EgoExoMoCap: Distributed Ego-Exo Human Motion Capture

Jiaxi Jiang[1,2*] Bharat Lal Bhatnagar[1] Nan Yang[1] Lingni Ma[1] Sebastian Starke[1]
Robin Kips[1] Nadine Bertsch[1] Christian Holz[2] Federica Bogo[1]
[1]Meta Reality Labs [2]ETH Zürich
https://siplab.org/projects/EgoExoMoCap

Figure 1. EgoExoMoCap is a lightweight, distributed human motion capture system. Given two or more subjects wearing head-mounted devices, the approach combines continuous egocentric signals with intermittent exocentric camera views to robustly handle challenges such as out-of-view motions (a) and severe body occlusions (b).

Abstract

Human motion capture from head-mounted devices (HMDs) offers a scalable way to acquire real-world human motion and interaction data, which is crucial for applications in embodied AI and VR/AR. Existing approaches focus on either egocentric body tracking, estimating the motion of the subject wearing the device, or exocentric tracking, capturing the movements of people in the wearer's surroundings. So far, these two paradigms have largely been explored in isolation. In this paper, we propose a novel distributed framework that jointly leverages ego- and exocentric multi-modal signals for human motion estimation from HMDs. Unlike traditional motion capture systems requiring bulky multi-camera setups or obtrusive mocap suits, our approach, EgoExoMoCap, is as simple as two (or more) people, each wearing a pair of smart glasses. The method leverages head (plus potentially wrist) tracking signals for accurate estimation of global motion in the 3D world and combines context-aware image features based on DINOv3 to achieve robustness in the presence of noise and occlusions. Extensive experiments on two in-the-wild datasets show that our approach can robustly reconstruct motion even in challenging scenarios.

1. Introduction

Accurate and robust human motion capture in the wild is important for many applications, including robotics, virtual and augmented reality (VR/AR), and human-computer interaction. Recent efforts have shown growing interest in using egocentric devices and videos as scalable sources of human demonstrations [35, 43, 70, 73, 98], yet accurate full-body motion capture from such devices remains challenging, as traditional motion capture systems often require

*This work was done during an internship at Meta.

bulky multi-camera setups or obtrusive mocap suits.

Several approaches in the literature [71, 74, 91, 92] focus on in-the-wild capture via *exocentric* body tracking: an external RGB camera (*e.g.*, from a phone) captures the performance of the subject and the resulting monocular video is used to infer their 3D motion. These methods often struggle in reconstructing global motion in world space, given the inherent ambiguity of the problem; furthermore, they suffer in the presence of body occlusions, fast camera motions, and image blur.

An alternative is offered by the recent proliferation of wearable devices, such as smart glasses equipped with cameras and inertial sensors (*e.g.*, Project Aria [26]). Lightweight and easily usable for hours, these devices capture multi-modal streams including head and hand trajectories plus stereo/RGB images. Recent work [12, 30, 38, 40] leverage these devices for *egocentric* body motion tracking and synthesis. Approaches commonly rely on head and wrist poses, plus potentially egocentric camera streams, to infer the wearer's pose. However, sensor noise and the limited visibility of the subject's body from the egocentric cameras make it difficult to faithfully reconstruct motions.

Recent progress in multi-human motion estimation [11, 15, 82, 95] highlights the need to capture coordinated human behaviors and interactions. However, existing methods mainly estimate people from external observations or individually worn sensors, without exploiting mutual observations among multiple users. More broadly, exocentric and egocentric sensing has been largely studied in isolation, with only a few efforts using third-person or external views as training-time supervision for egocentric pose estimation [22, 88]. In contrast, a multi-HMD setup turns each participant into both a motion subject and a mobile observer of others, creating a natural opportunity for collaborative motion capture. Enabled by modern HMDs that support distributed information exchange and accurate time alignment [1], we introduce EgoExoMoCap, a unified framework that jointly leverages egocentric self-motion cues and exocentric tracking signals from Aria glasses [26] for distributed human motion capture.

The first challenge is how to effectively combine heterogeneous signals: the head (plus potentially wrist) poses tracked by the glasses worn by one subject (*wearer*), plus the images coming from the glasses worn by another subject *observing* the wearer. Our approach first estimates an initial body pose from egocentric streams, which serves as initialization and provides reliable region proposals for person localization in exocentric views. Subsequently, we detect 2D keypoints [94] in the exocentric view, unproject them to 3D rays via the observer's camera parameters, and transform them into the egocentric coordinate frame via the wearer's camera parameters. We find this a simple yet effective solution to unify input streams into a consistent (egocentric) coordinate frame. It generalizes well across diverse motions and ensures scalability to scenarios encompassing multiple observers/exocentric images.

The second challenge is posed by the unreliability of exocentric streams. As the observer looks around, the wearer might be partially or totally out of the field of view. Frequent human-scene interactions, also studied in prior work on scene-aware motion modeling and human-scene interaction [8, 54, 72, 108], cause additional body occlusions and pose ambiguities; in these scenarios, even state-of-the-art 2D keypoint estimators are not robust enough. We propose to leverage the context around the subject, computing deep image features with DINOv3 [75] and using them to learn per-keypoint confidence scores. This effectively balances the contributions of egocentric and exocentric signals, relying more on one or the other depending on the reliability of the exocentric streams.

To summarize, our contributions are as follows:

(1) We propose a lightweight, portable solution for in-the-wild motion capture, which just relies on a set of people wearing HMDs such as Aria glasses [26].

(2) We design a multi-modal framework that combines heterogeneous ego- and exocentric signals, such as continuous head trajectories and intermittent image features, to ensure robust motion estimates. The approach works with as few as two subjects and naturally scales to multi-subject setups.

(3) We evaluate our approach on indoor and outdoor sequences from two in-the-wild datasets, Nymeria [64] and EgoHumans [44], achieving state-of-the-art performance in real-world scenarios.

By reducing reliance on centralized multi-camera setups and obtrusive motion-capture suits, EgoExoMoCap makes full-body motion capture more accessible outside controlled studios, enabling scalable collection of real-world whole-body motion and interaction data for embodied AI, VR/AR, and interactive agents.

2. Related Work

Egocentric human motion estimation. The problem of full-body pose reconstruction from HMDs has received growing attention in the past years [29, 38, 40, 60, 64, 107]. AvatarPoser [38] introduced a Transformer-based framework for full-body pose estimation from sparse HMD tracking signals, demonstrating that plausible articulated body motion can be recovered from only head and wrist observations. Follow-up work [112] improves robustness via a two-stage framework leveraging body joint correlations. QuestSim [52, 93] and SimXR [63] use physics simulation to generate plausible motions. MANIKIN [39] combined neural networks with analytical inverse kinematics, using biomechanical constraints and predicted swivel angles [80] to recover full-body poses There have also been

a number of generative approaches based on VAEs [23], normalizing flows [6], VQ-VAEs [27, 34, 76], and diffusion [13, 24, 25]. Most of these approaches assume head and wrist trajectories are always available as input. However, lightweight wearable glasses, when not accompanied by additional sensors like wristbands [64], cannot provide reliable wrist trajectories [7, 16, 40]. EgoPoser [40] proposes a system that is robust to intermittent hand tracking signals and can work in large scenes via global motion decomposition, while predicting also body shape. Similarly, DSPoser [16] and EgoAllo [100] use off-the-shelf hand pose estimators to predict hand motions, which then guide full-body motion estimation. EgoEgo [53] uses head motion, while HMD2 [30] relies on head trajectories plus egocentric camera streams. RPM [12] proposes a temporally causal rolling prediction framework that produces smoother hand trajectories. All these approaches deal with the limitation of just using egocentric signals: while they can synthesize plausible motions, they can hardly reconstruct lower body movement in a faithful way.

Exocentric human motion estimation. There is a rich literature on human pose and shape (HPS) estimation from images [18, 28, 41, 47, 48, 50, 56, 57, 66, 90, 105]. In the following, we focus in particular on monocular human motion estimation from unconstrained, monocular videos. Several approaches [19, 42, 46, 62, 77] propose to combine 3D human pose estimates (either 3D joints or body model parameters [68]) with temporal models to reconstruct smooth motions. Since camera extrinsics over time are in general unknown, these methods focus on retrieving 3D body motion in camera space – without returning coherent global motion in world coordinates. Recent methods try to overcome this limitation, considering dynamic camera captures and typically following a two-stage approach: they first estimate camera parameters via SLAM [31, 32, 78, 79], and then leverage human motion priors to optimize pose in world coordinates [49, 99, 103]. Other approaches [55, 71, 74] train temporal models to directly regress global human motion from image and camera features. Others [91, 111] solve for scale ambiguities via monocular metric depth. Most approaches assume the human body is fully visible in most frames – which often does not hold in real-world scenarios [45]. Some recent work tries to directly handle body occlusions [109], also considering HMD exocentric images [107]. LAMP [97] further leverages localized multi-camera HMD input for metric 3D people tracking in world coordinates. We propose to overcome the challenges of exocentric motion estimation by effectively leveraging the multi-modal streams offered nowadays by HMDs.

Human motion capture with body-worn sensors. There is a rich literature on reconstructing body motions from body-worn inertial sensors [36, 37, 83, 85, 101, 102, 110, 114]. A well-known challenge in these pipelines is drift, caused by the absence of absolute positioning data. Approaches in the literature try to tackle this challenge by combining sensor data with additional multi-modal streams [14]. RGB cameras are typically used to obtain more robust results: [67, 86] combine body-worn IMUs with body images captured with a moving camera; [20, 58] leverage IMUs and an egocentric camera, also achieving 3D scene reconstruction. EgoSim [33] simulates multiple body-worn cameras. Other approaches also leverage depth and plantar pressure sensors [106], LiDAR and event cameras [96], and Ultra-Wideband Units [10, 59, 95]. An interesting line of work focuses on motion capture "anywhere", by combining multi-modal streams from consumer devices [89]. EgoFormer [44] tracks humans based on RGB and grayscale images captured with Aria glasses. IMUPoser [65] leverages IMUs from smartphones, smartwatches, and earbuds. [51] relies on two smartwatches and a head-mounted camera. Similar in spirit to this line of work, our approach proposes a lightweight, easily portable system; in particular, our framework is distributed and scalable – exploiting multiple devices worn by two or more people.

3. Method

3.1. Problem Formulation

We focus on full-body motion capture in a distributed setup, where two or more subjects wear an HMD equipped with inertial sensors and exocentric cameras (*e.g.*, Aria glasses [26]). For simplicity, in the following we describe a two-person scenario, involving a *wearer* (target subject whose motion needs to be reconstructed) and an *observer* (nearby subject looking at the wearer). Our approach can be extended to scenarios with more observers.

Inputs. At each timestep $t \in \{1, \dots, T\}$, our method takes as input:

- an RGB image $\mathbf{I}_t^o \in \mathbb{R}^{H \times W \times 3}$ from the observer's head-mounted camera,
- the observer's head position $\mathbf{p}_t^o \in \mathbb{R}^3$ and orientation $\boldsymbol{\theta}_t^o \in \mathbb{R}^6$ (in 6D representation [113]),
- the wearer's head position $\mathbf{p}_t^{w,\text{head}} \in \mathbb{R}^3$ and orientation $\boldsymbol{\theta}_t^{w,\text{head}} \in \mathbb{R}^6$,
- optionally, position $\mathbf{p}_t^{w,\text{lrw}}$ and orientation $\boldsymbol{\theta}_t^{w,\text{lrw}}$ of the wearer's left and right wrists.

Head trajectories (positions and orientations) can be obtained from visual-inertial SLAM systems built into HMDs [3, 5, 26]. Wrist trajectories can be obtained, for example, either from camera-equipped wristbands using SLAM [64] or from handheld controllers commonly used in VR systems [4]. We consider both the case in which wrist trajectories are available (**3-point**) and the one in which only the head trajectory is known (**1-point**). We assume

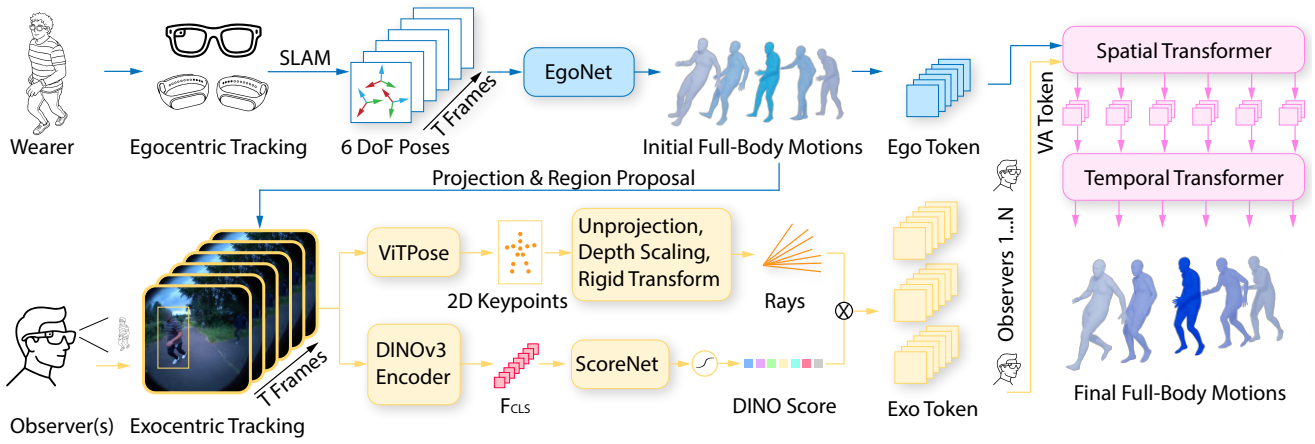

Figure 2. **Overview of EgoExoMoCap.** Given an egocentric and one or more exocentric streams from HMDs, we first roughly estimate 3D body poses from egocentric streams (EgoNet) to identify regions of interest in exocentric frames. From these, ViTPose [94]-extracted 2D keypoints are unprojected into 3D rays and softly weighted by DINOv3 [75]-based confidence scores to form Exo Tokens. A Spatial Transformer fuses Ego and Exo tokens into View-Aggregated (VA) Tokens, followed by a Temporal Transformer for smoothness to output final full-body motions.

camera extrinsics and intrinsics parameters are known for both the wearer and the observer (HMDs usually provide factory calibration information).

Outputs. Our model predicts the sequence of full-body poses of the wearer in global space. We use the SMPL body model [61], parameterized by pose parameters $\boldsymbol{\theta}_t^{w,\text{body}} \in \mathbb{R}^{J\times 6}$ (local joint rotations of J=21 joints, in 6D representation), together with root joint global orientation $\boldsymbol{\theta}_t^{w,\text{root}} \in \mathbb{R}^6$ and position $\mathbf{p}_t^{w,\text{root}} \in \mathbb{R}^3$. Note that the model does not predict SMPL shape parameters, but robustly handles different shapes as input – either subject-specific [12] or corresponding to the SMPL mean identity [100]. We obtain 3D joint positions at each timestamp t via forward kinematics.

3.2. Method Overview

Figure 2 provides an overview of our framework. Given the wearer's egocentric input streams, we employ a temporal network (EgoNet) to coarsely estimate full-body poses and project them onto the observer's exocentric images to localize the wearer and produce region proposals (Section 3.3). Within this region, we extract 2D body keypoints [94]. To account for different exo views (multiple observers) and observers' head motion, we lift 2D keypoints to 3D rays and transform them into the wearer's egocentric coordinate frame (Section 3.4). Since 2D keypoints (and therefore rays) are unreliable under occlusion, we leverage DINOv3 features [75] to associate keypoint predictions with learned confidence scores (Section 3.5). The resulting *exo* tokens are combined with *ego* ones (*i.e.*, egocentric input streams and EgoNet output) and processed with spatial and temporal transformers to predict the final body poses (Section 3.6).

3.3. Ego-Guided Wearer Localization

Egocentric signal encoding. At each timestep t, we encode the wearer's head position $\mathbf{p}_t^{w,\text{head}}$ and orientation $\boldsymbol{\theta}_t^{w,\text{head}}$, along with their velocities, into a 18D vector. When wrist tracking is available, we additionally incorporate wrist positions $\mathbf{p}_t^{w,\text{lrw}}$ and orientations $\boldsymbol{\theta}_t^{w,\text{lrw}}$, together with their velocities, extending the representation to 54D. We apply spatial normalization as in [38] and express wrist coordinates relative to the head, removing global position dependence; we also temporally normalize the head position at each frame $t > 1$ relative to the first frame (t=1), factoring out absolute starting location. After normalization, we include a 6D relative displacement encoding – expressing the head displacement on the XY plane at time t with respect to the position at timestep 1 – yielding a 60D egocentric feature vector $\mathbf{x}_t^{\text{ego}} \in \mathbb{R}^{60}$.

Coarse pose estimation and region proposal. To extract pose cues from exocentric images, we need to roughly localize the wearer in them. We observe that common bounding box detectors [87, 94] do not work well in scenarios with strong occlusions and viewpoint changes. Therefore we train an ego-only temporal network **EgoNet** that takes $\mathbf{x}_t^{\text{ego}}$ as input and predicts initial SMPL pose parameters (both local and global). EgoNet linearly projects the egocentric signals to a 512D hidden state with a sinusoidal temporal positional encoding [84], applies a single MLP-Mixer block [81] consisting of a token-mixing MLP across the time dimension for temporal dependencies followed by a channel-mixing MLP across features, and uses two MLP heads to predict global orientation and local body poses.

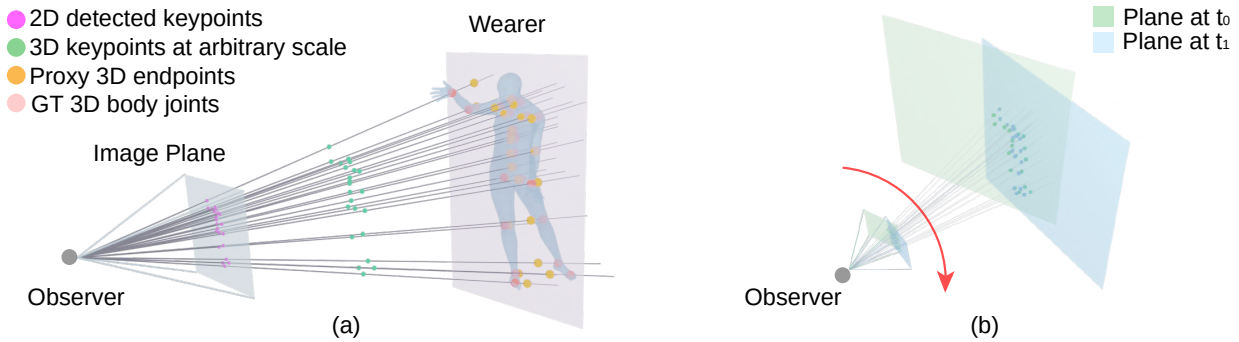

Figure 3. **Ray-based representation.** (a) We unproject 2D keypoints into 3D rays that can intersect planes at different depths (green, yellow). By using the wearer-observer distance, we obtain the correct depth and correctly scaled 3D endpoints (yellow). (b) Leveraging the observer's camera extrinsics ensures head-rotation invariance: proxy endpoints are the same if the observer moves but the wearer is still.

While coarse, these estimates provide a reasonable initialization. From them, we compute 3D joint positions in global space and project them on the exocentric image via the observer's camera parameters. The bounding box computed from the projected joints, expanded by a fixed margin, defines a region of interest. Subsequent exocentric processing – 2D keypoints (Section 3.4) and DINOv3 features (Section 3.5) – operates within this region, ensuring that the wearer is consistently tracked across frames.

3.4. Exocentric Ray-based Pose Canonicalization

Leveraging the bounding boxes obtained via EgoNet poses, we run ViTPose [94] to detect $K = 13$ body 2D keypoints, representing the wearer's 2D pose in the observer's image. Directly using these 2D coordinates as network input would inherently entangle the observer's viewpoint and camera parameters, limiting cross-scenario generalization. To reduce this camera dependence, we adopt a ray-based geometric representation, following camera-aware 3D pose formulations [17, 104]. Inspired by LAMP [97], which lifts 2D keypoints into 3D exocentric ray clouds using known HMD poses and calibration, we further extend this representation and develop a wearer-conditioned ray formulation for ego-exo fusion: exocentric keypoints are lifted with the observer pose, scaled by the observer–wearer head distance, and canonicalized in the wearer's head-local frame before being fused with continuous egocentric tracking signals, as illustrated in Fig. 3.

From 2D detections to 3D rays. Given a detected 2D keypoint $\mathbf{x}_{t,j}$ at time t, we utilize the observer's camera intrinsics $\mathbf{K}_{obs}$ and extrinsics rotation $\mathbf{R}_{obs}$ to unproject it into a 3D ray in global coordinates, then normalized into a unit vector (representing a direction):

$$\hat{\mathbf{d}}_{t,j} = \frac{\mathbf{R}_{obs}\mathbf{K}_{obs}^{-1}\mathbf{x}_{t,j}}{\|\mathbf{R}_{obs}\mathbf{K}_{obs}^{-1}\mathbf{x}_{t,j}\|} \tag{1}$$

Directly using rays in the observer camera frame would not work well in our scenario: the observer's fast and frequent head movements can cause severe high-frequency rotational variance, even when the wearer remains still. By leveraging $\mathbf{R}_{obs}$ to decouple ray and observer's ego-motion, $\hat{\mathbf{d}}_{t,j}$ provides a stable, rotation-invariant directional anchor.

Depth scaling. $\hat{\mathbf{d}}_{t,j}$ discards the spatial distance between the two subjects, which is critical for resolving scale ambiguities. We therefore scale $\hat{\mathbf{d}}_{t,j}$ by the Euclidean distance between observer and wearer head positions:

$$\tilde{\mathbf{d}}_{t,j} = \hat{\mathbf{d}}_{t,j} \cdot \|\mathbf{p}_t^{w,\text{head}} - \mathbf{p}_t^{o,\text{head}}\| \tag{2}$$

We then anchor this scaled vector to the observer's global head position to compute a proxy 3D endpoint position:

$$\mathbf{e}_{t,j} = \tilde{\mathbf{d}}_{t,j} + \mathbf{p}_t^{o,\text{head}} \tag{3}$$

Geometrically, $\mathbf{e}_{t,j}$ lies precisely along the observer's line of sight, at a depth proportional to the inter-person distance.

Ego-space canonicalization. Keeping these endpoints in the global frame makes the representation dependent on the wearer's global position and orientation: the same body pose performed at different locations would result in different representations, affecting generalization. We ensure a canonical representation by transforming each endpoint into the wearer's head-local coordinate frame:

$$\mathbf{r}_{t,j} = \mathbf{R}_{w,head}^{-1}(\mathbf{e}_{t,j} - \mathbf{p}_t^{w,\text{head}}) \tag{4}$$

In this canonical frame, the relationship between proxy endpoints and ground-truth body pose is invariant to the wearer's global trajectory and camera rotation. The concatenated representation $\mathbf{r}_t = [\mathbf{r}_{t,1}, \ldots, \mathbf{r}_{t,K}] \in \mathbb{R}^{3K}$ serves as our robust geometric exocentric feature. In summary, this representation accounts for wearer–observer distance, factors out observer camera rotations, and removes dependence on the wearer's absolute global position and orientation. We validate these design choices in Section 4.3.

3.5. Learned Visibility Gating

The reliability of the exocentric signal varies considerably over time: the wearer may be partially occluded, leave the observer's field of view, or be detected with poor accuracy at certain keypoints. Feeding noisy or absent rays into the fusion network without modulation would corrupt the prediction. Recent ray-based formulations (*e.g.*, LAMP [97]) attach 2D detector confidences to the lifted rays. However, we find these confidences are not always reliable as visibility estimates under occlusion: erroneous keypoints may still receive high scores. To address this, we learn a soft per-joint gating mechanism driven by the global semantic context of the observer's image. Specifically, from the same cropped region used for keypoint detection, we extract a DINOv3 CLS token $\mathbf{F}_{\text{CLS}} \in \mathbb{R}^{768}$ and map it to K per-joint confidence scores through **ScoreNet**, which consists of a two-layer MLP ($768 \to 512 \to K$), followed by a sigmoid function σ:

$$\mathbf{w} = \sigma(\text{MLP}(\mathbf{F}_{\text{CLS}})) \in [0,1]^K, \tag{5}$$

Each canonicalized ray $\mathbf{r}_{t,j}$ from Section 3.4 is then element-wise scaled by its corresponding confidence:

$$\hat{\mathbf{r}}_{t,j} = w_{t,j} \cdot \mathbf{r}_{t,j}, \quad j = 1, \ldots, K. \tag{6}$$

When the wearer is not well observable, the gate automatically suppresses the exocentric signal, encouraging the network to fall back on egocentric tracking.

We observed that a learned gating function can capture richer semantics from the holistic image context: for instance, it can recognize when the wearer is behind furniture even if the 2D detector still produces high-confidence but erroneous detections (Section 4.3).

3.6. Ego-Exo View Aggregation

Token construction. The egocentric signal and the coarse pose predicted by EgoNet (Section 3.3) are concatenated to form the Ego Token, which is projected to a 512-dimensional embedding. The gated exocentric rays $\hat{\mathbf{r}}_{t,j}$ from Section 3.5 are likewise projected to the same dimension, yielding the Exo Token. Each frame thus produces a pair of tokens encoding complementary information: the Ego Token carries accurate but spatially sparse device tracking together with a full-body prior, while the Exo Token provides visually grounded full-body geometry.

Spatial fusion. Per-frame Ego and Exo Tokens are arranged into a short sequence $[\mathbf{e}_t, \mathbf{o}_t]$ and processed by a Spatial Transformer Encoder. Self-attention allows the ego token to selectively attend to and absorb relevant information from the exocentric observation. We retain only the ego token's output $\mathbf{e}_t$ as the frame-level fused feature, enforcing an egocentric inductive bias: the egocentric signal serves as the primary representation, and the exocentric observation acts as an auxiliary enhancement. This ensures that even when the exocentric signal is entirely suppressed by the confidence gate (Section 3.5), the network can still produce a reasonable prediction from the egocentric tracking alone. This design also scales to N observers seamlessly by simply extending the token sequence to $[\mathbf{e}_t, \mathbf{o}_t^1, \ldots, \mathbf{o}_t^N]$ without any architectural change.

Temporal modeling and decoding. The fused features from all frames $\{\tilde{\mathbf{e}}_t\}_{t=1}^T$, augmented with learnable temporal positional embeddings, are passed through a Temporal Transformer Encoder that performs bidirectional self-attention over the full window of T=96 frames, capturing long-range motion dynamics. Each output frame is independently decoded by two lightweight MLP heads: one producing the root orientation $\hat{\boldsymbol{\theta}}_t^{w,\text{root}} \in \mathbb{R}^6$, and the other producing J=21 body joint rotations $\hat{\boldsymbol{\theta}}_t^{w,\text{body}} \in \mathbb{R}^{126}$, both in 6D rotation representation. The root translation is not predicted by the network; instead, we recover it analytically: given the known head world position $\mathbf{p}_t^{w,\text{head}}$ from the HMD and the predicted joint rotations $\hat{\boldsymbol{\theta}}_t$, we compute the head-to-root offset via forward kinematics and subtract it: $\mathbf{p}_t^{w,\text{root}} = \mathbf{p}_t^{w,\text{head}} - \text{FK}_{\text{head}}(\hat{\boldsymbol{\theta}}_t)$, where FK_{head} returns the head joint position relative to the root in the body's local frame [38, 40, 64].

3.7. Training

Two-stage training. We train the pipeline in two stages. In the first stage, EgoNet (Section 3.3) is trained using only the egocentric signal, without any exocentric input. In the second stage, we freeze the EgoNet's coarse predictions and train the full ego-exo fusion network end-to-end, including the ray embedding, DINO confidence gate, Spatial Transformer, and Temporal Transformer. The DINOv3 backbone is kept frozen throughout; only the gating MLP is trained.

Loss function. The training objective combines three complementary L1 losses:

$$\mathcal{L} = \lambda_{\text{orient}}\,\mathcal{L}_{\text{orient}} + \lambda_{\text{rot}}\,\mathcal{L}_{\text{rot}} + \lambda_{\text{pos}}\,\mathcal{L}_{\text{pos}}, \tag{7}$$

where $\mathcal{L}_{\text{orient}}$ supervises the root orientation in 6D rotation space, $\mathcal{L}_{\text{rot}}$ supervises the J=21 body joint rotations in 6D space, and $\mathcal{L}_{\text{pos}}$ penalizes the L1 error on 3D joint positions obtained via SMPL forward kinematics from the predicted rotations. The rotation and position losses are complementary: rotation loss provides direct supervision in the rotation space, while position loss propagates gradients through the kinematic chain and penalizes error accumulation at distal joints. We set $\lambda_{\text{orient}} = 0.02$, $\lambda_{\text{rot}} = 1.0$, and $\lambda_{\text{pos}} = 1.0$.

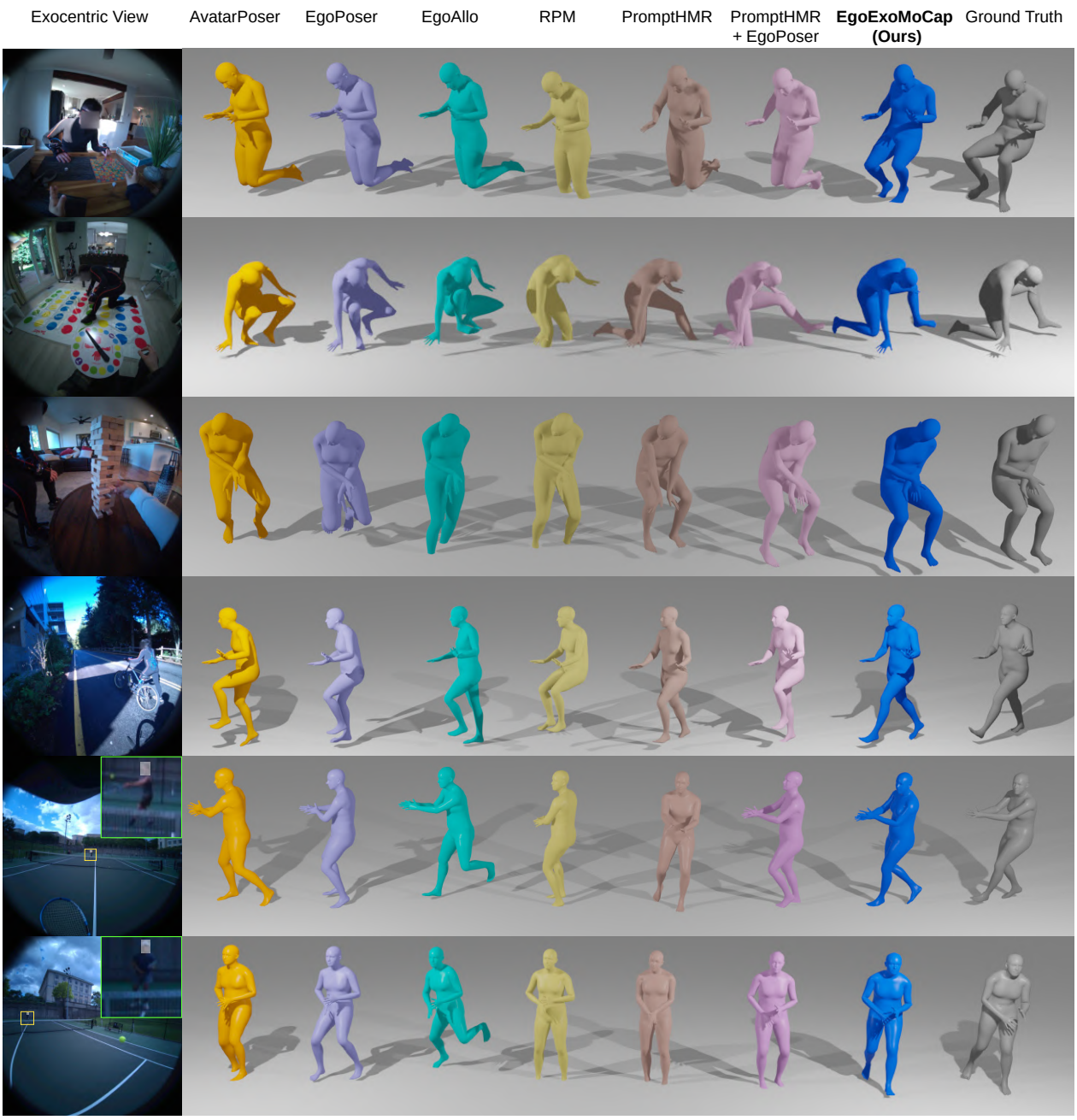

Figure 4. Qualitative comparison of EgoExoMoCap versus baselines. The first four rows compare diverse activities from Nymeria, while the last two rows are drawn from the EgoHumans dataset and show two players playing tennis together.

4. Experiments

4.1. Experimental Protocol

Datasets. We train and test on the large-scale **Nymeria** [64] dataset, which pairs multi-modal egocentric streams from *Aria* glasses [26] with ground-truth full-body motions obtained with the Xsens inertial system [2]. We use the SMPL representation provided by NymeriaPlus [21], which was retargeted from the original ground-truth motion. Nymeria features 300 hours of diverse indoor and outdoor activities. It also provides 6DoF wrist trajectories obtained by *miniAria* wristbands. We randomly split the dataset into 80% training and 20% test sets, ensuring no test-train overlap across subjects.

To assess generalization, we additionally perform cross-

dataset evaluation on **EgoHumans** [44], an outdoor multi-person dataset also captured with Aria glasses, featuring dynamic interactive activities such as fencing, basketball, and badminton. Ground-truth motions are obtained using a multi-view camera system, limiting the capture area but removing the requirement for subjects to wear an inertial suit as in Nymeria. Since EgoHumans sequences do not provide wrist tracking signals, we use the synthesized 6DoF tracking signals from ground-truth body parameters during evaluation.

Sensor setup. We evaluate our method under two tracking configurations based on the **wearer**'s device setup: (i) **3-point tracking**, where the wearer uses glasses plus two wrist-worn devices (or controllers) providing wrist trajectories; and (ii) **1-point tracking**, where the wearer uses only the glasses, providing head trajectory alone. In both setups, the **observer** wears only the glasses.

Metrics. We consider both positional and physical-plausibility evaluation metrics [12, 38, 40, 100]: Mean Per-Joint Position Error (**MPJPE**, cm), alongside Upper-body (**U-PE**) and Lower-body (**L-PE**) position errors; Mean Per-Joint Velocity Error (**MPJVE**, cm/s), measuring the difference between predicted and ground-truth joint velocities; Motion **Jitter** (10^2 m/s^3), calculated as the mean magnitude of the third derivative of position (jerk).

Baselines. We benchmark our method against several state-of-the-art pose estimation approaches, considering both egocentric and exocentric ones, using their open-sourced code. As egocentric baselines, we include regression models, such as AvatarPoser [38] and EgoPoser [40], alongside Diffusion-based frameworks like EgoAllo [100] and RPM [12]. We retrain and evaluate them under 1-point and 3-point settings on the same train-test split as ours. We partition the full sequence into non-overlapping T-frame segments during testing. Note that, for the experiments, we improved baselines [38] and [40] by predicting a sequence instead of just the last frame, which provides better results than their online tracking designs.

We consider PromptHMR [92] as an exocentric baseline. As its training code is not available, for a fairer comparison, we adapt the original model by freezing it and adding to it a Transformer-based output layer, trained on Nymeria. We feed PromptHMR with ground-truth observer camera parameters and ground-truth wearer global position and orientation. We report both the original PromptHMR output and a finetuned variant, PromptHMR-Finetuned, where the original model is frozen and followed by a Transformer-based output layer trained on Nymeria. Furthermore, we implement a custom baseline that integrates the outputs of PromptHMR and EgoPoser through an additional Transformer network. Following [12], our evaluation setup accounts for individual user dimensions, rather than assuming a universal average body shape [25, 38], using SMPL identity ground-truth parameters.

4.2. Results

Table 1 summarizes the quantitative results on Nymeria. Visual comparisons are provided in Figure 4. Across all metrics and both tracking setups, our method outperforms the baselines. The only exception is Jitter, where RPM achieves the best performance; we hypothesize this is due to its PCAF module [12], which balances smoothness with a potentially minor adherence to input signals.

In general, egocentric methods can return plausible motions (often exhibiting low jitter) but cannot faithfully reconstruct invisible parts (*e.g.*, kneeling or sitting poses). PromptHMR, even if fed with ground-truth root position and rotation, struggles with extreme occlusions (*e.g.*, intervals in which the wearer is not visible at all) and observer's HMD camera motion. The combination of PromptHMR and EgoPoser in our ego-exo baseline shows the benefits of combining both sources of information, reporting low MPJPE for both upper and lower body. However, its "naive" fusion of ego- and exocentric features results in decreased accuracy and reduced smoothness (higher MPJVE and Jitter). Early fusion of 2D and 3D features, as performed in our method, helps increase robustness: the estimate does not rely excessively on the exocentric input when this is noisy and unreliable (see *e.g.* first row in Figure 4, where naive fusion does not recover from the inaccurate PromptHMR estimate).

Tab. 2 reports quantitative results on EgoHumans, confirming the trend observed in Nymeria. We observe how scaling our approach to multiple observers can bring additional benefit. Fig. 5 shows a typical scenario: single observers may only have partial or far-away views of the wearer, leading to suboptimal pose estimates; combining their views, accuracy significantly improves. We also compare against a baseline using multi-view triangulated keypoints as exo tokens instead of fused 2D and DINO features: the baseline does not adequately account for confidence scores associated with different views and exhibits worse results. These results suggest the potential of distributed HMD-based setups for in-the-wild motion capture.

4.3. Ablation Studies

We ablate the components of our approach on the Nymeria test set in Table 3.

Ego+exo, wearer localization. Removing either the egocentric input signals (w/o Ego) or the exocentric images (w/o Exo) leads, as expected, to a significant metrics drop. Using a bounding box detector (YOLO [87]) instead of EgoNet predictions (w/o Ego BBX) worsens metrics. Detectors like YOLO typically struggle in the presence of strong body occlusions.

Table 1. **Quantitative evaluation on Nymeria.** We compare ego- and exocentric methods under the 1-point and 3-point setups. Best results highlighted in **boldface**.

Methods	Input Modality	Three-Point Tracking					One-Point Tracking				
		MPJPE	Upper	Lower	MPJVE	Jitter	MPJPE	Upper	Lower	MPJVE	Jitter
AvatarPoser [38]	Ego	8.16	3.82	14.43	13.62	2.20	12.38	9.22	16.94	18.90	4.07
EgoPoser [40]	Ego	7.74	3.44	13.96	13.94	2.43	11.86	9.08	15.87	20.26	3.43
EgoAllo [100]	Ego	8.77	3.18	16.84	12.81	1.72	12.53	8.27	18.68	20.17	1.96
RPM [12]	Ego	12.19	3.50	24.74	17.62	**1.29**	16.83	9.39	27.58	18.93	**1.15**
PromptHMR [92]	Exo	13.48	8.88	20.13	26.78	5.17	13.48	8.88	20.13	26.78	5.17
PromptHMR-Finetuned	Exo	10.65	7.63	15.01	23.55	4.98	10.65	7.63	15.01	23.55	4.98
PromptHMR+EgoPoser	EgoExo	6.47	3.41	10.90	17.49	3.08	9.03	6.74	12.34	17.52	3.17
EgoExoMoCap (Ours)	EgoExo	**5.72**	**2.72**	**10.05**	**12.77**	2.16	**8.28**	**6.10**	**11.44**	**17.23**	2.47

Table 2. **Quantitative evaluation on EgoHumans.** We compare ego- and exocentric methods under the 1-point and 3-point setups. We also evaluate our method in the multi-observer setup, on the EgoHumans subset(*) providing multi-observer streams. Best results highlighted in **boldface**.

Methods	Input Modality	Three-Point Tracking					One-Point Tracking				
		MPJPE	Upper	Lower	MPJVE	Jitter	MPJPE	Upper	Lower	MPJVE	Jitter
AvatarPoser [38]	Ego	9.60	4.34	17.18	32.94	3.08	14.17	10.38	19.64	54.60	3.84
EgoPoser [40]	Ego	9.03	3.87	16.48	33.12	3.55	13.96	10.49	18.96	57.04	5.05
EgoAllo [100]	Ego	10.19	3.82	19.39	36.82	1.94	14.38	11.27	18.87	51.77	1.85
RPM [12]	Ego	10.69	4.79	19.21	31.87	**0.85**	13.70	11.03	17.56	43.52	**1.14**
PromptHMR [92]	Exo	18.40	12.91	26.34	58.63	10.50	18.40	12.91	26.34	58.63	10.50
PromptHMR-Finetuned	Exo	18.15	11.79	27.34	46.84	5.17	18.15	11.79	27.34	46.84	5.17
PromptHMR+EgoPoser	EgoExo	8.62	4.24	14.94	34.01	2.79	13.27	8.96	19.50	41.88	2.71
EgoExoMoCap (Ours)	EgoExo	**7.62**	**3.45**	**13.64**	**30.94**	1.83	**11.54**	**8.18**	**16.39**	**41.04**	2.07
EgoExo-single-observer*	EgoExo	8.80	7.00	11.40	47.23	3.59	13.70	12.35	15.64	50.32	3.01
EgoExo-triangulation*	Ego-Multi-Exo	8.49	6.58	11.25	44.38	3.23	13.34	11.83	15.52	49.04	2.90
EgoExo-multi-observer* (Ours)	Ego-Multi-Exo	**7.11**	**5.89**	**8.87**	**37.31**	**2.58**	**10.48**	**9.59**	**11.77**	**45.50**	**2.83**

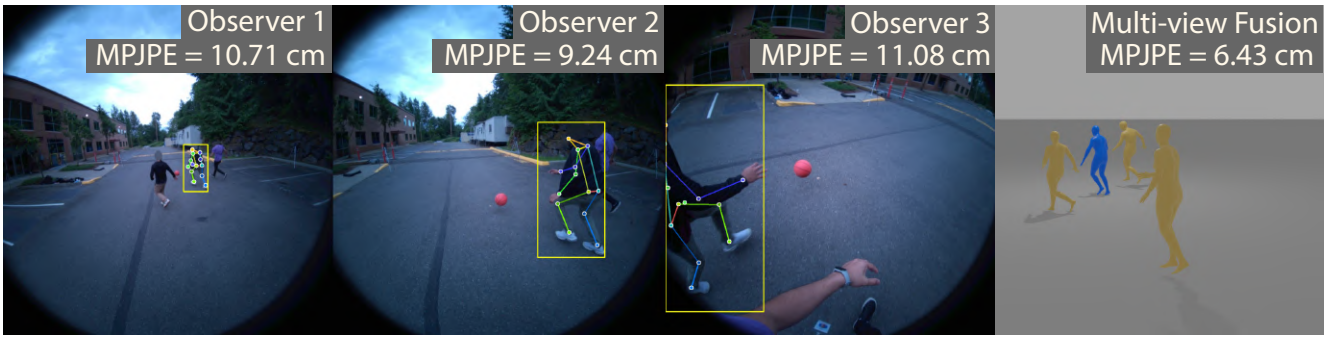

Figure 5. **Multi-observer fusion.** Egocentric tracking with a single exocentric view (Observers 1-3) may struggle with partial observations and severe body truncation, yielding MPJPEs around 9 to 11 cm. Our approach effectively aggregates these arbitrary views, dropping the final 3D pose error to 6.43 cm (right).

Ray-based pose representation. Omitting depth scaling (w/o Depth Scaling) degrades spatial reasoning and therefore accuracy. Not converting rays from the observer's frame to the wearer's frame (Ray in World-Space, Ray in Exo-Space) also negatively impacts performance. This transformation helps factor out the observer's head motion, ensuring better generalization.

Learned gating. To evaluate the effectiveness of learned

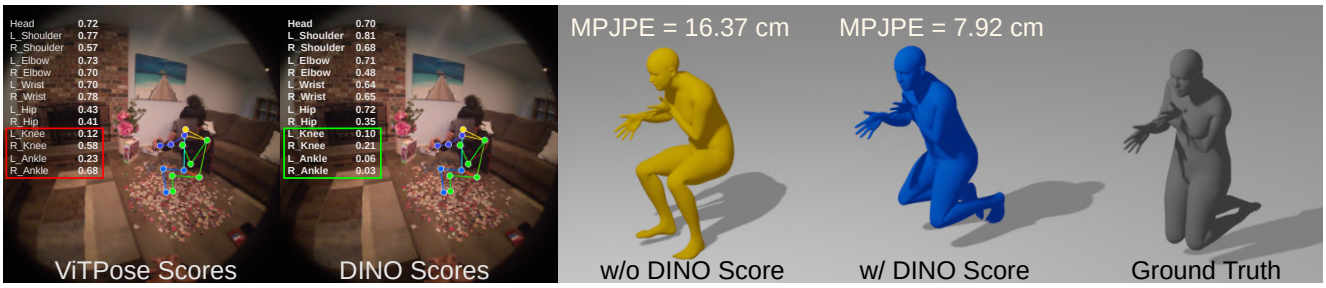

Figure 6. **ViTPose vs DINOv3 scores.** Learning DINO-based scores significantly improves robustness against 2D detections, typically observed under occlusion.

Table 3. Ablation study on Nymeria. Best results highlighted in **boldface**.

	Methods	MPJPE	Upper	Lower	MPJVE	Jitter
Default	EgoExoMoCap	**5.72**	**2.72**	**10.05**	**12.77**	**2.16**
Ego-Exo	w/o Ego	11.45	8.00	16.43	34.40	6.55
	w/o Exo	7.53	3.38	13.53	13.78	2.19
	w/o Ego BBX	6.43	3.05	11.31	15.54	2.80
Learned Gating	w/o DINO Score	6.30	3.00	11.08	14.42	2.38
	w/ ViT Score	6.29	2.92	11.16	13.91	2.23
	w/ Masking	6.32	2.99	11.14	15.31	2.68
Ray Representation	w/ Ray in World-Space	6.99	3.15	12.54	13.87	2.44
	w/ Ray in Exo-Space	6.14	2.92	10.78	13.22	2.21
	w/o Depth Scaling	6.26	2.99	10.97	13.26	2.19
	w/o Lifting	6.26	2.97	11.00	13.81	2.33

DINO scores, we remove them, keeping the original rays without gating (w/o DINO Score), replace them with ViT-provided confidences (w/ ViT score), and mask them out when ViT-provided confidence scores are smaller than 0.2 (w/ Masking). As Fig. 6 exemplifies, DINO scores help in the presence of noisy 2D keypoints, *e.g.* when the subject is occluded by furniture (recognized as not belonging to the body by DINO). Using ViT Score slightly improves some metrics over no gating, but remains worse than our learned DINO-based gating, especially for lower-body accuracy, given its unreliability in occluded scenarios.

Robustness to EgoNet. To further assess the dependence on EgoNet, we perturb its estimated joint positions with Gaussian noise of $\sigma = 1/2/5/10$ cm: the final full-body MPJPE increases by only 0.005/0.02/0.14/0.54 cm, respectively, indicating graceful degradation and limited sensitivity to EgoNet's coarse pose estimates.

4.4. Discussion

Our approach focuses on portable, lightweight human motion capture using a distributed HMD setup. We assume cameras are calibrated and synchronized by recent Aria tooling [1]. While we focus on in-the-wild ground-truth motion acquisition rather than online and real-time tracking, exploring real-time applications is an exciting direction for future work when hardware and tooling can support it.

Currently, we mainly focus on motion reconstruction and assume subject shape is provided when calculating the joint positions; it could potentially be estimated leveraging images or sensor-based calibrations [69]. We observed failures when the wearer is largely occluded by another subject – since the scenario may confuse DINO-based visibility scores. Long out-of-view intervals or persistently unreliable exo signals (*e.g.*, under heavy occlusion) might lower accuracy, especially for the lower body. We found learned gating helpful in assessing signal reliability in these scenarios: performance falls towards the ego-only method, leading to still plausible (but less faithful) motions. Better modeling of physical plausibility (foot-ground contact, body-self penetrations), together with better incorporation of scene context, could further enhance motion realism.

5. Conclusion

We presented EgoExoMoCap, a novel scalable approach effectively combining egocentric and exocentric tracking for in-the-wild human motion capture. Unlike traditional motion capture systems that rely on extensive hardware and complex setups, our method only leverages a set of HMDs, proposing a lightweight distributed solution for flexible and unobtrusive tracking in real-world environments. EgoExoMoCap requires just two subjects, each wearing a pair of glasses, and can naturally scale to multi-subject setups. Extensive experiments on two in-the-wild datasets show that the approach performs favorably with respect to egocentric and exocentric baselines, handling challenging scenarios like occlusions and out-of-view motions.

Future work should explore the combination of further multi-modal streams (*e.g.*, stereo cameras mounted on HMDs [26] or multiple body-worn sensors [9]) to provide richer motion cues. Tracklet association mechanisms from exocentric people tracking systems [97] could further extend EgoExoMoCap to crowded scenarios with non-HMD participants. Furthermore, leveraging the 3D scene reconstruction capabilities of HMDs [64] could lead to more accurate reconstruction of human-scene interactions.

References

[1] Aria gen2 glasses. https://ai.meta.com/blog/aria-gen-2-research-glasses-under-the-hood-reality-labs/. Accessed: 2026-02-28. 2, 10

[2] Movella xsens. https://www.movella.com/motion-capture/xsens-link-specifications. Accessed: 2026-02-28. 7

[3] Microsoft HoloLens 2. https://www.microsoft.com/hololens, 2019. Accessed: 2026-02-28. 3

[4] Meta Quest 3. https://www.meta.com/quest/quest-3/, 2023. Accessed: 2026-02-28. 3

[5] Apple Vision Pro. https://www.apple.com/apple-vision-pro/, 2024. Accessed: 2026-02-28. 3

[6] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13253–13262, 2022. 3

[7] Sadegh Aliakbarian, Fatemeh Saleh, David Collier, Pashmina Cameron, and Darren Cosker. HMD-Nemo: Online 3d avatar motion generation from sparse observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9622–9631, 2023. 3

[8] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 2

[9] Rayan Armani and Christian Holz. Accurately tracking relative positions on moving trackers based on uwb ranging and inertial sensing without anchors. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 10

[10] Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. Ultra Inertial Poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[11] 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. *European Conference on Computer Vision*, 2024. 2

[12] German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky Sijia He, Cristina Palmero, Sergio Escalera, Yuting Ye, et al. From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1850–1860, 2025. 2, 3, 4, 8, 9

[13] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4221–4231, 2023. 3

[14] Xiaodong Chen, Wu Liu, Qian Bao, Xinchen Liu, Quanwei Yang, Ruoli Dai, and Tao Mei. Motion capture from inertial and vision sensors. *arXiv preprint arXiv:2407.16341*, 2024. 3

[15] Seunggeun Chi, Hyung-gun Chi, Hengbo Ma, Nakul Agarwal, Faizan Siddiqui, Karthik Ramani, and Kwonjoon Lee. M2d2m: Multi-motion generation from text with discrete diffusion models. In *European conference on computer vision*, pages 18–36. Springer, 2024. 2

[16] Seunggeun Chi, Pin-Hao Huang, Enna Sachdeva, Hengbo Ma, Karthik Ramani, and Kwonjoon Lee. Estimating ego-body pose from doubly sparse egocentric video data. *Advances in neural information processing systems*, 2024. 3

[17] Hanbyel Cho, Yooshin Cho, Jaemyung Yu, and Junmo Kim. Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11169–11178, 2021. 5

[18] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020. 3

[19] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021. 3

[20] Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[21] Daniel DeTone, Federica Bogo, Eric-Tuan Le, Duncan Frost, Julian Straub, Yawar Siddiqui, Yuting Ye, Jakob Engel, Richard Newcombe, and Lingni Ma. Nymeriaplus: Enriching nymeria dataset with additional annotations and data. *arXiv preprint arXiv:2603.18496*, 2026. 7

[22] Ameya Dhamanaskar, Mariella Dimiccoli, Enric Corona, Albert Pumarola, and Francesc Moreno-Noguer. Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition*, 138:109358, 2023. 2

[23] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021. 3

[24] Kun Dong, Jian Xue, Zehai Niu, Xing Lan, Ke Lu, Qingyuan Liu, and Xiaoyu Qin. Realistic full-body motion generation from sparse tracking with state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 4024–4033. Association for Computing Machinery, 2024. 3

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 8

[26] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 3, 7, 10

[27] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 153–163, 2024. 3

[28] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Reconstructing and tracking humans with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

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

[30] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. HMD^2: Environment-aware motion generation from single egocentric head-mounted device. In *International Conference on 3D Vision (3DV)*, 2025. 2, 3

[31] Dorian F Henning, Tristan Laidlow, and Stefan Leutenegger. BodySLAM: joint camera localisation, mapping, and human motion tracking. In *European Conference on Computer Vision*, pages 656–673, 2022. 3

[32] Dorian F Henning, Christopher Choi, Simon Schaefer, and Stefan Leutenegger. BodySLAM++: Fast and tightly-coupled visual-inertial camera and human motion tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3781–3788. IEEE, 2023. 3

[33] 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. In *Advances in Neural Information Processing Systems*, 2024. 3

[34] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. EgoLM: Multi-Modal Language Model of Egocentric Motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[35] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 1

[36] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 3

[37] Andela Ilic, Jiaxi Jiang, Paul Streli, Xintong Liu, and Christian Holz. Human motion capture from loose and sparse inertial sensors with garment-aware diffusion models. *arXiv preprint arXiv:2506.15290*, 2025. 3

[38] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 4, 6, 8, 9

[39] Jiaxi Jiang, Paul Streli, Xuejing Luo, Christoph Gebhardt, and Christian Holz. Manikin: biomechanically accurate neural inverse kinematics for human motion estimation. In *European Conference on Computer Vision*, pages 128–146. Springer, 2024. 2

[40] 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*, 2024. 2, 3, 6, 8, 9

[41] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[42] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 3

[43] 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 *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025. 1

[44] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. EgoHumans: An Egocentric 3D Multi-Human Benchmark. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 8

[45] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: a video dataset for in-the-wild close human interactions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024. 3

[46] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 3

[47] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 11127–11137, 2021. 3

[48] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 3

[49] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and camera motion estimation from in-the-wild videos. In *International Conference on 3D Vision*, pages 397–408, 2024. 3

[50] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 3

[51] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2024. 3

[52] Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. QuestEnvSim: Environment-Aware Simulated Motion Tracking from Sparse Sensors. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[53] 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. 3

[54] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2

[55] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: Generative Models for Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[56] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606, 2022. 3

[57] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021. 3

[58] Bonan Liu, Handi Yin, Manuel Kaufmann, Jinhao He, Sammy Christen, Jie Song, and Pan Hui. EgoHDM: An Online Egocentric-Inertial Human Motion Capture, Localization, and Dense Mapping System. *ACM Trans. Graph.*, 43(6), 2024. 3

[59] Huakun Liu, Hiroki Ota, Xin Wei, Yutaro Hirao, Monica Perusquia-Hernandez, Hideaki Uchiyama, and Kiyoshi Kiyokawa. Umotion: Uncertainty-driven human motion estimation from inertial and ultra-wideband units. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7085–7094, 2025. 3

[60] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. In *2021 international conference on 3D vision (3DV)*, pages 930–939. IEEE, 2021. 2

[61] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

[62] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[63] Zhengyi Luo, Jinkun Cao, Rawal Khirodkar, Alexander Winkler, Kris Kitani, and Weipeng Xu. Real-time simulated avatar from head-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 571–581, 2024. 2

[64] 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, 2024. 2, 3, 6, 7, 10

[65] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. IMUPoser: Full-Body Pose Estimation Using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. 3

[66] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single RGB image. In *European Conference on Computer Vision*, pages 752–768, 2020. 3

[67] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. 3

[68] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 3

[69] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements. *IEEE transactions on visualization and computer graphics*, 25 (5):1887–1897, 2019. 10

[70] 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. 1

[71] 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. 2, 3

[72] Jingyu Shi, Rahul Jain, Seunggeun Chi, Hyungjun Doh, Hyung-gun Chi, Alexander J Quinn, and Karthik Ramani. Caring-ai: Towards authoring context-aware augmented reality instruction through generative artificial intelligence. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–23, 2025. 2

[73] Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Yinghui Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. *arXiv preprint arXiv:2602.10106*, 2026. 1

[74] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 2, 3

[75] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2, 4

[76] Sebastian Starke, Paul Starke, Nicky He, Taku Komura, and Yuting Ye. Categorical codebook matching for embodied character controllers. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024. 3

[77] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

[78] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems*, 34: 16558–16569, 2021. 3

[79] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[80] Deepak Tolani, Ambarish Goswami, and Norman I Badler. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical models*, 62(5):353–388, 2000. 2

[81] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34: 24261–24272, 2021. 4

[82] Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. MultiPhys: Multi-person physics-aware 3D motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2331–2340, 2024. 2

[83] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C. Karen Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2513–2523, 2024. 3

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[85] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, pages 349–360. Wiley Online Library, 2017. 3

[86] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 3

[87] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 4, 8

[88] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13157–13166, 2022. 2

[89] Wenjia Wang, Liang Pan, Huaijin Pi, Yuke Lou, Xuqian Ren, Yifan Wu, Zhouyingcheng Liao, Lei Yang, Rishabh Dabral, Christian Theobalt, and Taku. Komura. EmbodMocap: In-the-Wild 4D Human-Scene Reconstruction for Embodied Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 3

[90] Yufu Wang and Kostas Daniilidis. ReFit: Recurrent fitting network for 3D human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023. 3

[91] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487, 2024. 2, 3

[92] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 2, 8, 9

[93] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2

[94] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. VITPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 4, 5

[95] Ying Xue, Jiaxi Jiang, Rayan Armani, Dominik Hollidt, Yi-Chi Liao, and Christian Holz. Group inertial poser: Multi-person pose and global translation from sparse inertial sensors and ultra-wideband ranging. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 24910–24921, 2025. 2, 3

[96] Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, et al. Reli11d: A comprehensive multimodal human motion dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2250–2262, 2024. 3

[97] Nan Yang, Julian Straub, Fan Zhang, Richard Newcombe, Jakob Engel, and Lingni Ma. LAMP: Localization aware multi-camera people tracking in metric 3D world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3, 5, 6, 10

[98] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos, 2025. 1

[99] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023. 3

[100] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7072–7084, 2025. 3, 4, 8, 9

[101] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In *SIGGRAPH 2024 Conference Papers*, 2024. 3

[102] Xinyu Yi, Shaohua Pan, and Feng Xu. Improving global motion estimation in sparse imu-based motion capture with physics. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025. 3

[103] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11038–11049, 2022. 3

[104] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13116–13125, 2022. 5

[105] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 3

[106] He Zhang, Shenghao Ren, Haolei Yuan, Jianhui Zhao, Fan Li, Shuangpeng Sun, Zhenghao Liang, Tao Yu, Qiu Shen, and Xun Cao. Mmvp: A multimodal mocap dataset with vision and pressure sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21842–21852, 2024. 3

[107] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 180–200, 2022. 2, 3

[108] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7989–8000, 2023. 2

[109] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. RoHM: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14606–14617, 2024. 3

[110] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[111] Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-space camera and human reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1216–1226, 2024. 3

[112] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2

[113] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3

[114] Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, and Yipeng Qin. Loose inertial poser: Motion capture with imu-attached loose-wear jacket. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3