

EMHI: A Multimodal Egocentric Human Motion Dataset with HMD and Body-Worn IMUs

Zhen Fan*, Peng Dai*, Zhuo Su*, Xu Gao, Zheng Lv, Jiarui Zhang,
Tianyuan Du, Guidong Wang, Yang Zhang

PICO

{fanzhen.0315, daipeng.2022, suzhuo, gaoxu.1024, lvzheng.101, zhangjiarui.zjr123,
dutianyuan, guidong.wang, zhangyang.0621}@bytedance.com

Abstract

Egocentric human pose estimation (HPE) using wearable sensors is essential for VR/AR applications. Most methods rely solely on either egocentric-view images or sparse Inertial Measurement Unit (IMU) signals, leading to inaccuracies due to self-occlusion in images or the sparseness and drift of inertial sensors. Most importantly, the lack of real-world datasets containing both modalities is a major obstacle to progress in this field. To overcome the barrier, we propose EMHI, a multimodal Egocentric human Motion dataset with Head-Mounted Display (HMD) and body-worn IMUs, with all data collected under the real VR product suite. Specifically, EMHI provides synchronized stereo images from downward-sloping cameras on the headset and IMU data from body-worn sensors, along with pose annotations in SMPL format. This dataset consists of 885 sequences captured by 58 subjects performing 39 actions, totaling about 28.5 hours of recording. We evaluate the annotations by comparing them with optical marker-based SMPL fitting results. To substantiate the reliability of our dataset, we introduce MEPoser, a new baseline method for multimodal egocentric HPE, which employs a multimodal fusion encoder, temporal feature encoder, and MLP-based regression heads. The experiments on EMHI show that MEPoser outperforms existing single-modal methods and demonstrates the value of our dataset in solving the problem of egocentric HPE. We believe the release of EMHI and the method could advance the research of egocentric HPE and expedite the practical implementation of this technology in VR/AR products.

Introduction

Egocentric human pose estimation (HPE) has gained significant attention in computer vision, driven by the demand for accurate motion tracking in immersive VR/AR environments. Unlike traditional exocentric HPE which relies on external sensors, egocentric HPE employs body-worn sensors such as egocentric cameras or sparse IMUs. Although rapid progress has been made in this field, there remain challenges in obtaining accurate full-body poses from single-modal data due to issues like 1) self-occlusion and viewpoint variations in egocentric vision; and 2) sparsity and drifting

of IMU data. Most importantly, the lack of real-world multimodal training data poses the most significant challenge.

Previous works (Rhodin et al. 2016; Xu et al. 2019) introduced egocentric datasets using experimental fisheye camera setups to capture images and annotate 3D joints. However, these setups are impractical for real VR/AR products, which need compact, lightweight designs. Synthetic datasets (Tome et al. 2019; Akada et al. 2024; Cuevas-Velasquez et al. 2024) use physics engines for egocentric image rendering but suffer a domain gap with real images due to the complexity of human motion and environments. Meanwhile, the lower body may be occluded in a vision-based setting, and some body parts may fall outside the field of view (FOV) depending on the body pose. IMU-based datasets avoid occlusion but suffer from drift over time and ill-posed problems from sparse observations. Besides, existing methods (Jiang et al. 2022a; Zheng et al. 2023; Dai et al. 2024) typically use synthetic IMU data from AMASS (Mahmood et al. 2019), which may not accurately reflect real-world noise and drift. Some datasets (Trumble et al. 2017; Huang et al. 2018) provide real 3 Degrees of Freedom (3DoF, rotation) data from XSens, while others (Dai et al. 2024) include 6DoF (rotation and position) data of head, hands, and 3DoF for lower legs, but these are small-scale and primarily used for evaluation. Recently, several large-scale multimodal datasets (Ma et al. 2024; Grauman et al. 2024) have been released, offering RGB images, upper body IMU data, and motion narrations. However, the forward-facing camera limits the egocentric view, and missing lower-body IMU signals can cause ambiguity.

Combining egocentric cameras and body-worn IMUs offers a promising multimodal solution due to their lightweight and flexible design. This configuration is also commonly found in VR scenarios. Our proposed EMHI dataset, as shown in Fig. 1, features a VR headset with two downward-sloping cameras for egocentric image capture, 6DoF head and hand tracking, and additional IMUs on an actual VR device for lower-leg 3DoF tracking. We use a markerless multi-view camera system for SMPL (Loper et al. 2015) ground truth acquisition, with accuracy and consistency refinement using IMU data, and synchronization via OptiTrack. Furthermore, we propose a new baseline method, MEPoser, integrating egocentric images and IMU data to perform real-time HPE on a standalone VR headset.

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

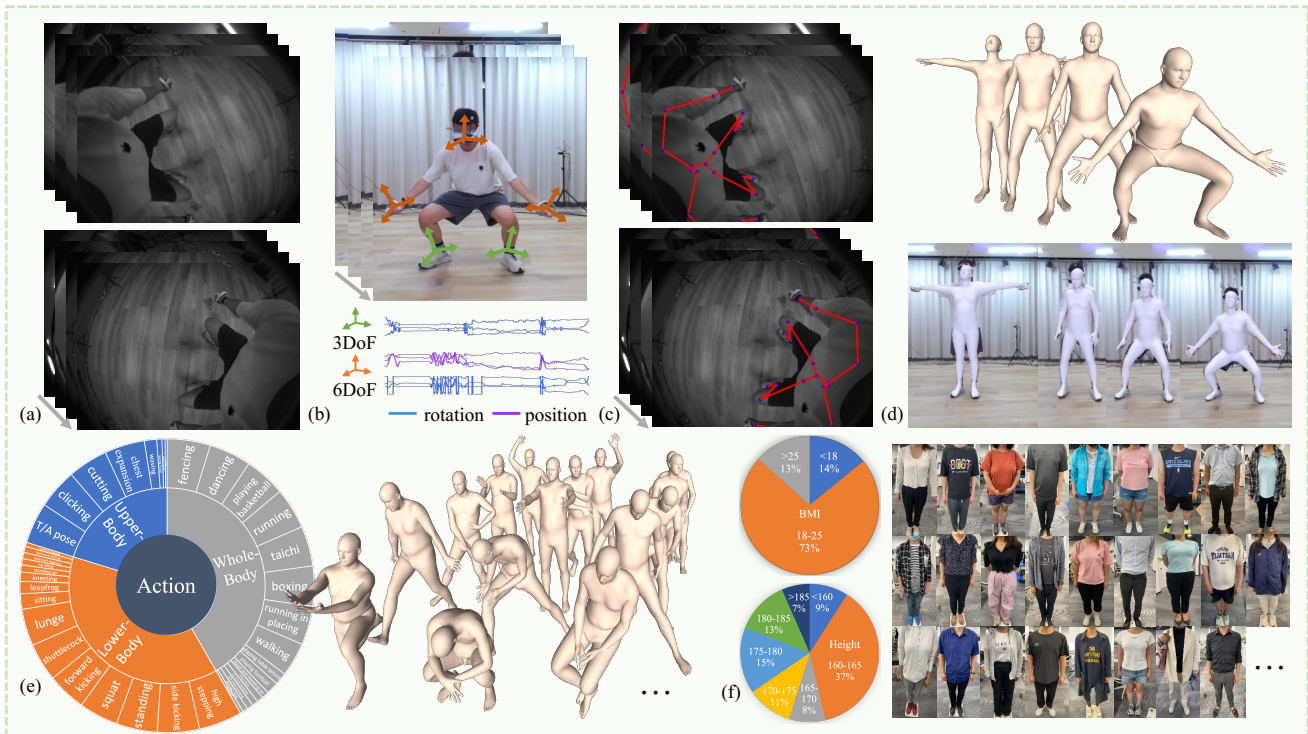


Figure 1: EMHI is a multimodal dataset that provides (a) stereo egocentric images and (b) IMU signals. The annotations include (c) 2D keypoints overlaid on egocentric images and (d) SMPL parameters in the world coordination. Each sequence is also annotated with (e) the action label, as well as (f) individual attributes such as height, BMI, and clothing descriptions.

The method employs a multimodal fusion encoder, a temporal feature encoder, and MLP-based regression heads to estimate SMPL body model parameters, effectively demonstrating the advantages of multimodal data fusion in enhancing pose accuracy and the value of our dataset. This approach paves the way for further research in egocentric HPE using multimodal inputs.

In summary, our work makes the following contributions:

- We first introduce a large-scale multimodal egocentric motion dataset EMHI on the real VR device, including stereo downward-sloping egocentric images, full-body IMU signals, and accurate human pose annotations.
- We propose a baseline method MEPoser, which employs a multimodal fusion encoder, temporal feature encoder, and MLP-based regression heads to perform real-time HPE on a standalone HMD.
- The experiment results demonstrate the rationality of our multimodal setting and the effectiveness of EMHI for addressing egocentric HPE.

Related Work

Egocentric Motion Dataset

As shown in Tab. 1, existing egocentric motion datasets can be divided into vision-based, IMU-based, and multimodal datasets depending on the input modality.

Vision-based egocentric motion datasets (Zhao et al. 2021; Wang et al. 2022, 2023; Liu et al. 2023) provide first-

person perspective images using head-mounted cameras, with the corresponding annotations of the wearer’s poses. Mo2Cap2 (Xu et al. 2019) and xR-EgoPose (Tome et al. 2019) made an early effort to build the synthetic monocular dataset with a downward-facing fisheye camera. EgoWhole-Body (Wang et al. 2024) is the latest synthetic dataset providing high-quality images and SMPL-X annotations. Ego-Cap (Rhodin et al. 2016) is a pioneer binocular dataset captured by helmet-mounted stereo cameras, containing 30K frames recorded in a lab environment. To relieve the dataset scale limitation, UnrealEgo (Akada et al. 2022) proposed a large-scale and highly realistic stereo synthetic dataset with 450K stereo views and was extended to 1.25M in UnrealEgo2 (Akada et al. 2024). SynthEgo (Cuevas-Velasquez et al. 2024) extended synthetic datasets with more identities and environments, annotated with SMPL-H for better body shape descriptions.

Sparse IMU-based datasets provide an alternative for this problem. AMASS (Mahmood et al. 2019) can provide a large-scale synthetic IMU dataset. TotalCapture (Trumble et al. 2017) and DIP-IMU (Huang et al. 2018) offered real IMU data captured by Xsens and SMPL pose annotations obtained by marker-based optical mocap system and IMU-based method (Von Marcard et al. 2017) respectively. PICO-FreeDancing (Dai et al. 2024) provided sparse IMU data with SMPL format GT fitting using OptiTrack data.

Multimodal datasets (Damen et al. 2022; Gong et al. 2023; Cha et al. 2021; Rai et al. 2021) have attracted significant










| Dataset | Device | Real/Synth | Sensor Modality | | SMPL(x) | Statistic | | |
|--------------------------------------|--|------------|-------------------------------|-----------------------------|---------|-----------|----------|--------|
| | | | Egocentric Vision | Inertial | | Actions | Subjects | Frames |
| <i>Mo²Cap²</i> |  | Synth. | Monocular Downward-Facing | - | - | 3K | 700 | 530K |
| EgoPW |  | Real. | Monocular Downward-Facing | - | - | 20 | 10 | 318K |
| EgoCap |  | Real | Binocular Downward-Facing | - | - | - | 8 | 30K |
| UnrealEgo |  | Synth. | Binocular Downward-Facing | - | - | 30 | 17 | 450K |
| DIP-IMU |  | Real | - | Full-Body 3DoF×6 | ✓ | 15 | 10 | 330K |
| FreeDancing |  | Real | - | Full-Body 6DoF×3, 3DoF×3 | ✓ | - | 8 | 532.8K |
| Nymeria |  | Real | Binocular Forward-Facing | Upper-Body 6DoF×3 | ✓ | 20 | 264 | 260M |
| Ego-Exo4D (Ego Pose) |  | Real | Binocular Forward-Facing | Head 6DoF×1 | - | - | - | 9.6M |
| Ours |  | Real | Binocular Downward-Sloping | Full-Body 6DoF×3, 3DoF×2 | ✓ | 39 | 58 | 3.07M |

Table 1: Comparison with existing egocentric motion datasets. EMHI is the first dataset that provides egocentric vision and full-body IMU signals captured by the real VR product suite, along with accurate SMPL annotations simultaneously.

attention in recent years due to the complementarity of different data modalities. Ego-Exo4D (Grauman et al. 2024), Nymeria (Ma et al. 2024) and SimXR (Luo et al. 2024) captured real-world images by Project Aria glasses (Somasundaram et al. 2023), along with the IMU data in upper-body. Ego-Exo4D provided up to 9.6M image frames with annotations of the body and hand joint positions. Nymeria further offered SMPL format data derived from Xsens mocap suits, with limited clothing diversity of the captured body. However, in these datasets, either the forward-facing perspective restricts the perception range of the wearer’s body, or they have not integrated downward-sloping perspectives and sparse full-body IMU signals on actual VR/AR devices.

Egocentric Human Pose Estimation Methods

Vision-based methods have been widely investigated (Rhodin et al. 2016; Wang et al. 2021, 2022, 2023; Akada et al. 2022; Liu et al. 2022, 2023; Kang et al. 2023; Park et al. 2023; Cuevas-Velasquez et al. 2024). For single egocentric pose estimation, Wang *et al.* (Wang et al. 2024) proposed an egocentric motion capture method that combines the vision transformer for undistorted image patch feature extracting and uses diffusion-based motion priors for pose refinement. However, 3D pose estimation from a single image remains challenging due to the lack of depth information. To

address this, EgoPoseFormer (Yang et al. 2024) introduced stereo egocentric skeleton tracking methods by a two-stage transformer-based coarse-to-fine pose estimator for 3D pose prediction. Hiroyasu Akada *et al.* (Akada et al. 2024) further enhanced this with a transformer-based model utilizing 3D scene information and temporal features. Despite these advancements, challenges with invisible body parts due to self-occlusion and out-of-view joints persist.

Methods using sparse tracking signals from body-worn IMUs have garnered significant attentions (Winkler, Won, and Ye 2022; von Marcard et al. 2017; Yi, Zhou, and Xu 2021; Yi et al. 2022; Jiang et al. 2022b; Du et al. 2023). In egocentric VR and AR scenarios, there are inherently three 6DoF tracking points for the head and hands, with the option to add two additional 3DoF IMUs on the legs. AvatarPoser (Jiang et al. 2022a) proposed a global pose prediction framework combining transformer structures with inverse kinematics (IK) optimization, while AvatarJLM (Zheng et al. 2023) introduced a two-stage approach that models joint-level features and uses them as spatiotemporal transformer tokens to achieve smooth action capture. HMDPoser (Dai et al. 2024) integrated these inputs, presenting a lightweight temporal-spatial learning method for full-body global 6DoF body action recovery. However, IMU-based data faces challenges such as drift and sparsity.

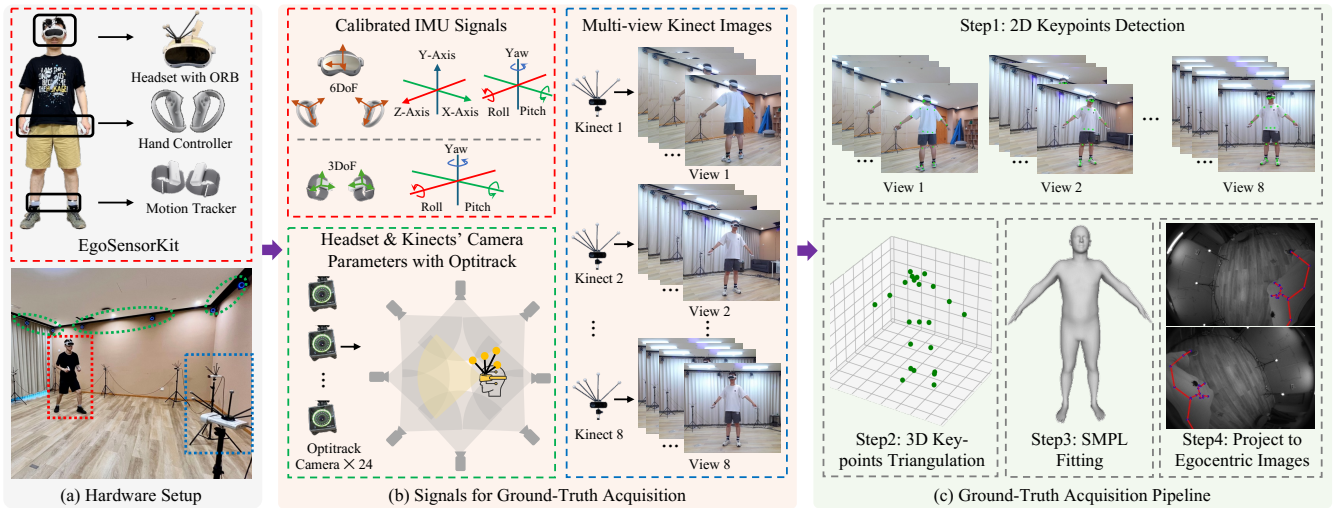


Figure 2: Hardware setup and ground-truth acquisition pipeline. (a) the data capture system consists of EgoSensorKit for egocentric images and calibrated IMU signals collection, eight Azure Kinects for multiple third-view image recording and an Optitrack system for spatiotemporal synchronization of the above signals. With the data collected in (b), (c) we produce the annotations including SMPL parameters and 2D keypoints on egocentric images automatically.

EMHI Dataset

EMHI is a multimodal egocentric motion dataset that contains 3.07M synchronized data pairs organized as 885 sequences recording at 30FPS. Each data pair contains stereo egocentric images (640×480), five IMUs data, and corresponding 3D SMPL pose and 2D keypoints. It is captured by 58 subjects, which are equally split into 29 male and 29 female, with a diverse range of body shapes. Each subject wears their daily clothing during data collection to ensure a wide variety of natural looks. We record 39 common actions of users experiencing games and social applications in VR scenarios and categorize them into upper-body motions, lower-body motions, and full-body motions. Additionally, this dataset is captured under three different environmental lighting conditions: dim light, natural light, and bright light for environment diversity.

Data Capture System

Hardware As shown in Fig. 2, the overall hardware consists of three subsystems: EgoSensorKit system to collect sensor data, with a PICO4 headset, two hand controllers, and two leg trackers; Kinect system to obtain SMPL annotations, with 8 cameras recording simultaneously from outside-in viewpoint; Optitrack system for spatiotemporal synchronization between the above two systems, with Optical Rigid Body (ORBs) mounted at the VR headset and all Kinect cameras, allowing all camera moving.

Temporal Synchronization Kinect and Optitrack systems rely on a signal transmitter device to trigger simultaneously, ensuring their inter-frame alignment. EgoSensorKit and Optitrack could also be synchronized offline with the headset IMU’s and ORB’s angular velocity, according to the motion correlation method in (Qiu et al. 2020). Finally, the data

frames of Kinect and EgoSensorKit are aligned via Optitrack as a bridge. As the recorded frame rate is 30Hz, the maximum synchronization deviation will reach up to 16.5 ms which might be notable during fast motion. So the annotations are further post-processed with linear interpolation to better align with the EgoSensorKit’s timestamps.

Spatial Alignment The 6DoF of the headset (IMU) in Optitrack coordination T_h^o could be obtained by $T_h^o = T_{rb}^o T_h^{rb}$, where T_{rb}^{rb} is a pre-calibrated rigid transformation between the IMU sensor in headset and its ORB, and the ORB’s 6DoF T_{rb}^o is tracked with Optitrack. Similarly, the extrinsic parameters of each Kinect RGB camera in the Optitrack coordinate system T_k^o can be determined using the same method. Then, the spatial transformation between Kinect cameras and headset could be obtained by $T_k^h = (T_h^o)^{-1} T_k^o$. Finally, the transformation matrix between the Kinects and egocentric cameras could be further calculated by $T_k^c = T_h^c T_k^h$, where T_h^c is also a constant spatial relationship between the headset (IMU) and its egocentric cameras.

Ground-Truth Acquisition

Keypoints Annotation We use HRNet (Sun et al. 2019) to detect 2D keypoints in multi-view Kinect RGB images with the body25 format (Cao et al. 2017). Then, we follow HuMMan (Cai et al. 2022) to derive 3D keypoints annotations P_{3D} by triangulation with camera parameters obtained in the spatial alignment, in which we also import smoothness and bone length constraints for P_{3D} to reduce temporal jitter and improve human shape consistency.

SMPL Fitting Multi-view SMPL fitting was a well-solved problem, with the inclusion of 3D joint, prior, smooth, and shape regularization errors. However, due to the occlusion of facial and hand areas by the EgoSensorKit (HMD

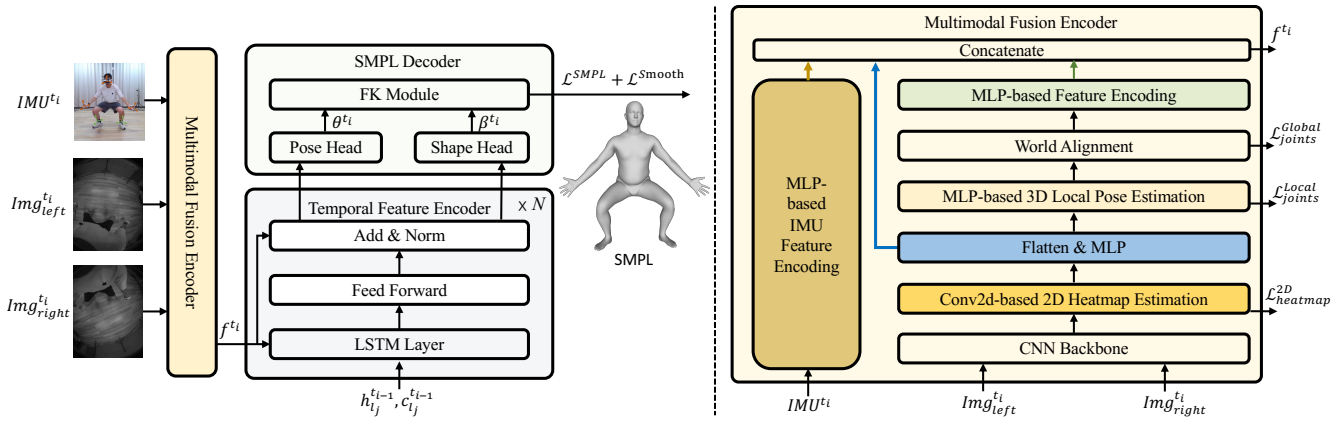


Figure 3: MEPoser. The proposed method consists of a multimodal fusion encoder for feature extracting and fusion of input signals, a temporal feature encoder for history information association, and an SMPL decoder for SMPL parameters prediction.

and controllers) and the Kinect camera’s limited resolution, it’s challenging to ensure the accuracy of the corresponding joint detection. This results in unreasonable SMPL fitting for wrist and head joint rotations. To tackle this problem, we incorporate the global rotation of head R_{head} and wrist joints R_{wrist} , which are transferred from the collected IMU rotations of hand controllers $R_{controller}$ and headset $R_{headset}$ by $R_{head} = R_{headset}^{head} R_{headset}$ and $R_{wrist} = R_{controller}^{wrist} R_{controller}$, where $R_{controller}^{wrist}$ and $R_{headset}^{head}$ are the constant transformation matrix obtained by statistical methods with a large amount of data collected in the standard sensor-wearing settings. Moreover, we leverage the calibrated leg motion tracker data, which represents the knee joint rotation R_{knee} , to constrain the lower leg pose. With the keypoint annotations and five joint rotations obtained above, we fit the SMPL parameters by minimizing the following energy function:

$$E(\theta, \beta) = \lambda_{rot} E_{rot} + \lambda_{joint} E_{joint} + \lambda_{prior} E_{prior} + \lambda_{smooth} E_{smooth} + \lambda_{reg} E_{reg}, \quad (1)$$

where $\theta \in \mathbb{R}^{75}$ and $\beta \in \mathbb{R}^{10}$ are optimized SMPL pose and shape parameters and λ_* are balance weight. For occluded joints, we introduce the rotation term to encourage the pose consistency with transferred IMU data as follows:

$$E_{rot} = \sum_j \|\mathcal{F}(\theta)_j - R_j\|, \quad (2)$$

where $j \in \{head, wrist, knee\}$ and \mathcal{F} indicate a forward kinematic (FK) to get the joint global rotation. Other energy terms are like previous works (Bogo et al. 2016; Cai et al. 2022; Pavlakos et al. 2019), in which E_{joint} minimizes the 3D distance between P_{3D} and regressed SMPL joints. $E_{prior}(\theta)$ is Vposer prior from SMPLify-X. E_{smooth} helps to keep smooth pose tracking, while the shape regularization term E_{reg} penalizes large shape variance. With the space alignment result, the SMPL results could transfer from world space to the egocentric camera coordinate and obtain 2D pose annotations on egocentric images.

A New Baseline Method: MEPoser

To demonstrate the significance of the EMHI dataset and to inspire new designs for multimodal egocentric HPE, we introduce a new baseline method called **Multi-modal Egocentric Pose Estimator (MEPoser)**. MEPoser takes multimodal inputs, including stereo egocentric images and inertial measurements, to extract multimodal representations and perform real-time HPE on a standalone HMD. As shown in Fig. 3, MEPoser consists of three components. (1) A multimodal fusion encoder extracts object representations at each frame from multimodal input data. (2) A temporal feature encoder composed of long short-term memory (LSTM) modules and feed-forward networks generates latent variables containing temporal information incorporated from past frames. (3) With the temporal aggregated multimodal features, two MLP-based (multi-layer perception) heads regress the pose and shape parameters of the SMPL model respectively.

Multimodal Fusion Encoder

The multimodal encoder first has separate feature encoders for different modalities, i.e., two weight-sharing CNN backbones for images and an MLP network for IMU data. To make MEPoser run in real-time on HMD, we use a lightweight RegNetY-400MF (Radosavovic et al. 2020) backbone, which takes stereo images $\{Img_{left}^t, Img_{right}^t\} \in \mathbb{R}^{640 \times 480 \times 1}$ as inputs, and generates 2D image features represented as $\{F_{left}^t, F_{right}^t\} \in \mathbb{R}^{80 \times 60 \times 256}$. These features are then concatenated and forwarded to a few convolution layers to infer a set of heatmaps $\{H_{left}^t, H_{right}^t\} \in \mathbb{R}^{80 \times 60 \times J}$. Here we predict 22 joints of the SMPL, i.e., $J = 22$. To train the RegNetY-400MF backbone, we calculate the binary cross-entropy with logits loss (BCEWithLogitsLoss) $\mathcal{L}_{heatmap}^{2D}$ between the GT heatmaps and the estimated 2D heatmaps. Then, the predicted heatmaps are flattened and forwarded to an MLP network to obtain the image feature. We have obtained the IMU and image features so far. To boost the performance of the pose estimation, we added a 3D module to estimate the 3D

| Dataset | Method | MPJRE↓ | MPJPE↓ | PA-MPJPE↓ | UpperPE↓ | LowerPE↓ | RootPE↓ | Jitter↓ |
|------------|--------------|------------|------------|------------|------------|------------|------------|--------------|
| Protocol 1 | UnrealEgo | - | 5.5 | 3.9 | 4.0 | 7.7 | 4.2 | 592.5 |
| | HMD-Poser | 4.6 | 5.8 | 2.8 | 4.8 | 7.1 | 5.8 | 114.9 |
| | MEPoser-CV | 5.4 | 4.5 | 2.9 | 3.3 | 6.3 | 3.8 | 511.0 |
| | MEPoser-IMU | 5.0 | 6.2 | 3.6 | 5.0 | 8.0 | 5.2 | 121.7 |
| | MEPoser-Full | 4.1 | 3.7 | 2.5 | 2.7 | 5.1 | 3.2 | 161.8 |
| Protocol 2 | UnrealEgo | - | 6.4 | 4.3 | 4.6 | 8.9 | 5.0 | 610.5 |
| | HMD-Poser | 4.9 | 7.0 | 3.4 | 5.2 | 9.7 | 7.2 | 165.7 |
| | MEPoser-CV | 6.0 | 5.4 | 3.5 | 3.8 | 7.8 | 4.4 | 566.5 |
| | MEPoser-IMU | 5.7 | 7.1 | 4.2 | 5.4 | 9.9 | 6.6 | 161.7 |
| | MEPoser-Full | 4.7 | 4.8 | 2.9 | 3.2 | 7.0 | 3.8 | 204.9 |

Table 2: Quantitative comparison between MEPoser and single-modal methods on our EMHI dataset

joint positions in both the local camera coordinate and the global world coordinate. Specifically, given the image features from stereo heatmaps, an MLP network first encodes them to estimate the 3D joint positions in the local camera coordinate $\hat{P}_{local}^{J \times 3}$. Then, these joints are transferred to the global SMPL coordinate $\hat{P}_{global}^{J \times 3}$ with the offline calibration results and the online headset’s 6DoF data. $\hat{P}_{local}^{J \times 3}$ and $\hat{P}_{global}^{J \times 3}$ are used to calculate the 3D joint loss $\mathcal{L}_{joints}^{Local}$ and $\mathcal{L}_{joints}^{Global}$, respectively. Next, the joint positions $\hat{P}_{global}^{J \times 3}$ are flattened and forwarded to an MLP network to obtain the 3D joint features. Finally, the IMU, image, and 3D joint features are concatenated to output the multimodal fused feature f^{t_i} .

Temporal Feature Encoder As demonstrated in HMD-Poser (Dai et al. 2024), temporal correlation information is the key to tracking accurate human motions. However, the multimodal fused features $\{f^{t_i}\}$ are still temporally isolated. To solve this problem, Transformer and RNN are adopted in existing methods. Although Transformer-based methods (Zheng et al. 2023) have achieved state-of-the-art results in HPE, their computational costs are much higher than those of RNN-based methods. To ensure our method runs in real-time on HMDs, we introduce a lightweight LSTM-based temporal feature encoder. Specifically, the encoder is composed of a stack of $N = 3$ identical blocks. And each block has two sub-layers. The first is an LSTM module to learn the temporal representation, and the second is a simple fully connected feed-forward network. We employ a residual connection followed by layer normalization.

SMPL Decoder

The SMPL decoder first adopts two regression heads to estimate the local pose parameters θ^{t_i} and the shape parameters β^{t_i} of SMPL. Both regression heads are designed as a 2-layer MLP. Then, it uses an FK module to calculate all joint positions $\hat{P}_{SMPL}^{J \times 3}$ with θ^{t_i} , β^{t_i} , and the online head’s 6DoF data from headset. We define the SMPL loss function \mathcal{L}^{SMPL} as a combination of root orientation loss \mathcal{L}_{ori} , local pose loss \mathcal{L}_{lrot} , global pose loss \mathcal{L}_{grot} and joint position loss \mathcal{L}_{joint} . All these losses are calculated as the mean of absolute errors (L1 norm) between the predicted results and the ground-truth values.

Training MEPoser

For the overall training loss, we combine a smooth loss \mathcal{L}_{smooth} with the above losses, including 2D heatmap loss $\mathcal{L}_{heatmap}^{2D}$, 3D joint loss $\mathcal{L}_{joints}^{Local}$, $\mathcal{L}_{joints}^{Global}$ and SMPL loss \mathcal{L}_{SMPL} . The smooth loss from HMD-Poser (Dai et al. 2024) is adopted to further enhance the temporal smoothness.

$$\mathcal{L} = \lambda_{hp} \mathcal{L}_{heatmap}^{2D} + \lambda_{l_joints} \mathcal{L}_{joints}^{Local} + \lambda_{g_joints} \mathcal{L}_{joints}^{Global} + \lambda_{smpl} \mathcal{L}_{SMPL} + \lambda_{smooth} \mathcal{L}_{smooth}, \quad (3)$$

Experiment

Dataset Splitting

We split the dataset into three parts: one for training (70%) and two separate testing sets based on different protocols, as follows. The training set comprises 615 sequences captured by 38 individuals, covering 20 daily actions involving upper-body, lower-body, and full-body movements. For testing, **Protocol 1** (16%) contains 141 sequences with the same set of actions but performed by 8 different subjects than those in the training set, to evaluate cross-subject generalization. **Protocol 2** (14%) is designed to assess the model’s effectiveness and robustness in more general scenarios, consisting of 129 sequences involving 19 unseen actions and 20 unseen subjects not present in the training set.

Comparison

To validate the dataset and the corresponding baseline methods, we conducted comparisons using our dataset between MEPoser against the latest single-modal methods. The quantitative results in Tab. 2 show that MEPoser outperforms existing single-modal methods for egocentric HPE. Compared to UnrealEgo (Akada et al. 2022) which takes stereo egocentric images as inputs, our method reduces the MPJPE (Mean Per Joint Position Error, cm) by 32.7% and 25% in protocol1 and protocol2, respectively. Notably, MEPoser significantly enhances the smoothness of estimation results by using a temporal LSTM structure. Besides, our method obtains 27.5% and 23.8% enhancement in joint location precision in contrast to EgoPoseFormer (Yang et al. 2024). In comparison with whole-body IMU-based method HMD-Poser (Dai

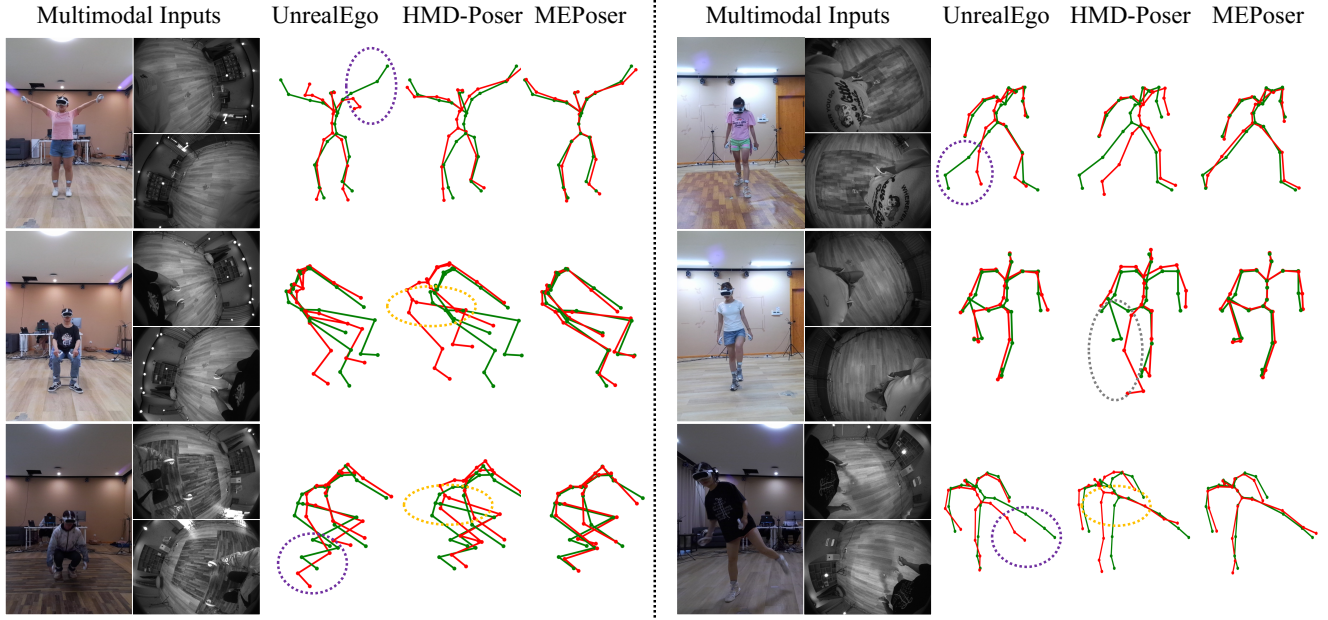


Figure 4: Qualitative comparison between ours and single-modal methods (GT: green, estimation: red). Ours relieves joint invisibility in egocentric images (purple), IMU data drifting (yellow), and ambiguous measurements in slow motions(gray).

et al. 2024), MEPoser shows 36.2% and 31.4% reduction in MPJPE on two test sets by combining the egocentric image features and slight improvement of joint rotation accuracy according to the MPJRE (Mean Per-Joint Rotation Error, $^{\circ}$). The results also demonstrate the improved generalizability of MEPoser across subjects and actions, validating the effectiveness of the multimodal setting and the value of our dataset in solving egocentric HPE. We also include ablation experiments to investigate the impact of different components in our network and the results demonstrate that each modality contributes to performance improvements.

Qualitative comparisons are shown in Fig. 4 using various test sequences, featuring different actions and environments. MEPoser could relieve the limitations of the single-modal methods by exploiting the complementary between the vision and IMU signals. Specifically, our method could deal with issues like self-occlusion and out-of-FOV problems in egocentric images by utilizing IMU features and temporal information to get more accurate pose results. Additionally, our approach mitigates the sparsity, drifting, and ambiguous measurements in slow motions of IMU signals by incorporating visible body joints in egocentric images.

Annotation Cross-validation

To further assess the accuracy of our dataset annotations, we randomly capture 9 motion sequences for cross-validation by simultaneously using both our system and an optical marker-based motion capture system. The subject wears a suit attached with reflective markers tracked by the OptiTrack system. Marker-based SMPL parameters are then derived from the MoCap data using MoSh++ (Loper, Mahmood, and Black 2014), with a temporal filter applied to reduce jittering. As shown in Tab. 3, the low error metrics

| Action | MPJPE | Action | MPJPE |
|---------------------|-------|----------------|-------------|
| walking | 2.61 | kicking | 2.12 |
| hand-waving | 2.72 | lunge | 2.23 |
| taichi | 2.98 | boxing | 2.73 |
| shuttlecock-kicking | 2.65 | dancing | 2.49 |
| marching-in-place | 2.08 | Average | 2.51 |

Table 3: Cross-validation for dataset annotations

and variance demonstrate the robustness of our annotation pipeline and the high quality of the dataset. Besides the validation, every sequence of our dataset has been inspected manually to eliminate the data with erroneous annotations.

Conclusion

In this paper, we introduce EMHI, a novel multimodal human motion dataset designed for egocentric HPE. It includes synchronized egocentric images and IMU signals from a real VR product suite, with SMPL annotations in the same world coordinate system. To enhance generalization in real-world applications, we collected a diverse range of data across various actions and individuals. We also present MEPoser, a new baseline HPE method that combines image and IMU inputs for real-time HPE on a standalone HMD. MEPoser effectively demonstrates the benefits of multimodal fusion, improving accuracy and addressing the limitations of previous single-modal methods. This approach serves as an initial exploration, inviting further research of egocentric HPE with multimodal data. We believe releasing this dataset and method will accelerate the practical implementation of HPE with body-worn sensors in future VR/AR products.

References

- Akada, H.; Wang, J.; Golyanik, V.; and Theobalt, C. 2024. 3d human pose perception from egocentric stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 767–776.
- Akada, H.; Wang, J.; Shimada, S.; Takahashi, M.; Theobalt, C.; and Golyanik, V. 2022. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, 1–17. Springer.
- Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, 561–578. Springer.
- Cai, Z.; Ren, D.; Zeng, A.; Lin, Z.; Yu, T.; Wang, W.; Fan, X.; Gao, Y.; Yu, Y.; Pan, L.; et al. 2022. Humman: Multimodal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, 557–577. Springer.
- Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Real-time multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.
- Cha, Y.-W.; Shaik, H.; Zhang, Q.; Feng, F.; State, A.; Ilie, A.; and Fuchs, H. 2021. Mobile. Egocentric human body motion reconstruction using only eyeglasses-mounted cameras and a few body-worn inertial sensors. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 616–625. IEEE.
- Cuevas-Velasquez, H.; Hewitt, C.; Aliakbarian, S.; and Baltrušaitis, T. 2024. Simpleego: Predicting probabilistic body pose from egocentric cameras. In *2024 International Conference on 3D Vision*, 1446–1455. IEEE.
- Dai, P.; Zhang, Y.; Liu, T.; Fan, Z.; Du, T.; Su, Z.; Zheng, X.; and Li, Z. 2024. 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*, 874–884.
- Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 1–23.
- Du, Y.; Kips, R.; Pumarola, A.; Starke, S.; Thabet, A. K.; and Sanakoyeu, A. 2023. Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 481–490.
- Gong, X.; Mohan, S.; Dhingra, N.; Bazin, J.-C.; Li, Y.; Wang, Z.; and Ranjan, R. 2023. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6481–6491.
- Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. 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*, 19383–19400.
- Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; Hilliges, O.; and Pons-Moll, G. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics*, 37(6): 1–15.
- Jiang, J.; Strelci, P.; Qiu, H.; Fender, A.; Laich, L.; Snape, P.; and Holz, C. 2022a. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, 443–460. Springer.
- Jiang, Y.; Ye, Y.; Gopinath, D.; Won, J.; Winkler, A. W.; and Liu, C. K. 2022b. Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. In *SIGGRAPH Asia 2022 Conference Papers*, 3:1–3:9. ACM.
- Kang, T.; Lee, K.; Zhang, J.; and Lee, Y. 2023. Ego3dpose: Capturing 3d cues from binocular egocentric views. In *SIGGRAPH Asia 2023 Conference Papers*, 1–10.
- Liu, Y.; Yang, J.; Gu, X.; Chen, Y.; Guo, Y.; and Yang, G.-Z. 2023. EgoFish3D: Egocentric 3D Pose Estimation From a Fisheye Camera via Self-Supervised Learning. *IEEE Transactions on Multimedia*, 25: 8880–8891.
- Liu, Y.; Yang, J.; Gu, X.; Guo, Y.; and Yang, G.-Z. 2022. Ego+ x: An egocentric vision system for global 3d human pose estimation and social interaction characterization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5271–5277. IEEE.
- Loper, M.; Mahmood, N.; and Black, M. J. 2014. MoSh: motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6): 220–1.
- Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.
- Luo, Z.; Cao, J.; Khirodkar, R.; Winkler, A.; Kitani, K.; and Xu, W. 2024. Real-Time Simulated Avatar from Head-Mounted Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 571–581.
- Ma, L.; Ye, Y.; Hong, F.; Guzov, V.; Jiang, Y.; Postyeni, R.; Pesqueira, L.; Gamino, A.; Baiyya, V.; Kim, H. J.; et al. 2024. Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild. *arXiv preprint arXiv:2406.09905*.
- Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.
- Park, J.; Kaai, K.; Hossain, S.; Sumi, N.; Rambhatla, S.; and Fieguth, P. 2023. Domain-guided spatio-temporal self-attention for egocentric 3D pose estimation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1837–1849.
- Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. 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*, 10975–10985.
- Qiu, K.; Qin, T.; Pan, J.; Liu, S.; and Shen, S. 2020. Real-time temporal and rotational calibration of heterogeneous sensors using motion correlation analysis. *IEEE Transactions on Robotics*, 37(2): 587–602.
- Radosavovic, I.; Kosaraju, R. P.; Girshick, R.; He, K.; and Dollár, P. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10428–10436.
- Rai, N.; Chen, H.; Ji, J.; Desai, R.; Kozuka, K.; Ishizaka, S.; Adeli, E.; and Niebles, J. C. 2021. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11184–11193.
- Rhodin, H.; Richardt, C.; Casas, D.; Insafutdinov, E.; Shafiei, M.; Seidel, H.-P.; Schiele, B.; and Theobalt, C. 2016. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics*, 35(6): 1–11.
- Somasundaram, K.; Dong, J.; Tang, H.; Straub, J.; Yan, M.; Goesele, M.; Engel, J. J.; De Nardi, R.; and Newcombe, R. 2023. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561*.
- Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.
- Tome, D.; Peluse, P.; Agapito, L.; and Badino, H. 2019. xregopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7728–7738.
- Trumble, M.; Gilbert, A.; Malleon, C.; Hilton, A.; and Colloso, J. P. 2017. Total capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC*, 5, 1–13. London, UK.
- Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, 2, 349–360. Wiley Online Library.
- von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Comput. Graph. Forum*, 36(2): 349–360.
- Wang, J.; Cao, Z.; Luvizon, D.; Liu, L.; Sarkar, K.; Tang, D.; Beeler, T.; and Theobalt, C. 2024. Egocentric whole-body motion capture with fisheye and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 777–787.
- Wang, J.; Liu, L.; Xu, W.; Sarkar, K.; Luvizon, D.; and Theobalt, C. 2022. 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*, 13157–13166.
- Wang, J.; Liu, L.; Xu, W.; Sarkar, K.; and Theobalt, C. 2021. Estimating Egocentric 3D Human Pose in Global Space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11500–11509.
- Wang, J.; Luvizon, D.; Xu, W.; Liu, L.; Sarkar, K.; and Theobalt, C. 2023. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13031–13040.
- Winkler, A.; Won, J.; and Ye, Y. 2022. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, 1–8.
- Xu, W.; Chatterjee, A.; Zollhoefer, M.; Rhodin, H.; Fua, P.; Seidel, H.-P.; and Theobalt, C. 2019. Mo2cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5): 2093–2101.
- Yang, C.; Tkach, A.; Hampali, S.; Zhang, L.; Crowley, E. J.; and Keskin, C. 2024. EgoPoseFormer: A Simple Baseline for Egocentric 3D Human Pose Estimation. *arXiv preprint arXiv:2403.18080*.
- Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 2022. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13167–13178.
- Yi, X.; Zhou, Y.; and Xu, F. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4): 1–13.
- Zhao, D.; Wei, Z.; Mahmud, J.; and Frahm, J.-M. 2021. Ego-glass: Egocentric-view human pose estimation from an eye-glass frame. In *2021 International Conference on 3D Vision*, 32–41. IEEE.
- Zheng, X.; Su, Z.; Wen, C.; Xue, Z.; and Jin, X. 2023. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14678–14688.