn = synthetic. #Ego: head-mounted cameras; #Wrist: wrist-mounted cameras; Body-fac.: at least one egocentric camera observes the wearer's body. Loco: locomotion beyond forward walking (backward, lateral, turning, speed). L-M: loco-manipulation, e.g., carrying or transport while walking. Frame counts (fr) are given when hours are not reported.

Dataset	Year	Hours	#Subj	Body GT	#Ego	Body-fac.	#Wrist	#Exo	IMU	Hand	Obj. 6D	Scene	Audio	Loco	L-M
Motion capture and pseudo-labelled motion															
AMASS [38]	2019	40+	344	Opt	0	✗	✗	✗	✗	(✓)	✗	✗	✗	✓	(✓)
LAFAN1 [17]	2020	4.6	5	Opt	0	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
GRAB [51]	2020	1.6M fr	10	Opt	0	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
100STYLE [40]	2022	22.1	1	Opt	0	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
OMOMO [29]	2023	10	17	Opt	0	✗	✗	✗	✗	✗	✓	✗	✗	(✓)	✓
Motion-X [31]	2023	144.2	–	Ps	0	✗	✗	✗	✗	(✓)	✗	✗	✗	(✓)	✗
HUMOTO [32]	2025	2.2	1	Opt	0	✗	✗	✗	✗	✓	✓	✗	✗	(✓)	(✓)
Egocentric body motion															
EgoCap [47]	2016	0.7	8	MV	2	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Mo^{2}Cap^2 [56]	2019	530K fr	2	Syn/MV	1	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
xR-EgoPose [52]	2019	380K fr	–	Syn	1	✓	✗	✗	✗	(✓)	✗	✗	✗	✗	✗
HPS [16]	2021	3+	7	IMU	1	✗	✗	✗	✓	✗	✗	✓	✗	(✓)	✗
EgoBody [60]	2022	220K fr	36	MV	1	✗	✗	3–5	✗	(✓)	✗	✓	✗	✗	✗
ADT [42]	2023	6.6	–	MV	3	✗	✗	✗	✓	✗	✓	✓	✓	✗	(✓)
UnrealEgo2 [1]	2024	2.5M fr	17	Syn	2	✓	✗	✗	✗	(✓)	✗	(✓)	✗	(✓)	✗
UnrealEgo-RW [1]	2024	1.4	16	MV	2	✓	✗	✗	✗	✗	✗	(✓)	✗	(✓)	✗
EgoWholeBody [54]	2024	870K fr	14	Syn	1	✓	✗	✗	✗	✓	✗	✗	✗	(✓)	✗
Ego-Exo4D [15]	2024	1286	740	Ps	3	✗	✗	4–5	✓	(✓)	✗	✓	✓	(✓)	(✓)
Nymeria [36]	2024	300	264	IMU	3	✗	2	1	✓	✗	✗	(✓)	✓	(✓)	(✓)
EgoBody3M [61]	2024	30+	120	MV	4	✗	✗	✗	✗	✗	✗	✗	✗	(✓)	✗
EMHI [13]	2025	28.5	58	MV+IMU	2	✓	✗	✗	✓	✗	✗	✗	✗	(✓)	✗
Ego-Elec [63]	2025	9	–	IMU	3	✓	✗	✗	✓	✗	✗	✗	✗	(✓)	✗
FRAME [7]	2025	7.4	16	MV	2	✓	✗	✗	✗	✗	✗	✗	✗	(✓)	✗
Eva-3M [10]	2026	3.0M fr	31	MV	2	✓	✗	✗	✗	✗	✗	✗	✗	(✓)	✗
OmniEgoDB [9]	2026	3	4	IMU	1–3	✗	✗	–	✓	(✓)	✗	✗	–	(✓)	–
Human data for robot learning															
Humanoid-X [39]	2025	240.3	–	Ps	0	✗	✗	–	✗	✗	✗	✗	✗	(✓)	✗
PHUMA [27]	2026	73.0	–	Opt+Ps	0	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
EgoDex [20]	2026	829	–	✗	1	✗	✗	✗	–	(✓)	✗	✗	✗	✗	✗
OmniRetarget [57]	2026	∼9	–	Opt	0	✗	✗	✗	✗	✗	(✓)	(✓)	✗	(✓)	✓
BONES-SEED [34]	2026	288	522	Opt	0	✗	✗	✗	✗	–	✗	✗	✗	✓	(✓)
EgoHTR [6]	2026	1.37	8	IMU	3	✗	✗	2	✓	✗	✗	✓	✓	✓	✗
EgoVerse [44]	2026	1362	2087	✗	1+	✗	✗	✗	–	(✓)	✗	✗	–	✗	✗
HumanVerse-500 [55]	2026	500	–	Ps	1	✗	✗	✗	–	(✓)	✗	–	–	(✓)	✓
Gen-HumanWholeBody [14]	2026	600	–	Ps	6	–	–	–	–	(✓)	–	–	–	(✓)	✓
EgoExplorer [62]	2026	100	–	Ps	1	✗	✗	✗	–	✓	–	✓	–	(✓)	✓
ACE-Data-0 [8]	2026	150	50	Opt	4	✓	✗	8	✓	✓	✓	✗	✓	(✓)	✓
RoboMoCap (Ours)	**2026**	**200+**	**14**	**Opt**	**6**	✓	**2**	**1**	✓	(✓)	✓	(✓)	✓	✓	✓

tometer (Fig. 2-(A)). The camera configuration comprises two fisheye stereo pairs (front-facing and downward-facing eye cameras) and two lateral monocular wide-angle cameras, providing complementary views of the surrounding environment, hands, and body. All six cameras record at a resolution of 1920 × 1080 pixels and 30 Hz, with hardware-triggered synchronisation. The two IMUs record accelerometer and gyroscope measurements at 200 Hz, while the auxiliary magnetometer operates at 100 Hz.

Wrist-mounted Cameras. Participants additionally wear a RoboWrist [5] on each wrist (Fig. 2-(B)). Each device integrates a wide-angle RGB camera, an IMU, and a magnetometer, providing complementary hand-centric observations during locomotion and object interaction. The cameras record at 1920 × 1080 pixels and 30 Hz, while the IMUs capture accelerometer and gyroscope measurements at 200 Hz and the magnetometers operate at 100 Hz. Both RoboWrist devices are paired with RoboCap for coordinated recording, with their independently clocked streams aligned during subsequent data processing.

Exocentric Camera. A single exocentric RGB camera captures a third-person view of participants within the motion capture studio, providing complementary visual observations of whole-body movements and interactions with the surrounding environment. The camera records at a resolution of 1920 × 1080 pixels and 30 Hz.

Optical Motion Capture System. To capture high-fidelity whole-body motion independently of egocentric visibility,

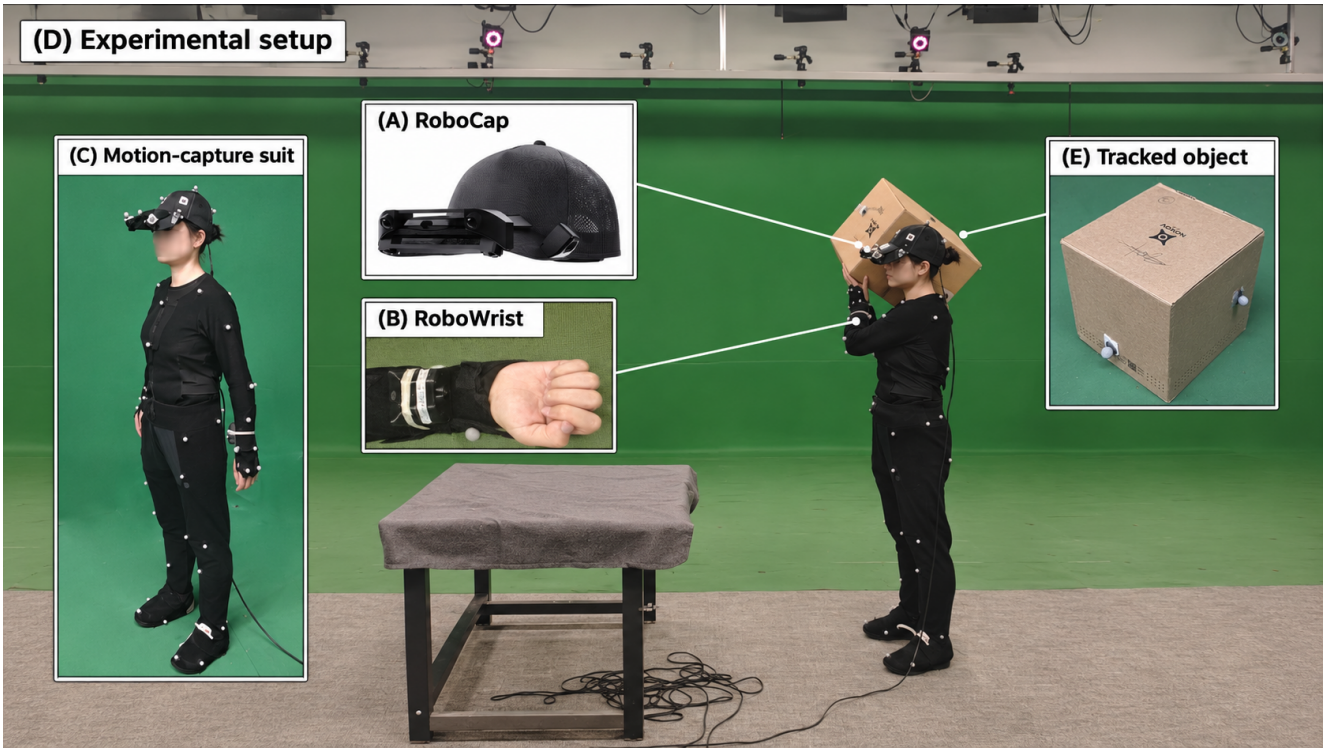


Figure 2. **RoboMoCap hardware setup.** (A) RoboCap [2] for head-mounted egocentric sensing. (B) RoboWrist [5] for wrist-mounted visual capture. (C) Motion capture suit with reflective markers for NOKOV optical tracking. (D) Recording studio equipped with NOKOV optical tracking cameras. (E) Example rigid object fitted with reflective markers for 6-DoF pose tracking.

each participant wears a motion-capture suit equipped with 53 reflective markers (Fig. 2-(C)), tracked using a NOKOV optical motion capture system [41]. The system employs multiple infrared cameras distributed around the recording area to capture three-dimensional marker trajectories at 120 Hz (Fig. 2-(D)). Additionally, four reflective markers are rigidly attached to the RoboCap headset, forming an independently tracked rigid body, denoted as *Tracker0*. This configuration enables 6-DoF tracking of the headset in the optical motion capture coordinate system, providing a reference for temporal synchronisation and spatial registration between the wearable sensors and optical motion capture.

Object Tracking. We first acquire a 3D scan of each rigid object using the LiDAR-based Reality Composer app [3] on an iPhone 14 Pro. We then attach four reflective markers to each object (Fig. 2-(E)), enabling the NOKOV optical motion capture system to track its 6-DoF pose at 120 Hz. This setup provides metric object trajectories throughout manipulation and transportation tasks, allowing analysis of object motion alongside synchronised whole-body movements.

Audio. A microphone attached to the RoboCap records audio during data collection, capturing audible cues associated with contact events, spoken instructions, and ambient environmental sounds.

3.2. Capture Environment

A central design consideration of RoboMoCap is to capture diverse whole-body movements while maintaining reliable optical motion tracking. Data collection is conducted in NOKOV's motion capture studio, comprising a 7×7 m recording area (Fig. 2-(D)). Unlike conventional manipulation-centred capture setups [20, 44, 51] that primarily focus on stationary workspaces, our recording environment accommodates continuous locomotion, changes in movement direction, dynamic postural transitions, and interactions with surrounding objects. The recording area is configured to support a wide range of locomotion behaviours, including straight-line walking, turning, circular and figure-eight trajectories, and repeated shuttle movements. Task-specific arrangements introduce spatial constraints and environmental variations, including narrow passages, obstacles, inclined surfaces, and changes in elevation. For loco-manipulation activities, interaction areas are arranged to accommodate object pickup, transportation, and placement at different locations and heights. Objects with varying physical properties and carrying requirements are incorporated to capture the coordination between locomotion, posture, and manipulation. These configurations allow both controlled motion primitives and continuous whole-body activities to be recorded within a common sensing en-

vironment.

4. RoboMoCap Dataset

4.1. Design Principles and Action Taxonomy

RoboMoCap is designed to support egocentric whole-body motion understanding and humanoid robot learning, with a particular emphasis on locomotion and its coordination with manipulation. We adopt four key design principles to guide data collection.

(i) Locomotion-centric diversity: Human locomotion involves substantially more than steady forward walking. Everyday movement requires changes in speed, direction, body orientation, and foot placement to accommodate environmental constraints. We therefore emphasise diverse locomotion behaviours, including turning, lateral and backward walking, obstacle avoidance, terrain traversal, and dynamic postural transitions. This systematic coverage aims to capture a broad distribution of lower-body movements relevant to humanoid locomotion.

(ii) Egocentric motion observability: A fundamental challenge in egocentric motion reconstruction is that the lower body, particularly the feet, is frequently outside the camera's field of view. Moreover, similar head movements can correspond to different body configurations and foot placements. We design activities that expose these ambiguities, including changes in gaze, walking direction, and body orientation. By pairing egocentric observations with optical whole-body motion measurements, RoboMoCap provides supervision for both visible and unobserved body movements.

(iii) Locomotion–manipulation coordination: Humanoid robots must coordinate locomotion with object interaction rather than execute these behaviours independently. We therefore incorporate tasks involving object pickup, transportation, placement, and manipulation during locomotion. Variations in carrying configuration, object load, and navigation constraints capture how upper-body activities influence whole-body coordination and locomotion strategies.

(iv) Structured and continuous behaviours: We combine controlled motion primitives with extended, multi-stage activities to capture both individual movement patterns and their temporal transitions. Structured task variations enable comparisons across movement conditions, while continuous recordings preserve natural transitions between locomotion, postural adjustments, and object interactions. This design supports both fine-grained motion analysis and longer-horizon embodied tasks.

Action Taxonomy. To systematically capture diverse whole-body behaviours, we organise RoboMoCap around three core action categories: locomotion primitives, posture and whole-body dynamics, and loco-manipulation. Beyond these core protocols, the recorded dataset also includes static manipulation, environment-coupled whole-body activities, dynamic recovery, human–human loco-manipulation, and long-horizon composite tasks. These additional activities extend the dataset towards more complex whole-body interactions and multi-stage behaviours. Tab. 2 summarises the taxonomy and its principal variations. The collection protocol comprises 291 predefined tasks, systematically varying locomotion patterns and body configurations.

4.2. Collection Protocol

Data collection follows predefined protocols specifying task instructions, scene configurations, repetitions, and movement variations. Participants are instructed to move naturally at a comfortable pace without imitating robotic movements. Task conditions are systematically varied through movement direction, path geometry, obstacle arrangements, object loads, and interaction targets. Left- and right-sided variations are explicitly instructed to encourage balanced coverage.

Repeated movements are recorded in continuous takes rather than isolated repetitions, preserving natural transitions between motion phases. Loco-manipulation tasks capture object acquisition, transport, and placement, including activities requiring simultaneous walking and manipulation. To isolate the influence of locomotion, selected static manipulation tasks are paired with corresponding loco-manipulation protocols, using the same participant and objects. Long-horizon tasks specify initial scene configurations and goals while allowing participants to determine their own execution strategies.

Each take is associated with participant and action identifiers, task conditions, and execution metadata. Objects are reset between takes, while carried loads and reaching heights are adjusted to individual comfort. Equipment checks and task-specific safety procedures are incorporated for dynamic movements and recovery activities.

4.3. Synchronisation and Calibration

Temporal Synchronisation. Synchronising RoboCap's visual-inertial recordings with the optical motion capture system is essential for establishing accurate correspondences between egocentric observations and whole-body motion. Since the two systems operate on independent clocks, we estimate their temporal alignment by exploiting the head motion observed simultaneously by RoboCap's IMU and the optically tracked rigid body (Tracker0) attached to the headset.

We first construct a reference timebase for the optical motion capture system using its frame indices and the recording frequency of 120 Hz, with the origin defined at the first successfully solved body frame. We then compute angular-speed magnitudes from the RoboCap gyro-

Table 2. **RoboMoCap action taxonomy.** The dataset defines 291 action protocols across eight categories, spanning locomotion, loco-manipulation, dynamic recovery, collaborative interactions, and long-horizon activities. The dataset comprises over 200 hours of multi-modal recordings from 14 participants.

Action category	#	Representative tasks	Task variations
Locomotion primitives	44	Variable-speed walking, turns and figure-eights, backward and sideways walking, narrow passages, ramps and stairs	Speed, acceleration, travel direction, path curvature, stance width, obstacles and terrain
Locomotion posture and dynamics	60	Sit-to-stand, squatting, kneeling, lying and rising, jumping, mimed sports and combat, dance, running and cutting	Posture, support configuration, movement intensity, in-place vs. travelling, gait transitions
Loco-manipulation	86	Height-paired transport, varied-load carries, carrying over terrain and stairs, manipulation while walking, bin loading, precision placement, deformable loads	Object weight and shape, carry position, pick and place height, route and terrain, concurrent manipulation
Static manipulation	44	Stationary pick-and-place, drawer, lid and latch operation, folding and pouring, tool use, reach-envelope mapping	Reach height and direction, articulation type, uni- vs. bimanual coordination, object rigidity
Long-horizon composite	11	Delivery shelving, order picking, restocking, palletising, sorting, kit assembly and room tidying	Task order, object selection, goal specification, start-state and resource perturbations
Dynamic and recovery	15	Recovery stepping, trips, surface changes, reduced friction, load perturbations, controlled falls and rising	Perturbation source and direction, support surface, carried load and recovery strategy
Human–human loco-manipulation	14	Static and walking handovers, two-person lifting and carrying, relay chains, assisted lifts, passing in narrow spaces	Partner movement, object weight and length, relative position, load sharing and route constraints
Environment-coupled whole-body	17	Cart pushing and pulling, cart slalom and parking, sled dragging, bracing against fixed loads, hoisting and lowering	Resistance type, load, direction, obstacle layout and sustained force
Total	**291**		

scope and the temporal derivatives of Tracker0's orientation trajectory. By comparing angular-speed magnitudes rather than rotation vectors, the alignment is invariant to the relative orientation between the IMU and optical tracker, eliminating the need for their spatial calibration during temporal alignment. We clip both signals at $800^\circ/\mathrm{s}$ to suppress spurious peaks and estimate the initial clock offset by maximising their temporal cross-correlation.

To improve robustness against periodic movements and ambiguous correlation peaks, we repeat the alignment over 20-second temporal windows and aggregate the estimated offsets using their median. We additionally examine the evolution of window-wise offsets to detect clock drift. When a statistically significant trend exceeds 3 ms over the recording, we estimate an affine time mapping; otherwise, we retain an offset-only model. For each session s, the resulting transformation is

$$t_{\mathrm{mocap}} = a_s t_{\mathrm{rig}} + b_s, \tag{1}$$

where t_{rig} denotes session-relative RoboCap time, b_s is the estimated temporal offset, and a_s accounts for relative clock-rate differences ($a_s = 1$ in the absence of measurable drift).

To obtain frame-level correspondence, we recover the timestamp of each video frame from the recording's device-clock start time, its presentation timestamp (PTS), and the camera–IMU temporal offset obtained during device calibration. Discontinuities in video timestamps are examined using consistency between image-derived motion and gyroscope measurements to distinguish dropped frames from timestamp artefacts. Finally, Eq. 1 maps each frame to the optical motion capture timebase. All modalities retain their native sampling rates and individual timestamps, without temporal resampling or video re-encoding. This provides a common timeline for the aligned egocentric observations, inertial measurements, and optical body motion.

Spatial Calibration. To establish spatial correspondence between the wearable cameras and optical motion capture, we use the per-device camera calibration provided by RoboCap [2]. This includes camera intrinsics and inter-camera extrinsics, with camera projection modelled using the Kannala–Brandt [25] fisheye distortion model.

To register the egocentric cameras with the optical motion capture coordinate system, we establish a fixed rigid transformation between each camera and the optically tracked headset rigid body, *Tracker0*. The world-frame pose of camera C_k at time t is obtained as

$$T_{WC_k}(t) = T_{WS}(t) T_{SC_k}, \tag{2}$$

where W denotes the canonical world coordinate sys-

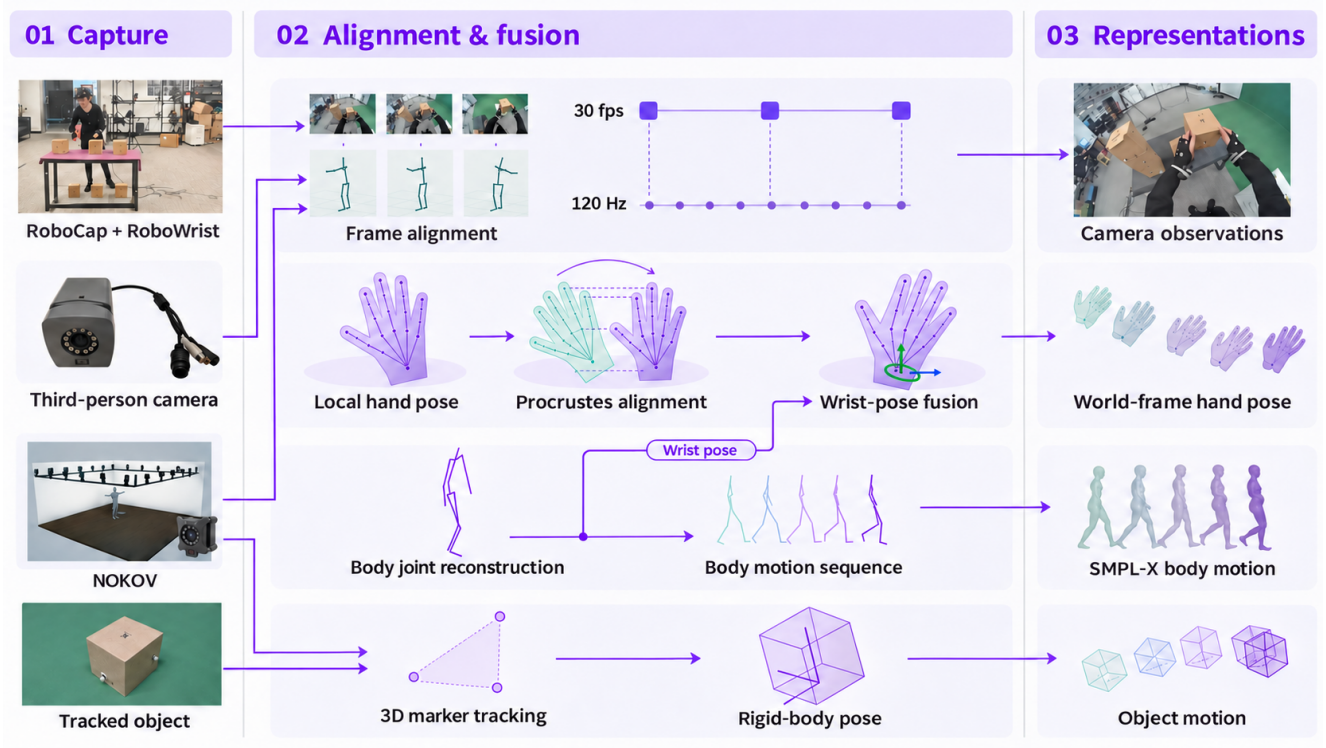


Figure 3. **Overview of the RoboMoCap capture and annotation pipeline. (1) Capture:** A multimodal recording system combines head-mounted and wrist-mounted cameras, third-person video, and NOKOV optical motion capture to record synchronised egocentric observations, full-body motion, and object interactions. **(2) Alignment and fusion:** Visual observations and optical motion capture are temporally aligned across 30 fps video and 120 Hz motion measurements. Local hand-pose estimates are aligned and fused with tracked wrist poses, while body joint trajectories and rigid-object poses are reconstructed from optical MoCap. **(3) Representations:** The resulting annotations provide calibrated camera observations, world-frame hand poses, SMPL-X body motion sequences, and tracked object motion for egocentric whole-body motion understanding and humanoid learning.

tem, S represents the Tracker0 coordinate frame, $T_{WS}(t)$ is its optically tracked 6-DoF pose, and T_{SC_k} is the calibrated rigid transformation from camera C_k to Tracker0.

This registration provides metric camera trajectories aligned with the optical whole-body motion, enabling direct correspondence between egocentric observations and 3D body annotations. The wrist-mounted and exocentric cameras are calibrated separately, as they are not rigidly attached to the head-mounted tracking assembly.

Coordinate Canonicalisation. Raw optical observations are recorded in millimetres in the motion capture system's Y-up coordinate convention. We transform these measurements into a right-handed, Z-up world coordinate system expressed in metres, using a consistent coordinate conversion and block-level levelling transformation. The same canonical frame is used for optical marker trajectories, body representations, and registered camera poses. Importantly, the levelled world origin is not assumed to coincide exactly with the physical ground-contact surface; foot-contact analysis must account for marker attachment and support-surface geometry.

4.4. Multimodal Annotation Generation

Figure 3 illustrates the RoboMoCap capture and annotation pipeline, which integrates synchronised egocentric observations with optical motion capture to generate consistent representations of full-body motion, articulated hand poses, and object trajectories.

Whole-Body Motion Reconstruction. To obtain a consistent parametric representation of human motion, we fit the SMPL-X body model [43] to the optical motion capture measurements. The fitting process leverages the articulated skeleton reconstructed by NOKOV for pose initialisation, while the observed 3D marker trajectories provide geometric constraints on the body surface.

We first estimate body shape from individual recording sessions and aggregate these estimates to obtain a consistent participant-specific shape. With the body shape fixed, we optimise the time-varying body pose and global translation, together with session-specific marker attachment offsets, to minimise discrepancies between observed marker positions and their corresponding locations on the fitted SMPL-X body.

The resulting SMPL-X parameters provide a consistent representation of articulated body motion, including global position, body orientation, joint articulation, and mesh geometry. Unlike motion estimated solely from egocentric observations, the fitting process is constrained by optical MoCap measurements of the entire body, including the legs and feet, which are frequently outside the wearable cameras' field of view. This enables the derivation of lower-body kinematics and foot trajectories for locomotion-oriented analysis and evaluation.

Hand Motion Estimation. To complement the optical whole-body measurements with articulated hand motion, we use Grounded API [2]. Its multi-view hand tracking model processes calibrated egocentric camera observations to estimate metric 3D hand poses represented using the MANO hand model [48]. By combining information from multiple camera views, the method estimates hand articulation under challenging egocentric conditions, including partial visibility and hand–object occlusion. Unlike the optically measured body markers, these hand annotations are vision-derived estimates. We therefore distinguish estimated finger articulation from optical body motion measurements in the released annotations.

To produce hand reconstructions that are consistent with the body meshes, we align Grounded API's local hand pose with the SMPL-X wrist pose. We fine-tune the model on a small training set (34.9 minutes), use the validation set to select the best checkpoint, and report the Procrustes-Aligned Mean Per-Joint Position Error (PA-MPJPE) on a test set (9.2 minutes). We use the same subjects for these splits as in the rest of the dataset. On the test split, Grounded API reaches a PA-MPJPE of 4.5 mm.

Scene Reconstruction. We reconstruct the static 3D environment scene using MapAnything [26], a feed-forward metric 3D reconstruction model. For this, we use the front-facing ego stereo cameras, together with their calibrated intrinsics and camera poses obtained from NOKOV optical tracking, to estimate scene depth. The predicted depth maps are back-projected into 3D and transformed into the NOKOV world coordinate system using the corresponding optically tracked camera poses. Consequently, the reconstructed scene geometry is directly expressed in the same coordinate frame as the captured human motion, without requiring any additional point-cloud alignment.

Task Annotations. We provide detailed textual descriptions for each task, specifying the intended activity and its execution conditions. Each recording is additionally annotated with the start and end timestamps of individual task episodes, enabling the extraction of temporally localised action sequences.

Data Quality. Each session passes automated checks on stream integrity, calibration and marker labelling before release. Ego and mocap streams are aligned by cross-correlating the head-IMU gyroscope with the angular speed of the head-mounted rigid body (median peak correlation 0.985); offsets estimated in independent windows of a take deviate from the per-take estimate by a median of 0.5 ms (95th percentile 26 ms). The fitted soles rest a median of 10 mm above the floor (IQR 4–16 mm), consistent with footwear on a barefoot body model. Per-frame validity masks are provided for the body, marker, skeleton, head-rig, camera-pose and object streams.

4.5. Dataset Statistics and Analysis

4.5.1. Dataset Statistics.

Action Distribution. Fig. 5 summarises the distribution of 200+ h across eight action categories. Locomotion primitives, posture and dynamics, and loco-manipulation collectively account for 161 hrs ($\approx$ 80.6%) of the recordings, reflecting the dataset's emphasis on locomotion-centric whole-body behaviours.

Participant and Recording Coverage. Participants contribute a median of 13.67 h (range: 9.79–21.95 h), with the largest contribution accounting for 11.4% of the dataset. Recording sessions have a median duration of 127 s (IQR: 106.8–135.0 s), preserving continuous motion rather than isolated frames.

Optical Annotation Quality. Optical capture achieves 99.81% valid skeleton frames and 99.74% marker visibility. SMPL-X fitting yields a median marker reconstruction error of 4.78 mm, demonstrating close agreement between fitted body models and observed markers.

4.5.2. Locomotion and Loco-Manipulation Diversity

Recording duration and the number of action categories alone do not fully characterise the diversity of human locomotion. In particular, locomotion involves variations in translation, rotation, foot placement, and the relationship between head orientation and body movement. To quantify these properties, we analyse the optical motion capture annotations across locomotion primitives (P), posture and dynamics (D), and loco-manipulation (LM) activities. Figure 4 summarises three complementary aspects of the recorded motion.

Locomotion Dynamics. Figure 4-(A) characterises the joint distribution of pelvis translational speed and turning rate. Given the pelvis position $\mathbf{p}_t$ and heading ψ_t, we define horizontal speed and absolute turning rate as

$$v_t = \frac{\|\mathbf{p}^{xy}_{t+\Delta t} - \mathbf{p}^{xy}_t\|_2}{\Delta t}, \tag{3}$$

$$\omega_t = \frac{|\operatorname{wrap}(\psi_{t+\Delta t} - \psi_t)|}{\Delta t}. \tag{4}$$

The analysis uses moving frames sampled at 10 Hz, with speed and turning-rate bins of 0.05 m/s and 5°/s, respectively. Each heatmap displays log-transformed bin counts

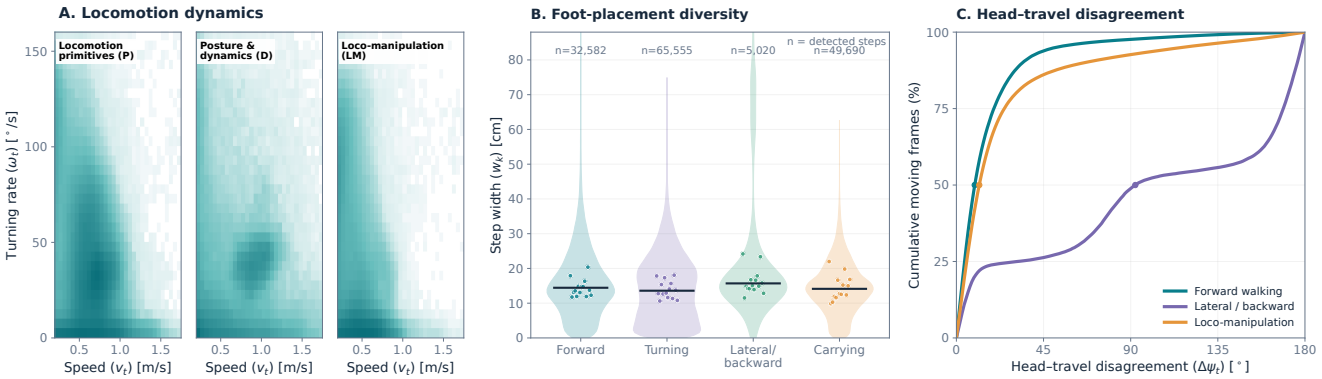


Figure 4. **Locomotion and loco-manipulation diversity in RoboMoCap.** (A) Distributions of pelvis translational speed and turning rate across locomotion primitives, posture and dynamics, and loco-manipulation activities. (B) Step-width distributions across forward walking, turning, lateral/backward walking, and object carrying. Violin widths indicate step density, horizontal lines denote medians, and dots represent participant-level medians (n denotes detected steps). (C) Cumulative distributions of angular disagreement between head orientation and travel direction across movement conditions, highlighting the challenges of inferring locomotion from egocentric observations.

Table 3. **Egocentric full-body motion reconstruction on RoboMoCap.** Comparison of existing methods across joint position and rotation accuracy, temporal motion consistency, and head pose estimation.

	Joint Position				Rotation	Temporal Motion			Head	
Method	MPJPE (mm) ↓	PA-MPJPE (mm) ↓	W-MPJPE (mm) ↓	WA-MPJPE (mm) ↓	MPJRE (°) ↓	MPJVE (cm/s) ↓	Accel. Error (cm/s^2) ↓	Jitter (m/s^3)↓	Position (mm) ↓	Orientation (°) ↓
AGRoL [12]	82.3	**51.2**	**95.4**	**64.9**	**28.2**	15.1	285.7	119.31	**30.4**	23.0
EgoPoser [24]	88.8	56.1	105.8	71.4	**28.2**	17.1	251.5	96.48	**30.4**	23.0
EgoEgo [28]	130.9	97.2	224.6	118.8	33.5	27.5	426.7	157.73	138.2	**22.9**
EgoAllo [58]	**80.8**	69.2	103.9	79.3	30.3	**14.5**	**214.5**	**72.06**	74.8	**22.9**

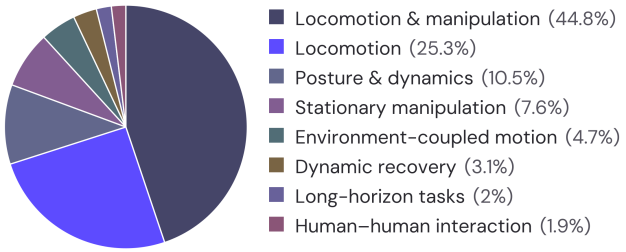


Figure 5. **Distribution of recorded data across action categories.** RoboMoCap emphasises locomotion and loco-manipulation while covering diverse whole-body activities.

normalised within its action family, allowing the shape of the motion distribution to be compared across categories.

The three action families exhibit different locomotion profiles. Locomotion primitives have a median speed of 0.65 m/s and median turning rate of 27.1°/s, whereas loco-manipulation activities have a lower median speed of 0.45 m/s and turning rate of 7.7°/s. Nevertheless, loco-manipulation includes substantial rotational motion, with a 95th-percentile turning rate of 96.3°/s. These observations show that the dataset captures diverse combinations of translation and rotation, including slower object-related movement and dynamic changes in heading.

Foot-Placement Diversity. Figure 4-(B) examines step-width distributions across four movement conditions: forward walking, turning, lateral/backward walking, and object carrying. Steps are extracted from the optically tracked lower-body motion, and step width is measured as the lateral separation of successive foot placements in the pelvis-relative coordinate frame. The violin plots represent the distributions of individual detected steps, horizontal bars indicate their medians, and dots show participant-level median step widths.

We analyse 32,582 forward-walking steps, 65,555 turning steps, 5,020 lateral/backward steps, and 49,690 carrying steps. The corresponding median step widths are 14.4, 13.6, 15.7, and 14.1 cm, respectively. Lateral/backward locomotion exhibits the largest median step width among these conditions, while all four groups show substantial within-condition variation.

These distributions demonstrate that the dataset contains diverse lower-body stepping patterns rather than only repeated forward walking. Importantly, the recorded foot

Table 4. **Human-to-humanoid retargeting: kinematic fidelity on RoboMoCap.** Comparison of human-to-Unitree G1 motion retargeting methods across 543 test sessions (≈18 h), evaluating global and root-relative positional errors for matched key bodies, feet, and hands/wrists, alongside orientation accuracy.

	Matched Key Bodies		Feet		Hands/Wrists		Orientation
Method	$E_{\text{g-mpbpe}}$ (mm) ↓	E_{mpbpe} (mm) ↓	Global (mm) ↓	Root-rel. (mm) ↓	Global (mm) ↓	Root-rel. (mm) ↓	Geodesic (rad) ↓
GMR [4]	90.7	78.6	84.8	87.7	131.5	121.7	**0.385**
PHC [33]	173.9	64.9	160.6	62.7	186.1	88.8	0.435
OmniRetarget [57]	**34.8**	**36.3**	**49.5**	**61.8**	**29.3**	**28.8**	0.470

Table 5. **Retargeting feasibility and contact preservation on RoboMoCap.** Comparison of human-to-Unitree G1 motion retargeting methods across 543 test sessions (≈18 h), evaluating foot skating, ground penetration, floating, contact preservation, joint-limit and self-collision violations, and clip-level success.

	Foot Skating		Ground Penetration		Floating	Contact F1	Constraints		Success
Method	Rate (%) ↓	p99 speed (m/s) ↓	Frames (%) ↓	Max depth (cm) ↓	Mean gap (cm) ↓	Foot (%) ↑	Joint-limit (%) ↓	Self-coll. (%) ↓	Clips (%) ↑
GMR [4]	12.3	0.36	72.5	6.8	1.5	83.0	87.2	10.7	7.4
PHC [33]	18.5	0.56	20.9	37.1	1.4	80.4	82.9	**1.1**	**79.6**
OmniRetarget [57]	**2.2**	**0.26**	**0.0**	**0.2**	**0.1**	**87.7**	**71.8**	10.0	53.6

placements provide motion-capture supervision for movements where the feet may be outside the egocentric cameras' field of view. This is particularly relevant for evaluating foot-placement estimation, support transitions, and whole-body reconstruction.

Head–Travel Disagreement. Figure 4-(C) quantifies the disagreement between the horizontal heading of the headset and the actual direction of travel. Let ψ_t^h denote the headset heading and ψ_t^v the direction of horizontal motion. Their absolute angular difference is

$$\Delta\psi_t = \left|\text{wrap}\left(\psi_t^h - \psi_t^v\right)\right|. \tag{5}$$

We compute the empirical cumulative distribution of $\Delta\psi_t$ over moving frames, separately for forward walking, lateral/backward walking, and loco-manipulation activities.

The median disagreement is $9.5°$ during forward walking, indicating that head orientation and travel direction are usually aligned. In contrast, the median reaches $92.4°$ during lateral/backward locomotion, demonstrating that the wearer can travel in a direction substantially different from the direction the headset faces. Loco-manipulation exhibits a median disagreement of $11.8°$, but its 90th percentile reaches $66.0°$, indicating substantial disagreement in a subset of moving frames.

These findings highlight a fundamental difficulty in egocentric full-body reconstruction: head orientation cannot be treated as a direct proxy for locomotion direction. Even when head motion is accurately measured, different lower-body movements and foot placements may remain compatible with similar head observations. By pairing egocentric recordings with optical full-body motion capture, RoboMoCap provides supervision for studying this ambiguity and evaluating motion reconstruction under challenging locomotion conditions.

5. Benchmarks

5.1. Egocentric Full-Body Motion Reconstruction

Task. We evaluate the ability of existing egocentric motion reconstruction methods to recover full-body human motion from ego observations. Unlike conventional pose estimation benchmarks, RoboMoCap provides optically grounded whole-body motion across diverse locomotion and loco-manipulation activities, enabling evaluation of both local articulation and global motion. We use MoCap-derived SMPL-X motion as the reference and evaluate publicly available pretrained methods without training or fine-tuning on RoboMoCap.

Metrics. We evaluate joint position accuracy using Mean Per Joint Position Error in local (MPJPE) and world-frame (W-MPJPE), sequence-aligned MPJPE (WA-MPJPE) [49], and Procrustes-aligned MPJPE (PA-MPJPE) [22]. Joint rotation accuracy is measured using mean per-joint rotation error (MPJRE). To assess temporal consistency, we report joint velocity error (MPJVE), acceleration error, and motion jitter [50], measured as the magnitude of joint jerk. We

Table 6. **Foot contact, placement, and physical plausibility on RoboMoCap.** Contact predictions are derived from reconstructed foot joints using a shared height- and velocity-based rule, calibrated on training participants and evaluated against sole-marker ground truth. Footprint and Step measure planted-foot position and step displacement errors, respectively; Skate measures foot velocity during ground-truth contact, and Penetr. measures ground penetration. GT SMPL-X is included as a reference.

	Contact	Foot placement		Plausibility	
Method	F1 (%) ↑	Footprint (mm) ↓	Step (mm) ↓	Skate (mm/s) ↓	Penetr. (mm) ↓
EgoEgo [28]	74.3	220.1	234.7	291.1	36.0
EgoAllo [58]	**90.5**	**85.1**	**154.1**	**54.8**	**8.5**
EgoPoser [24]	68.1	140.3	266.3	314.4	64.1
AGRoL [12]	75.3	113.8	237.0	273.7	66.8
GT SMPL-X	*93.6*	–	–	*43.6*	*0.0*

additionally report head position and orientation errors to characterise the accuracy of the estimated global motion.

Baselines. We evaluate on four methods: EgoEgo [28], EgoAllo [58], EgoPoser [24], and AGRoL [12]. EgoEgo and EgoAllo estimate full-body motion using egocentric visual and head-motion observations, whereas EgoPoser and AGRoL reconstruct body motion from sparse head and hand tracking. All methods are evaluated using their released pretrained models. Since their required observations differ, we interpret the results under their respective input configurations.

Results. Tab. 3 presents the quantitative results. EgoAllo achieves the lowest MPJPE (80.8 mm), whereas AGRoL obtains the lowest world-frame joint error (95.4 mm) and sequence-aligned error (64.9 mm). EgoAllo additionally achieves the lowest joint velocity error (14.5 cm/s), acceleration error (214.5 cm/s^2), and motion jitter (72.06 m/s^3). These results reveal a discrepancy between local pose accuracy and global motion reconstruction: methods that recover accurate body articulation do not necessarily preserve the corresponding motion in world coordinates. Moreover, favourable joint-position and temporal errors alone do not establish accurate foot placement or physically plausible locomotion. This motivates our dedicated evaluation of foot contact, foot placement, and lower-body motion plausibility in Sec. 5.2.

5.2. Foot Contact and Placement Estimation

Task. We further evaluate lower-body reconstruction through foot contact, placement accuracy, and kinematic plausibility for the Sec. 5.1 baselines. Contact is estimated from reconstructed foot joints using a shared height- and velocity-based rule, with thresholds calibrated on the training participants.

Metrics. We evaluate foot contact with the F1 score against contact labels derived from sole markers, following prior contact evaluation protocols [45]. Since the baselines do not predict contact explicitly, we derive it from their foot joints using a single height and speed rule, calibrated once on ground-truth bodies of the training participants and shared by all methods. Following [46, 58, 59], we measure foot skating as the horizontal speed of the foot during reference stance (*Skate*), and ground penetration as the mean depth of the lowest foot joint below the floor [50, 59]. We further introduce two footstep-level measures: the position error of each planted foot relative to the ground-truth body (*Footprint*), and the drift-free step displacement error, expressed in the frame of the preceding footprint (*Step*).

Results. Tab. 6 reports results on 543 test sessions (≈18 h), covering approximately 50k stance intervals. EgoAllo performs best across all metrics, achieving 90.5% contact F1, 85.1 mm footprint error, 154.1 mm step displacement error, and 54.8 mm/s foot skating. Despite this, substantial foot-placement errors persist. Even reference SMPL-X motion achieves only 93.6% contact F1 against sole-marker labels, highlighting the limitations of kinematics-derived contact estimation and the need for explicit lower-body evaluation.

5.3. Human-to-Humanoid Motion Retargeting

Task. We evaluate the transfer of human motion to a 29-DoF Unitree G1 [53], assessing both kinematic fidelity and geometric feasibility. RoboMoCap provides MoCap-derived SMPL-X motion and sole-marker-based foot-contact annotations, enabling evaluation beyond pose similarity. We retarget 543 test sessions (≈18 h, 30 Hz), using identical input sequences and a shared robot collision model for all methods.

Metrics. *Kinematic fidelity* (Tab. 4) is evaluated against height-scaled SMPL-X motion using 13 corresponding body segments. We report mean per-body position error with a constant horizontal offset removed per sequence ($E_{\text{g-mpbpe}}$) and relative to the hip midpoint (E_{mpbpe}) [33], alongside foot and wrist position errors and geodesic orientation error. *Feasibility and contact preservation* (Tab. 5) are assessed using the robot's collision geometry. We report foot-contact F1 against sole-marker labels, foot skating (speed > 0.1 m/s during shared stance), ground penetration (above 5 mm and maximum depth), and foot floating during human stance. We additionally measure near-joint-limit frames (within 0.01 rad) and self-collision frames (penetration > 5 mm). A clip is considered successful if its root-relative error is below 15 cm for at least 99% of frames, maximum ground penetration is below 5 cm, and self-collision occurs in at most 5% of frames.

Baselines. We compare three retargeting approaches: GMR [4], which solves inverse kinematics on scaled human targets; PHC [33], whose shape-fitting and motion-

optimisation procedure we adapt to G1; and OmniRetarget [57], which incorporates contact and kinematic constraints.

Results. OmniRetarget achieves the highest kinematic fidelity, reducing global body-position error from 90.7 mm (GMR) to 34.8 mm, with approximately fivefold lower wrist error. It also achieves the best contact preservation (87.7% F1), lowest skating rate (2.2%), and minimal ground penetration (0.2 cm maximum). PHC achieves low root-relative error (64.9 mm) but substantially higher global error (173.9 mm) and ground penetration (up to 37.1 cm). GMR exhibits ground penetration in 72.5% of frames, yielding only 7.4% clip success. Despite its superior motion fidelity, OmniRetarget achieves 53.6% clip success, compared with 79.6% for PHC, which exhibits substantially fewer self-collisions (1.1%). All methods approach joint limits in over 70% of frames, while OmniRetarget and GMR exhibit approximately 10% self-collision rates. These findings reveal a trade-off between motion fidelity and geometric feasibility, highlighting the challenges of retargeting diverse whole-body movements to humanoid robots.

6. Conclusion

We introduced RoboMoCap, a large-scale multimodal dataset that bridges egocentric perception and high-fidelity whole-body motion capture, with a particular emphasis on locomotion and loco-manipulation. By combining synchronised wearable sensing with marker-based optical motion capture, RoboMoCap provides accurate motion supervision for lower-body movements and foot placements that are often unobservable from egocentric views. Its structured collection protocol captures diverse locomotion patterns, whole-body dynamics, and object interactions, enabling systematic investigation of how humans coordinate locomotion and manipulation. Our benchmarks provide reference evaluations for egocentric full-body motion reconstruction, foot contact and placement estimation, and human-to-humanoid retargeting. By bridging egocentric perception and accurate whole-body motion, RoboMoCap opens new opportunities for learning coordinated locomotion and manipulation from human demonstrations.

Limitations. Despite its scale and motion diversity, RoboMoCap is collected within an instrumented optical motion capture environment, limiting the range of environments and terrain conditions compared with unconstrained real-world settings. Although marker-based tracking provides precise motion measurements, occasional marker occlusions and body-model fitting errors may affect the resulting annotations. Furthermore, the dataset covers a finite set of participants, objects, and interaction scenarios, and does not represent the full variability of human locomotion and manipulation. Finally, differences in human and humanoid morphology, kinematics, and physical capabilities mean that recorded motions cannot always be directly transferred to robotic platforms without retargeting and control adaptations. Extending RoboMoCap to more diverse environments and validating its utility through downstream humanoid execution are promising directions for future work.

References

[1] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 767–776. IEEE, 2024. 3, 4

[2] Qasim Ali, Nikash Bhardwaj, Kheri Hughes, Sam Cho, Ethan Lee, Paul Luo, Shinn Chong, Sunny Yu, Owen Akhibi, Alex Proshkin, Haotian Zhan, Niresh Dravin, Vivek Sharma, Jonathan Victor, Amanda Young, Michael Cho, Ademi Adeniji, and Vincent Liu. Robocap: A new platform for egocentric robot learning, 2026. 3, 5, 7, 9

[3] Apple Inc. Reality Composer. Apple App Store, 2023. Object Capture feature, version 1.6. 5

[4] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6097–6104. IEEE, 2026. 11, 12

[5] BitRobot. BitRobot Deployer: RoboCap & RoboWrist Bundle. `https://shop.bitrobot.ai/products/robocap-robowrist-bundle`, 2026. Official product page, accessed 4 November 2026. 4, 5

[6] Alex Brandes, Haig Conti Georges Sajelian, Manthan Patel, Dominik Hollidt, Chenhao Li, Matthias Heyrman, Oliver Hausdoerfer, Manuel Kaufmann, Xi Wang, Jonas Frey, et al. Egohtr: Egocentric 4d demonstrations of human terrain traversal. *arXiv preprint arXiv:2607.13472*, 2026. 2, 3, 4

[7] Andrea Boscolo Camiletto, Jian Wang, Eduardo Alvarado, Rishabh Dabral, Thabo Beeler, Marc Habermann, and Christian Theobalt. Frame: Floor-aligned representation for avatar motion from egocentric video. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17497–17507. IEEE, 2025. 3, 4

[8] Yukang Cao, Haozhe Xie, Beichen Wen, Runmao Yao, Yinghao Liu, Yue Huang, Zhichao Liao, Yunxiang Wang, Haiheng Liu, Xingshun Tian, et al. Ace-data-0: Human-centric ambient capture as embodied data engine. *arXiv preprint arXiv:2607.28625*, 2026. 2, 3, 4

[9] Kyungwon Cho, Jeonghyeon Na, and Hanbyul Joo. Omniegocap: Device-agnostic sequence-level egocentric motion reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. 4

[10] Peng Dai, Yu Zhang, Feng Yiqiang, Zhen Fan, and Yang Zhang. Egocentric visibility-aware human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7047–7057, 2026. 3, 4

[11] Daniel DeTone, Federica Bogo, Eric-Tuan Le, Duncan Frost, Julian Straub, Yawar Siddiqui, Yuting Ye, Jakob Engel, Richard Newcombe, and Lingni Ma. Nymeriaplus: Enrich-

ing nymeria dataset with additional annotations and data. *arXiv preprint arXiv:2603.18496*, 2026. 3

[12] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–490. IEEE, 2023. 10, 12

[13] Zhen Fan, Peng Dai, Zhuo Su, Xu Gao, Zheng Lv, Jiarui Zhang, Tianyuan Du, Guidong Wang, and Yang Zhang. Emhi: A multimodal egocentric human motion dataset with hmd and body-worn imus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2879–2887, 2025. 3, 4

[14] GenRobot.AI. Gen-humanwholebody. https://huggingface.co/datasets/genrobot2025/Gen-HumanWholeBody, 2026. 4

[15] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19383–19400, 2024. 3, 4

[16] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 3, 4

[17] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 3, 4

[18] Tairan He et al. Learning human-to-humanoid real-time whole-body teleoperation. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 8944–8951, 2024. 2

[19] Dominik Hollidt, Paul Streli, Jiaxi Jiang, Yasaman Haghighi, Changlin Qian, Xintong Liu, and Christian Holz. Egosim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. *Advances in Neural Information Processing Systems*, 37:106607–106627, 2024. 3

[20] Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, pages 4218–4237, 2026. 3, 4, 5

[21] Yaoqing Hu, Xuguo He, Daohan Cui, Yukun Shi, and Kai Chen. Egowholeview: Camera-only head-mounted capture of body and bimanual hand motion for embodied demonstrations. *arXiv preprint arXiv:2610.16256*, 2026. 3

[22] 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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 11

[23] Taoyang Jia, Weikai Huang, Linxin Song, Jared Darlington, Jieyu Zhao, Yue Wang, Jiafei Duan, Zhongzheng Ren, and Ranjay Krishna. Reflex: Visually grounded reactive humanoid control. *arXiv preprint arXiv:2610.12772*, 2026. 2

[24] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pages 277–294. Springer, 2024. 10, 12

[25] Juho Kannala and Sami Brandt. A generic camera calibration method for fish-eye lenses. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pages 10–13. IEEE, 2004. 7

[26] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. In *2026 International Conference on 3D Vision (3DV)*, pages 499–509. IEEE, 2026. 9

[27] Kyungmin Lee, Sibeen Kim, Youngdo Lee, Minho Park, Hyunseung Kim, Dongyoon Hwang, Donghu Kim, Hojoon Lee, and Jaegul Choo. Phuma: Physically reliable humanoid locomotion dataset. In *Conference on Robot Learning (CoRL)*, 2026. 3, 4

[28] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 10, 12

[29] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 3, 4

[30] Qiayuan Liao et al. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *Sci. Robot.*, 11(117):eadx8924, 2026. 2

[31] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 4

[32] Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10886–10897, 2025. 3, 4

[33] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 11, 12

[34] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. 3, 4

[35] Zhengyi Luo et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Sci. Robot.*, 11(117): eaed4592, 2026. 2

[36] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2024. 2, 3, 4

[37] Abhiram Maddukuri and Georgios Pavlakos. Ego-exo4d human meshes dataset: 4d human motion reconstruction for ego-exo captures, 2026. 2

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 3, 4

[39] Jiageng Mao, Siheng Zhao, Siqi Song, Chuye Hong, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, et al. Universal humanoid robot pose learning from internet human videos. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2025. 3, 4

[40] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022. 3, 4

[41] NOKOV. NOKOV Optical Motion Capture System. https://docs.nokov.com/xingying/XingYing4.7-EN/, 2026. Official system documentation, accessed 2026. 5

[42] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20076–20086. IEEE, 2023. 3, 4

[43] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 8

[44] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, RiZhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026. 3, 4, 5

[45] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European conference on computer vision*, pages 71–87. Springer, 2020. 12

[46] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11468–11479. IEEE, 2021. 12

[47] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 3, 4

[48] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 9

[49] 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 Conference Papers*, pages 1–11, 2024. 11

[50] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 11, 12

[51] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 3, 4, 5

[52] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7727–7737. IEEE, 2019. 3, 4

[53] Unitree Robotics. Unitree G1 Humanoid Robot. https://www.unitree.com/g1/. 12

[54] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 777–787. IEEE, 2024. 3, 4

[55] Chongyang Xu, Zhao Wu, Jin Chen, Yiming Jiang, Jinhui Ye, Yuming Jiang, Shifeng Zhang, Ziliang Feng, Mu Xu, Yilun Chen, et al. Towards a general humanoid loco-manipulation model via egocentric whole-body human data pretraining. *arXiv preprint arXiv:2610.00438*, 2026. 3, 4

[56] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 3, 4

[57] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11691–11698. IEEE, 2026. 3, 4, 11, 13

[58] 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7072–7084. IEEE, 2025. 10, 12

[59] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15964–15975. IEEE, 2023. 12

[60] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people

from head-mounted devices. In *European Conference on Computer Vision*, pages 180–200. Springer, 2022. 3, 4

[61] Amy Zhao, Chengcheng Tang, Lezi Wang, Yijing Li, Mihika Dave, Lingling Tao, Christopher D Twigg, and Robert Y Wang. Egobody3m: Egocentric body tracking on a vr headset using a diverse dataset. In *European Conference on Computer Vision*, pages 375–392. Springer, 2024. 3, 4

[62] Yunsong Zhou, Xuekun Jiang, Haiqi Shen, Hangxu Liu, Chaojian Shi, Yujie Liu, Jingyuan Li, Haoran Yang, Jiafei Cao, Dongzhan Zhou, Dong Wang, Yao Mu, Chunhua Shen, and Weinan Zhang. Egoexplorer: A vision-motion-touch observatory for human loco-manipulation in situ, 2026. 3, 4

[63] Siqi Zhu, Yixuan Li, Junfu Li, Qi Wu, Zan Wang, Haozhe Ma, and Wei Liang. Human motion estimation with everyday wearables. *arXiv preprint arXiv:2512.21209*, 2025. 3, 4