Group Inertial Poser: Multi-Person Pose and Global Translation from Sparse Inertial Sensors and Ultra-Wideband Ranging

Ying Xue, Jiaxi Jiang, Rayan Armani, Dominik Hollidt, Yi-Chi Liao, and Christian Holz
Department of Computer Science, ETH Zürich, Switzerland

https://siplab.org/projects/GroupInertialPoser

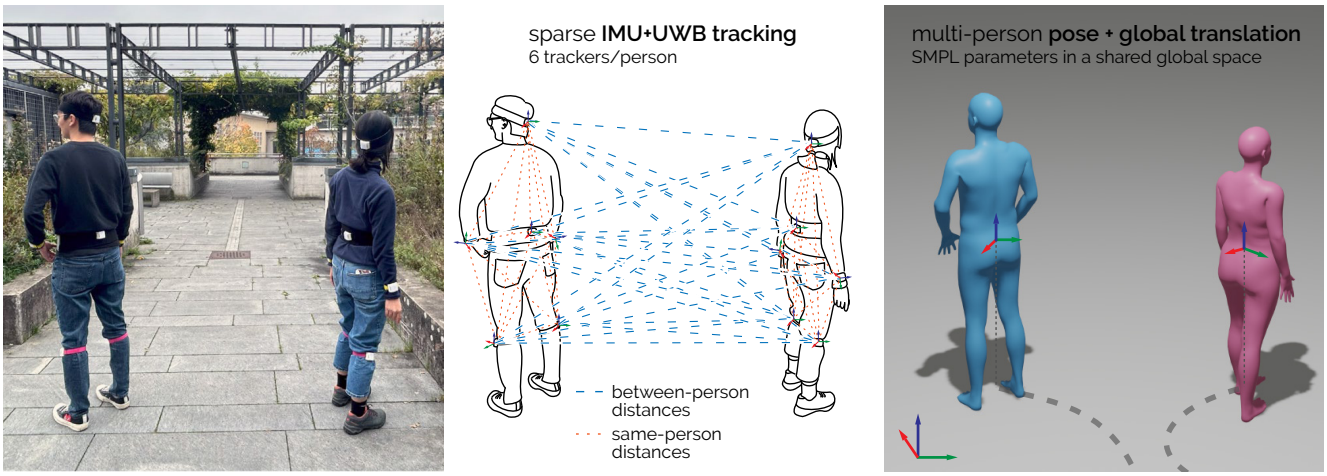

Figure 1. Our method *Group Inertial Poser* estimates 3D full-body poses and global translation for multiple humans using inertial measurements from a sparse set of wearable sensors, augmented by the distances between the sensors via ultra-wideband ranging. Our approach overcomes the challenge of drift in previous inertial pose estimators to track translation, as we leverage information from body motions *across* multiple people. Our IMU+UWB method stabilizes and improves individual pose estimates, relative translation estimates between people, and global translation estimates, thereby preserving meaningful interaction dynamics.

Abstract

Tracking human full-body motion using sparse wearable inertial measurement units (IMUs) overcomes the limitations of occlusion and instrumentation of the environment inherent in vision-based approaches. However, purely IMU-based tracking compromises translation estimates and accurate relative positioning between individuals, as inertial cues are inherently self-referential and provide no direct spatial reference for others. In this paper, we present a novel approach for robustly estimating body poses and global translation for multiple individuals by leveraging the distances between sparse wearable sensors—both on each individual and across multiple individuals. Our method Group Inertial Poser *estimates these absolute distances between pairs of sensors from ultra-wideband ranging (UWB) and fuses them with inertial observations as input into structured state-space models to integrate temporal motion patterns for precise 3D pose estimation. Our novel two-step optimization further leverages the estimated distances for accurately tracking people's global trajectories through the world. We also introduce* GIP-DB*, the first IMU+UWB dataset for two-person tracking, which comprises 200 minutes of motion recordings from 14 participants. In our evaluation, Group Inertial Poser outperforms previous state-of-the-art methods in accuracy and robustness across synthetic and real-world data, showing the promise of IMU+UWB-based multi-human motion capture in the wild. Code, models, dataset: github.com/eth-siplab/GroupInertialPoser.*

1. Introduction

Accurate motion tracking is a long-standing goal in computer vision. Single-person motion tracking has been extensively studied using camera-based approaches [10, 14, 23, 34, 39, 63, 76]. Extending such tracking to multi-person motions introduces computational complexity and new challenges. While vision-based approaches are accurate [13, 19, 40, 43, 58, 71, 77], they struggle with occlusion

in crowded environments and face challenges when individuals are close. Their frequent reliance on stationary camera setups further constrains the effective tracking range.

Body-worn sensor-based approaches offer a promising alternative to these challenges. Using sparse sets of wearable motion sensors, typically inertial measurement units (IMUs), has become popular to capture individuals' movements independent of environmental factors such as lighting, occlusion, or dynamic backgrounds. Recent IMU-based approaches [81, 82, 84] estimate body poses from six body-worn sensors in controlled environments, but they often exhibit drift in their predictions and, thus, struggle with estimating global translation. Extending these approaches to multi-person constellations would further complicate accurately estimating relative positions *between* individuals. This prevents them from capturing inter-personal dynamics and spatial relationships between people—aspects that are highly interesting to reconstruct when aiming to understand interactions in real social scenarios.

In this paper, we introduce *Group Inertial Poser* (GIP), a novel approach for robust multi-person 3D pose and global translation estimation from sparse inertial sensing with inter-sensor distances (as illustrated in Figure 1). It first uses structured state-space models to estimate individual body poses and translations from inertial signals and same-person sensor distances. Then, a two-step optimization process refines the translation estimates: (a) Relative position optimization aligns individuals in a shared world frame using between-person distances, eliminating the need for calibrated starting positions; (b) Trajectory optimization further improves global translation accuracy.

To validate GIP in *real-world settings*, we introduce *GIP-DB*, a novel motion dataset that captures diverse activities from pairs of 14 participants who interacted during recording. Evaluating GIP on GIP-DB, our results show that GIP estimates more accurate poses and translations with reduced drift—despite high noise levels in between-person distances. Finally, we demonstrate GIP for a four-person setting, where our method's performance improves as it leverages more distances. By analyzing the reconstructed inter-human motions, we demonstrate that GIP effectively captures relative spatial relationships and preserves meaningful inter-personal motion dynamics—an essential capability for modeling human-to-human interaction that previous approaches fail to achieve.

Contributions

1. *Group Inertial Poser* (GIP), a novel method to incorporate between-sensor distances and inertial signals to estimate 3D full-body poses and global translations for multiple people. *GIP* is the first IMU+UWB-based solution for estimating multi-person motion in a *shared reference frame* and sets new state-of-the-art results.
2. A structured state-space network that efficiently models sequential data to improve human pose estimation. To our knowledge, we are the first to adapt state-space models for inertial-based human motion estimation.
3. An optimization-based method to estimate people's initial world positions, allowing tracked users to start at arbitrary locations and eliminating the need for calibration, manual setup, or aligning synchronized motions.
4. *GIP-DB*, the first IMU+UWB two-person motion dataset with diverse activities from 14 participants who interact and perform everyday movements, totaling over 200 minutes of captured motions. GIP-DB comprises synchronized IMU signals, UWB-based distance measurements, and SMPL body motion parameters.

2. Related Work

Multi-Person Motion Capture. Human motion capture, for both single and multiple individuals, has traditionally relied on camera-based approaches, with marker-based systems [58, 71] offering high precision but requiring costly setups confined to indoor environments. Recent advances in computer vision have enabled human pose estimation from sparse images [8, 10, 14, 34] or videos [23, 35, 39, 70, 77, 86]. Frames from third-person [27], floor [7, 11], or egocentric cameras [32, 66] extend naturally to multi-person scenarios, with challenges mainly in accurate person detection, tracking, and pose estimation. Efforts such as RTMO [50] and HigherHRNet [15] proposed efficient bottom-up approaches for multi-actor tracking, while PETR [62] uses transformer-based models and AlphaPose [19, 42] offers joint pose estimation-tracking frameworks. However, tracking in crowded or occluded scenarios remains challenging due to obstruction and inconsistent tracking [18], which wearable inertial motion capture systems can address by overcoming subject identification and occlusion issues.

Motion Capture with Wearable Sensors. Wearable inertial sensors (IMU) have emerged as an alternative approach to motion capture, given their low power, small form factor, and affordability. Commercial systems, such as XSens [78] or Noitom [56] use 17 to 19 IMUs with underlying biomechanical models to estimate body pose. With the availability of large motion capture datasets [52, 68], learning-based methods using only sparsely worn IMUs [28, 29, 53, 55, 73] are emerging. Recent approaches use 6 IMUs and estimate both body pose and translation through ground contact points [33, 81] and physical constraints [82, 84, 85]. Several methods have further reduced input requirements to just the upper body alone [1, 22, 30–32, 44, 51, 64, 75, 80].

IMUs detect relative acceleration and angular velocity, so IMU-only methods have lower pose estimation

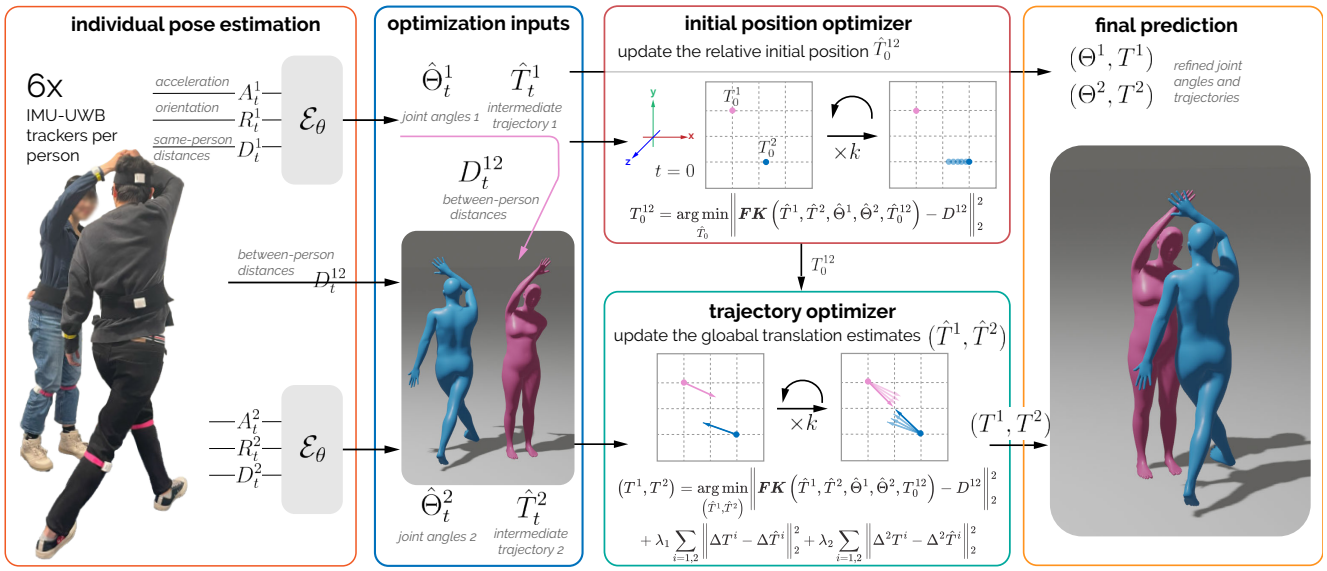

Figure 2. Overview of Group Inertial Poser (GIP). Our pipeline consists of three key steps. It begins with individual pose estimation using an SSM-based model $\mathcal{E}_\theta$ to generate a full-body SMPL pose and translations. The optimization steps then refine these estimates by minimizing the discrepancy between predicted and actual between-sensor distances. First, GIP optimizes the initial relative positions (T_0^{12}); second, it fine-tunes the translations for both users (T^1, T^2).

and translation accuracy compared to vision-based alternatives. To address this, researchers have explored hybrid approaches incorporating external [59, 65, 74] or body-worn [26, 45, 83] cameras. Alternative sensing modalities have also been investigated, including wearable ultrasonic [48, 72] and electromagnetic sensors [36, 37]. Recently, UIP [5] proposed leveraging distance measurements from ultra-wideband ranging to constrain IMU drift. Yet, these approaches focus on single-person tracking as sensors worn by different actors are treated as disjoint problems. We propose a novel approach to IMU+UWB-based pose estimation that, for the first time, enables multi-human relative pose estimation solely from inertial sensors and UWB.

Ultra-Wideband Ranging. UWB is characterized by its large bandwidth and very short waveforms, which is well-suited for ranging applications, such as asset tracking [6, 69, 87], robotic localization [4, 12, 41, 57, 88] and collaboration [16, 60]. It is increasingly available in commercial devices such as smartphones, smartwatches, and tags (*e.g.* AirTags) [3, 17, 61]. A challenge in UWB ranging is the noise in non-line of sight (NLOS) conditions. This is especially relevant in human motion tracking, where the human body is an obstacle between two ranging UWB radios. Researchers have addressed this by analyzing raw channel impulse response on UWB radios [2, 9, 67] or via sensor fusion with IMUs [20, 25, 54, 57] or cameras [60, 79], and effectively filtering distance estimates [4]. Building on this approach, GIP estimate distances in any constellation of trackers, worn by one or multiple people.

3. Method

3.1. Problem Formulation

GIP addresses multi-person pose estimation using sparse inertial sensors and between-sensor distances. GIP is a generic approach that can be applied to any number of users. For simplicity and clarity, we consider a scenario with two users in our notation, each equipped with $S = 6$ sparse sensors placed on the head, pelvis, wrists and knees. Each sensor includes a 6-DoF IMU and a single UWB sensor. For each user $i \in \{1, 2\}$ and at each frame t, we obtain 3D orientations $R_t^i \in \mathbb{R}^{S\times 3}$ and accelerations $A_t^i \in \mathbb{R}^{S\times 3}$, both represented in a shared world frame. Additionally, we denote the pairwise same-person sensor distances as $D_t^i \in \mathbb{R}^{S\times S}$ for user i, and the between-person distances as $D_t^{12} \in \mathbb{R}^{S\times S}$ at timestamp t. Given a sequence of length N represented by $[A^1, A^2, R^1, R^2, D^1, D^2, D^{12}]$, we predict the SMPL [49] pose parameters $\boldsymbol{\Theta}^i \in \mathbb{R}^{N\times 3J}$ and the translation $T^i \in \mathbb{R}^{N\times 3}$ for each user i, all in a shared frame of reference. Here, $J = 24$ represents the number of joints.

3.2. Method Overview

Figure 2 shows an overview of our proposed method Group Inertial Poser (GIP). The first step of our pipeline is **Individual Pose Estimation**, which involves estimating each person's full-body pose independently using a learning-based pose estimator. We design a human pose estimator based on State Space Models [21] to produce sequences of full-body SMPL poses $\hat{\Theta}^i$ and root translation trajectories $\hat{T}^i$ using the acceleration, orientation, and same-person sen-

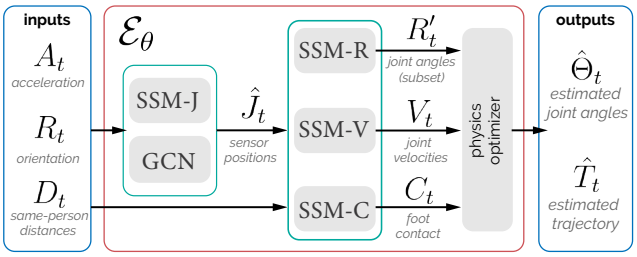

Figure 3. Our SSM-based model takes the global acceleration, rotation, and between-sensor distances to estimate the pose and translation via SMPL.

sor distances from six IMUs per person. To extend from individual pose to multi-person pose, our method introduces two optimization steps: The **Initial Position Optimization** refines the initial positions of the two trajectories. We define one user's initial position P_1 as the world origin $(0, 0, 0)$. Our goal is to optimize the second user's position P_2, which is relative to P_1. This step aligns trajectories into a shared world frame and provides a stable initialization for the next optimizer. The **Trajectory Optimization** further refines the full trajectories by integrating between-human distance D^{ij} constraints through a trajectory optimizer. This step enforces consistency with the relative distances between individuals. Note that, directly optimizing the entire trajectory (without the first optimization step) results in unstable convergence to unrealistic paths.

3.3. Individual Pose Estimation

GIP starts from individual pose estimation and is performed separately for each user. As shown in Figure 3, the human pose estimator ($\mathcal{E}_\theta$) is state-space model based, which uses acceleration $\mathbf{A}^i$, orientation $\mathbf{R}^i$, and same-person distances D^i to predict SMPL pose parameters (joint orientations) $\hat{\mathbf{\Theta}}^i$ [49] and translations $\hat{T}^i$ which are relative to their own initial frame:

$$\hat{\mathbf{\Theta}}^i, \hat{T}^i = \mathcal{E}_\theta(\mathbf{A}^i, \mathbf{R}^i, D^i) \tag{1}$$

Since the pose estimator $\mathcal{E}_\theta$ processes each user independently, SMPL poses and translations for each user are estimated separately when multiple users are present.

Inspired by recent advancements in structured state-space sequence models (S4), which have shown remarkable proficiency in sequence modeling [21], we design our pose estimator based on state-space models (SSMs). The S4 model leverages SSMs to efficiently capture complex patterns in sequential data, making it highly effective for modern, large-scale applications. Compared to traditional models like LSTMs [24], S4 offers improved scalability, long-range dependency modeling, and parallelization. SSMs rely on a classical continuous-time system that maps an input sequence $x(t) \in \mathcal{R}$, through intermediate implicit states $h(t) \in \mathcal{R}^N$ to an output $y(t) \in \mathcal{R}$. The aforementioned process can be formulated as a linear Ordinary Differential Equation (ODE): $h'(t) = \mathrm{A}h(t) + \mathrm{B}x(t)$, and $y(t) = \mathrm{C}h(t)$, where $\mathrm{A} \in \mathcal{R}^{N\times N}$ denotes the state transition matrix, while $\mathrm{B} \in \mathcal{R}^{N\times 1}$ and $\mathrm{C} \in \mathcal{R}^{1\times N}$ represents the projection parameters. The S4 model discretizes this continuous system, making it suitable for deep learning scenarios. Specifically, it introduces a timescale parameter Δ and applies fixed discretization rules to transform A and B into discrete parameters $\overline{\mathrm{A}}$ and $\overline{\mathbf{B}}$. Typically, zero-order hold (ZOH) is employed as the discretization rule, defined as follows: $\overline{\mathrm{A}} = \exp(\Delta\mathrm{A}), \overline{\mathrm{B}} = (\Delta\mathrm{A})^{-1}(\exp(\Delta\,\mathrm{A})-\mathrm{I})\cdot\Delta\mathrm{B}$. After discretization, the SSM can be computed through linear recurrence, described as

$$\begin{aligned} h(t) &= \overline{\mathrm{A}}h(t-1) + \overline{\mathrm{B}}x(t), \\ y(t) &= \mathrm{C}h(t) \end{aligned} \tag{2}$$

Our pose estimator consists of four structured SSM (S4) modules, one graph convolutional network (GCN) module, and a physics optimizer. Since GCNs have been effective in capturing between-sensor distance information [5], we incorporate a GCN to process orientation and between-sensor distances, predicting sensor positions. Meanwhile, the SSM-J module takes orientation and acceleration as inputs to estimate sensor positions. We learn adaptive weights for SSM-J and GCN to fuse their predictions effectively. Next, given the predicted sensor positions and the input sensor acceleration, orientation, and same-person sensor distances, SSM-R, SSM-V, and SSM-C predict joint angles, joint velocities, and foot contact states, respectively. Finally, these predicted values are passed to a physics-based optimization module to ensure physical correctness, following [5, 82]. This produces the final estimated joint angles $\hat{\mathbf{\Theta}}^i$ and translation $\hat{T}^i$.

3.4. Initial Position Optimization

GIP initializes by setting the initial position of the first person, T_0^1, as the origin in the global coordinate system, while setting the initial position of the second person as T_0^2. Given these two positions, the relative position between them is represented as T_0^{12}. The objective of this step is to determine the optimal T_0^{12} by minimizing the differences between two sets of between-person distances: one is derived from *prediction* (denoted as PD_t) and the other is from *UWB* sensing (denoted as D_t). Specifically, the *predicted* distances are computed with the individual pose estimator with forward kinematics. The *UWB-based distances* are direct measurements. Since each user wears 6 sensors, the optimization considers only the 6×6 between-person sensor pairs, excluding any within-user pairs, to refine the spatial alignment of the two users.

To generate the predicted between-person differences, at each timestamp t, we compute the global positions of all

sensors for each person using forward kinematics applied to SMPL parameters $\hat{\mathbf{\Theta}}_t^i$ and the translation T_t^i, which are relative to each person's initial position P_i. The predicted sensor positions, denoted $SP_t^i \in \mathbb{R}^{S\times 3}$, are calculated as follows:

$$SP_t^i = P_i + \hat{T}_t^i + fk(\hat{\mathbf{\Theta}}_t^i), \tag{3}$$

where $fk(\hat{\mathbf{\Theta}}_t^i)$ is the forward kinematics function that computes the body mesh based on joint rotations and estimates sensor positions from the corresponding mesh vertices. From the predicted sensor coordinates, we calculate the predicted between-person sensor distances $PD_t \in \mathbb{R}^{S\times S}$. In detail, each distance is computed as:

$$PD_t(j,k) = \left\| SP_t^1(j) - SP_t^2(k) \right\|_2, \tag{4}$$

where j and k index the sensors on user 1 and user 2, respectively. The predicted distances are compared with the actual distances $D_t^{12} \in \mathbb{R}^{S\times S}$ obtained from UWB measurements.

To streamline the full process, we define FK as the composite function that takes SMPL parameters, translations, and initial positions as input, applies forward kinematics via fk, and computes pairwise distances PD_t:

$$PD_t = FK(\Theta^1, \Theta^2, T^1, T^2, T_0^{12}) \tag{5}$$

The goal here is to adjust T_0^{12} so that the predicted and actual between-person sensor distances align as closely as possible over the entire trajectory of T timestamps:

$$(T_0^{12}) = \arg\min_{\hat{T}_0^{12}} \sum_{t=1}^{T} \left\| FK(\hat{\Theta}_t^1, \hat{\Theta}_t^2, \hat{T}_t^1, \hat{T}_t^2, \hat{T}_0^{12}) - D_t^{12} \right\|_2^2, \tag{6}$$

3.5. Trajectory Optimization

The initial positions optimization effectively aligns the trajectories of two people into a shared world frame. The trajectory optimization step then improves trajectories of both users simultaneously.

$$T_t^1, T_t^2 = \arg\min_{(\hat{T}_t^1, \hat{T}_t^2)} \sum_{t=1}^{T} \left\| FK\left(\hat{\Theta}_t^1, \hat{\Theta}_t^2, \hat{T}_t^1, \hat{T}_t^2, T_0^{12}\right) - D_t^{12} \right\|_2^2 \tag{7}$$

$$+ \lambda_1 \sum_{i=1,2} \left\| \Delta T^i - \Delta \hat{T}^i \right\|_2^2 \quad + \lambda_2 \sum_{i=1,2} \left\| \Delta^2 T^i - \Delta^2 \hat{T}^i \right\|_2^2 \tag{8}$$

In this formulation: The first term (7) enforces alignment between predicted and observed between-person sensor distances. The two regularization terms (8) promote translation smoothness. Let $\Delta\mathbf{T}$ and $\Delta^2\mathbf{T}$ denote the first- and second-order differences (velocity and acceleration) of a person's translation:

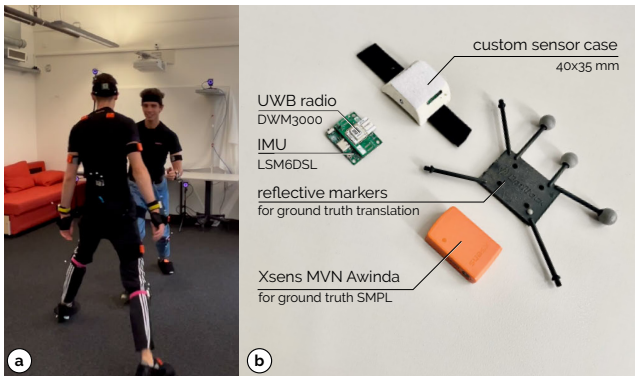

Figure 4. (a) Our dataset includes pairs of participants equipped with (b) various motion capture sensors. These sensors capture acceleration, orientation, and distance data, along with ground-truth SMPL pose parameters and translations for each participant.

$$\Delta\mathbf{T} = \mathbf{T}_{:,2:T} - \mathbf{T}_{:,1:T-1} \tag{9}$$

$$\Delta^2\mathbf{T} = \Delta\mathbf{T}_{:,2:T-1} - \Delta\mathbf{T}_{:,1:T-2} \tag{10}$$

$$= \mathbf{T}_{:,3:T} - 2 \times \mathbf{T}_{:,2:T-1} + \mathbf{T}_{:,1:T-2} \tag{11}$$

One term penalizes the difference between true and predicted velocities(ΔT^i vs. $\Delta \hat{T}^i$), and the other does the same for accelerations ($\Delta^2 T^i$ vs. $\Delta^2 \hat{T}^i$). Together, they encourage smoother motion by constraining predicted speed and acceleration.

3.6. Implementation Details

We integrate four S4-based neural networks, namely SSM-J, SSM-R, SSM-V, and SSM-C (see Figure 3), each designed to process different output features. These models share a consistent architecture, beginning with a linear encoder that maps the input features to a hidden dimension of 256. The exception is SSM-C, where the hidden dimension is reduced to 32, due to the smaller output size (2). Each network then passes through two residual S4 layers, which incorporate LayerNorm and dropout (0.2) for regularization. Finally, a linear decoder maps the output representations to the respective task-specific output spaces. We train the model on a single NVIDIA 4090 GPU with a batch size of 256, a sequence length of 200, and a learning rate of 1×10^{-3}, which decays by a factor of 0.33 every 20 epochs. The *initial position optimizer* and *trajectory optimizer* utilize the L-BFGS optimizer [47], with a maximum of four iterations per optimization step and a history size of 10. We use *strong Wolfe* line search to ensure robust step size selection and stable convergence [47].

4. Group Inertial Poser Dataset (GIP-DB)

To evaluate our method on real-world data, we collected a motion capture dataset featuring seven pairs of participants

(10 male, 4 female) with heights ranging from 160 cm to 195 cm. The dataset includes everyday movements such as walking, stretching, and jogging in place, as well as interactive activities like close conversation, sparring, handshakes, and dancing. Each participant was equipped with an Xsens MVN Awinda suit [78] consisting of 17 IMUs, which provided ground-truth SMPL pose parameters via Xsens' proprietary software. Additionally, they wore six custom wireless sensors placed on the head, pelvis, wrists, and knees. Each custom sensor integrates a UWB radio (DWM3000) and a 6DoF IMU (LSM6DSL). The UWB system, running a ranging protocol and filtering pipeline from existing work [4], measured distances at 40 Hz across 12 sensor pairs: (15x2) same-person sensor distances and 36 between-person sensor distances. Meanwhile, each IMU recorded acceleration A_t and orientation R_t at 104 Hz. We additionally captured ground-truth translation using a 20-camera OptiTrack system [58], using the pelvis sensor as the reference point. All sensors were calibrated according to their respective documentation [4, 78]. Each recording session began and ended with a T-pose and a jump to synchronize data across sources (Xsens, OptiTrack, and custom sensors). Figure 4 illustrates our setup. In our GIP-DB dataset, the average UWB RMSE is 5 cm for same-person measurements, increasing to 15 cm for between-person measurements due to greater occlusions.

5. Experiments

To assess the advantages of *GIP* for multi-human inertial pose estimation, we conduct experiments on both synthetic and real-world data. We compare our method against previous inertial sensing-based approaches, namely PIP [82] and UIP [5]. There are fundamental differences between GIP and these prior methods. Notably, neither PIP nor UIP was designed to estimate inter-human translation. To address this limitation, we initialize their translation at frame zero using the ground truth. Additionally, PIP was originally designed to work solely with IMU data and does not leverage between-sensor distances. To ensure a fair comparison, we provide PIP with additional sensor distance measurements following the same approach as UIP [5].

Datasets. Our evaluation aims to show the benefits of *GIP* in both synthetic and real-world scenarios. All methods are trained from scratch to estimate individual poses using synthesized AMASS data [52]. Specifically, we generate synthetic data by computing IMU measurements (acceleration and rotation) and between-sensor distances on virtual sensors, following prior work [5, 28, 82]. We use two datasets for evaluation: InterHuman [46] and our real-world GIP-DB dataset, both of which contain two-person interactions. For InterHuman, we synthesize IMU and UWB data using the same procedure as for AMASS.

Metrics. Following previous work on human pose estimation from inertial sensors [5, 81, 82], we evaluate the predicted poses and trajectory of SMPL using the following metrics: *SIP Error (°)* evaluates the global joint angle error of arms (shoulder angle) and legs (hip angle). *Angle Error (°)* evaluates all global joint angle errors. *Joint Error (cm)* evaluates the root aligned mean per joint position error (MPJPE). *Trans Error @{3m, 6m} (cm)* The global root translation error is computed over all movement pairs that span a distance of x meters. This metric quantifies the deviation from the true trajectory over fixed-distance intervals. Indicates the diversion from the true trajectory over time. We introduce *Dist Err-{4s, 8s, 12s, 16s, 20s}(cm)*, which captures the distance error between two individuals over the corresponding time. It quantifies the multi-human relative translation error, providing more meaningful insights compared to global translation errors for interaction scenarios.

5.1. Evaluation Results

Quantitative Results on the InterHuman Dataset. As shown in Table 1, *GIP* outperforms the baselines across all pose and translation metrics. Our model, incorporating trajectory optimization, produces significantly more accurate and consistent relative translations over time. Additionally, our SSM-based pose estimation method reduces full-body joint angle error by 22% compared to UIP and 33% compared to PIP. Notably, when initialized using the *initial position optimizer*, GIP closely matches the performance of the ground-truth-initialized version, further validating the effectiveness of our approach. Figure 5 shows that our method achieves the lowest cumulative translation error.

Quantitative Results on the GIP-DB Dataset. Compared to finite difference-based simulations of accelerations or distances between virtual sensors, GIP-DB's recordings of *actual* UWB distances and IMU values are noisier as it was recorded from real-world behavior. This in turn makes accurate predictions more challenging. Table 2 highlights *GIP's* benefits over PIP and UIP when dealing with real world noisy data. We improve all pose and translation metrics by a substantial margin, specifically, we reduce the distance error by 72% at 20s. Again, the initial pose optimizer proves its benefit and yields comparable results as the ground truth initialized experiment. We also show that when we use the GIP-DB data to finetune the model, the model performance on angular prediction could be significantly improved.

General Qualitative Results. Figure 6 presents a visual comparison of our proposed method, *GIP*, and the state-of-the-art method, *UIP*. It is clear that *UIP* struggles to estimate relative translations accurately, making it difficult to

Table 1. Results for training on AMASS and evaluation on the InterHuman dataset. When directly comparing with PIP and UIP, we assume the initial position is known for all methods (upper table), as PIP and UIP cannot predict the initial relative translations. We also report results when GIP is not initialized with ground truth but using our *initial position optimizer*, which only affects the translation errors and shows comparable performance to the ground truth initialization. The best results are in **bold**.

Method	SIP Err (°)	Angle Err (°)	Joint Err (cm)	Vertex Err (cm)	Dist Err-4s (cm)	Dist Err-8s (cm)	Dist Err-12s (cm)	Dist Err-16s (cm)	Dist Err-20s (cm)
PIP + D	20.95	14.89	7.98	9.40	48.82	58.25	59.71	63.39	63.72
UIP	19.25	12.89	6.60	7.31	49.61	62.51	59.10	83.16	81.44
GIP (ours)	**18.30**	**9.94**	**5.74**	**6.53**	**3.08**	**3.73**	**4.40**	**1.36**	**1.91**
GIP (ours init.opt.)	**18.30**	**9.94**	**5.74**	**6.53**	3.19	4.00	4.70	4.90	**1.91**

Table 2. Results for training on AMASS data and evaluation on the real-world GIP-DB dataset. When directly comparing with PIP and UIP, we assume the initial position is known for all methods, as PIP and UIP cannot predict the initial translation.

Method	SIP Err (°)	Angle Err (°)	Joint Err (cm)	Vertex Err (cm)	Dist Err-4s (cm)	Dist Err-8s (cm)	Dist Err-12s (cm)	Dist Err-16s (cm)	Dist Err-20s (cm)
PIP + D	30.55	27.40	11.43	12.37	33.22	37.83	46.65	54.56	73.70
UIP	30.18	26.16	10.88	11.50	31.11	37.49	44.24	55.73	74.93
GIP (ours)	27.77	23.34	9.45	10.21	**23.06**	23.79	**21.86**	19.82	20.69
GIP (ours init.opt.)	27.77	23.34	9.45	10.21	23.57	24.40	22.79	**19.34**	20.71
GIP (finetuned)	**18.04**	**17.57**	**8.70**	**9.60**	23.36	**23.27**	22.07	19.56	**19.59**

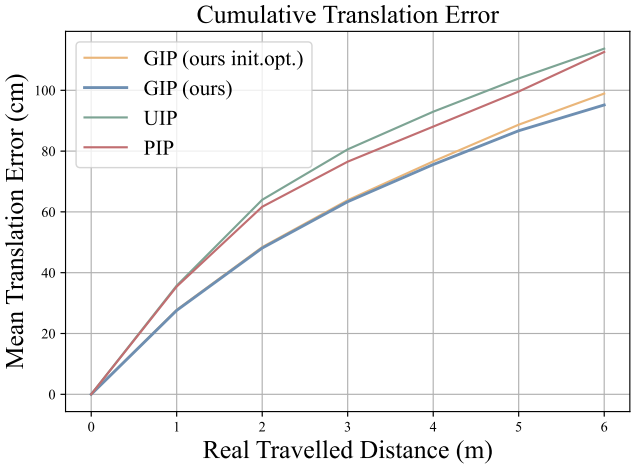

Figure 5. Comparison of translation error on InterHuman.

capture inter-personal interactions. In contrast, *GIP* effectively improves both the relative translations between individuals and the global translations.

Multi-User Scenarios. We conducted an experiment on a synthetic four-person EgoHumans [38] dataset. We keep the test subject fixed and add more people to the optimization process. For example, with three people—p1, p2, and p3—we jointly optimize the group (p1, p2, p3). When a fourth person, p4, is added, we perform two separate optimizations: first on (p1, p2, p3), then on (p1, p2, p4). Adding more people improves translation estimation for the same person thanks to the additional spatial constraints (Table 3).

Table 3. Translation error decreases with more people involved.

N_{people}	1	2	3	4
Trans Error (m)	2.34	1.52	1.20	1.15

Table 4. Ablation study on the InterHuman [46] testset.

Method	SIP Err (°)	Angle Err (°)	Trans-3m (cm)	Trans-6m (cm)	RMSE (cm)	MAE (cm)
w/o init. opt.	18.30	9.94	99.50	116.01	76.50	69.16
w/o traj. opt.	18.30	9.94	76.07	115.32	40.00	36.21
w/LSTM	19.24	12.88	66.60	103.74	42.19	38.53
Ours	**18.30**	**9.94**	**63.85**	**98.93**	**34.86**	**31.11**

6. Ablation Studies

Table 4 shows the impact of *GIP*'s components on pose and translation accuracy. All evaluations assume that the initial position is unknown. The result highlights the importance of the *initial position optimization*, as removing this step significantly degrades translation accuracy (1st row). The trajectory optimizer further improves the translations in a shared frame, improving translation accuracy (2nd row), reducing the translation error @6m by up to 16 cm. To assess the effectiveness of SSM, we substitute SSMs with LSTMs (4th row) while keeping the rest of the pipeline intact (optimizations included). We demonstrate that LSTMs perform worse in both angle errors and translations.

Figure 6. Visual comparison of *GIP* and *UIP*. *GIP* effectively corrects trajectory errors and preserves inter-personal interaction dynamics.

7. Limitations and Discussion

Our method has demonstrated significant improvements over prior approaches for inertial-based motion capture across multiple datasets. Nevertheless, several limitations remain. First, as noted in prior work [5], the UWB noise remains a significant challenge, particularly in multi-person interactions where signal obstruction is common. Nonetheless, our results suggest that future advancements in UWB precision could further enhance motion accuracy. Second, our approach estimates body pose while assuming a mean body shape, without accounting for inter-individual shape variation. Future research could extend this work by incorporating body shape estimation from sparse observations, as explored in [32]. Third, the use of optimization-based inference introduces computational overhead. Although our method converges in fewer than 10 iterations and processes a 30-second motion sequence in 2.04 seconds, this remains a limiting factor for resource-constrained applications. Finally, our method does not explicitly mitigate foot sliding, which can partially arise from our trajectory optimization.

8. Conclusion

Accurate multi-person tracking using sparse sensing is an essential step toward generalized motion tracking and capturing meaningful inter-personal interactions. For this purpose, *Group Inertial Poser* overcomes the drift and lack of positional references in previous inertial methods and demonstrates multi-person motion tracking by augmenting inertial measurements with between-sensor distances. We leverage these novel constraints to mitigate drift and improve relative translation accuracy. Beyond improved single-person pose estimation with a novel SSM model, GIP robustly tracks two individuals using sparse IMU+UWB sensors and accurately estimates relative trajectories. Unlike previous methods, GIP allows people to start from arbitrary positions and automatically determines their initial relative positions via a two-step optimization. GIP offers quantitative and qualitative advantages—it improves accuracy but also preserves crucial inter-personal interaction dynamics. Additionally, we introduce *GIP-DB*, the first IMU+UWB dataset designed for multi-person tracking using sparse inertial sensors. Evaluated with InterHuman and GIP-DB, our approach consistently outperforms existing methods in accuracy and robustness across synthetic and real-world data. Collectively, our work highlights the potential of IMU+UWB fusion for multi-person motion tracking, opening new opportunities for real-world applications.

Acknowledgments

We thank all participants of our data capture. We thank Adnan Harun Dogan for his help with preparing the dataset. Yi-Chi Liao was supported by the ETH Zurich Postdoctoral Fellowship Program. Ying Xue was in part supported by the Swiss National Science Foundation (Grant No. 10004941).

References

[1] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5 (2):1–23, 2021. 2

[2] Simone Angarano, Vittorio Mazzia, Francesco Salvetti, Giovanni Fantin, and Marcello Chiaberge. Robust ultra-wideband range error mitigation with deep learning at the edge. *Engineering Applications of Artificial Intelligence*, 102:104278, 2021. 3

[3] Apple. Nearby interaction with uwb, 2023. 3

[4] 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)*. 3, 6

[5] 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*, New York, NY, USA, 2024. Association for Computing Machinery. 3, 4, 6, 8

[6] Aditya Arun, Tyler Chang, Yizheng Yu, Roshan Ayyalasomayajula, and Dinesh Bharadia. Real-time low-latency tracking for uwb tags. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, page 611–612, New York, NY, USA, 2022. Association for Computing Machinery. 3

[7] Thomas Augsten, Konstantin Kaefer, René Meusel, Caroline Fetzer, Dorian Kanitz, Thomas Stoff, Torsten Becker, Christian Holz, and Patrick Baudisch. Multitoe: high-precision interaction with back-projected floors based on high-resolution multi-touch input. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, page 209–218, New York, NY, USA, 2010. Association for Computing Machinery. 2

[8] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 2

[9] Valentín Barral, Carlos J. Escudero, José A. García-Naya, and Roberto Maneiro-Catoira. Nlos identification and mitigation using low-cost uwb devices. *Sensors*, 19(16), 2019. 3

[10] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 1, 2

[11] Alan Bränzel, Christian Holz, Daniel Hoffmann, Dominik Schmidt, Marius Knaust, Patrick Lühne, René Meusel, Stephan Richter, and Patrick Baudisch. Gravityspace: tracking users and their poses in a smart room using a pressure-sensing floor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 725–734, New York, NY, USA, 2013. Association for Computing Machinery. 2

[12] Yanjun Cao, Chenhao Yang, Rui Li, Alois Knoll, and Giovanni Beltrame. Accurate position tracking with a single uwb anchor. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2344–2350, 2020. 3

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[14] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 1, 2

[15] Bowen Cheng, Bin Xiao, Jingdong Wang, Humphrey Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394, 2019. 2

[16] J. A. Corrales, F. A. Candelas, and F. Torres. Hybrid tracking of human operators using imu/uwb data fusion by a kalman filter. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 193–200, 2008. 3

[17] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. Smartposer: Arm pose estimation with a smartphone and smartwatch using uwb and imu data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2023. Association for Computing Machinery. 3

[18] Andreas Döring, Di Chen, Shanshan Zhang, Bernt Schiele, and Jürgen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20963–20972, 2022. 2

[19] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[20] Daquan Feng, Chunqi Wang, Chunlong He, Yuan Zhuang, and Xiang-Gen Xia. Kalman-filter-based integration of imu and uwb for high-accuracy indoor positioning and navigation. *IEEE Internet of Things Journal*, 7(4):3133–3146, 2020. 3

[21] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3, 4

[22] 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. *arXiv preprint arXiv:2409.13426*, 2024. 2

[23] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019. 1, 2

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[25] Jeroen D. Hol, Fred Dijkstra, Henk Luinge, and Thomas B. Schon. Tightly coupled uwb/imu pose estimation. In *2009 IEEE International Conference on Ultra-Wideband*, pages 688–692, 2009. 3

[26] 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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2025. 3

[27] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional directed graph convolution for 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 602–611, 2021. 2

[28] 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. 2, 6

[29] 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. 2

[30] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 443–460. Springer, 2022. 2

[31] 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.

[32] 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*. Springer, 2024. 2, 8

[33] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[34] 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 conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1, 2

[35] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[36] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11510–11520, 2021. 3

[37] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 3

[38] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19807–19819, 2023. 7

[39] 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. 1, 2

[40] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019. 1

[41] Sangmin Lee, Seungho Yoo, Joon Yeop Lee, Seongjoon Park, and Hwangnam Kim. Drone positioning system using uwb sensing and out-of-band control. *IEEE Sensors Journal*, 22(6):5329–5343, 2022. 3

[42] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 2

[43] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11354–11361, 2020. 1

[44] 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. 2

[45] Shuang Li, Jiaxi Jiang, Philipp Ruppel, Hongzhuo Liang, Xiaojian Ma, Norman Hendrich, Fuchun Sun, and Jianwei Zhang. A mobile robot hand-arm teleoperation system by vision and imu. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10900–10906. IEEE, 2020. 3

[46] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21, 2024. 6, 7

[47] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989. 5

[48] Huajun Liu, Xiaolin Wei, Jinxiang Chai, Inwoo Ha, and Taehyun Rhee. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*, pages 133–140, 2011. 3

[49] 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. 3, 4

[50] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation, 2024. 2

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

[52] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 6

[53] 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*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[54] Mark W. Mueller, Michael Hamer, and Raffaello D'Andrea. Fusing ultra-wideband range measurements with accelerometers and rate gyroscopes for quadrocopter state estimation, 2015. 3

[55] Deepak Nagaraj, Erik Schake, Patrick Leiner, and Dirk Werth. An rnn-ensemble approach for real time human pose estimation from sparse imus. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, pages 1–6, 2020. 2

[56] Noitom. https://www.noitom.com/, 2024. 2

[57] Aitor Ochoa-de Eribe-Landaberea, Leticia Zamora-Cadenas, Oier Peñagaricano-Muñoa, and Igone Velez. Uwb and imu-based uav's assistance system for autonomous landing on a platform. *Sensors*, 22(6):2347, 2022. 3

[58] Optitrack. https://wwww.optitrack.com/, 2023. 1, 2, 6

[59] 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*, New York, NY, USA, 2023. Association for Computing Machinery. 3

[60] Jorge Peña Queralta, Li Qingqing, Fabrizio Schiano, and Tomi Westerlund. Vio-uwb-based collaborative localization and dense scene reconstruction within heterogeneous multi-robot systems, 2022. 3

[61] Samsung. Galaxy s22 ultra specifications, 2023. 3

[62] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11059–11068, 2022. 2

[63] 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. 1

[64] 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. 2

[65] Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. Hoov: Hand out-of-view tracking for proprioceptive interaction using inertial sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023. 3

[66] D. Tome, P. Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7727–7737, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2

[67] Vu Tran, Zhuangzhuang Dai, Niki Trigoni, and Andrew Markham. Deepcir: Insights into cir-based data-driven uwb error mitigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13300–13307, 2022. 3

[68] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017. 2

[69] Ubisense. https://ubisense.com/, 2023. 3

[70] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[71] Vicon. https://wwww.vicon.com/, 2023. 1, 2

[72] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3):35–es, 2007. 3

[73] 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. 2

[74] 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

[75] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13031–13040, 2023. 2

[76] 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. Springer, 2024. 1

[77] 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. 1, 2

[78] Xsens. https://www.xsens.com, 2024. 2, 6

[79] Hao Xu, Yichen Zhang, Boyu Zhou, Luqi Wang, Xinjie Yao, Guotao Meng, and Shaojie Shen. Omni-swarm: A decentralized omnidirectional visual–inertial–uwb state estimation system for aerial swarms. *IEEE Transactions on Robotics*, 38 (6):3374–3394, 2022. 3

[80] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, pages 265–275. Wiley Online Library, 2021. 2

[81] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2, 6

[82] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 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*, pages 13167–13178, 2022. 2, 4, 6

[83] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 3

[84] 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. 2

[85] 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. 2

[86] 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. 2

[87] Minghui Zhao, Tyler Chang, Aditya Arun, Roshan Ayyalasomayajula, Chi Zhang, and Dinesh Bharadia. Uloc: Low-power, scalable and cm-accurate uwb-tag localization and tracking for indoor applications. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3), 2021. 3

[88] Shuaikang Zheng, Zhitian Li, Yuanli Yin, Yunfei Liu, Haifeng Zhang, Pengcheng Zheng, and Xudong Zou. Multi-robot relative positioning and orientation system based on uwb range and graph optimization. *Measurement*, 195: 111068, 2022. 3