ToF-IP: Time-of-Flight Enhanced Sparse Inertial Poser for Real-time Human Motion Capture

Yuan Yao[1] **Shifan Jiang**[1] **Yangqing Hou**[1] **Chengxu Zuo**[1] **Xinrui Chen**[1]

Shihui Guo[1]* **Yipeng Qin**[2]

[1] School of Informatics, Xiamen University, China
[2] School of Computer Science & Informatics, Cardiff University, UK

Abstract

Sparse inertial measurement units (IMUs) provide a portable, low-cost solution for human motion tracking but struggle with error accumulation from drift and sensor noise when estimating joint position through time-based linear acceleration integration (i.e., *indirect* measurement). To address this, we propose ToF-IP, a novel 3D full-body pose estimation system that integrates Time-of-Flight (ToF) sensors with sparse IMUs. The distinct advantage of our approach is that ToF sensors provide *direct* distance measurements, effectively mitigating error accumulation without relying on *indirect* time-based integration. From a hardware perspective, we maintain the portability of existing solutions by attaching ToF sensors to selected IMUs with a negligible volume increase of just 3%. On the software side, we introduce two novel techniques to enhance multi-sensor integration: (i) a Node-Centric Data Integration strategy that leverages a Transformer encoder to explicitly model both intra-node and inter-node data integration by treating each sensing node as a token; and (ii) a Dynamic Spatial Positional Encoding scheme that encodes the continuously changing spatial positions of wearable nodes as motion-conditioned functions, enabling the model to better capture human body dynamics in the embedding space. Additionally, we contribute a 208-minute human motion dataset from 10 participants, including synchronized IMU-ToF measurements and ground-truth from optical tracking. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches such as PNP, achieving superior accuracy in tracking complex and slow motions like Tai Chi, which remains challenging for inertial-only methods.

1 Introduction

Sparse inertial measurement units (IMUs) have emerged as a promising solution for human motion tracking due to their portability, low cost, and camera-free nature [8, 11, 20, 42]. However, despite their potential, sparse IMUs face inherent numerical challenges due to their *indirect* method of position estimation. Specifically, sparse IMUs estimate velocity and position by time-based linear acceleration integration, a process highly prone to error accumulation from drift and sensor noise. These errors are further amplified by the task's reliance on human body forward kinematics, where positional inaccuracies in intermediate joints propagate along the kinematic chain, leading to greater errors at terminal joints. Consequently, accurately tracking subtle positional changes of key joints during low-velocity motions (where the motion-signal-to-noise ratio is low) and enhancing the long-term stability of sparse IMU systems remain persistent and critical challenges in this field [37].

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

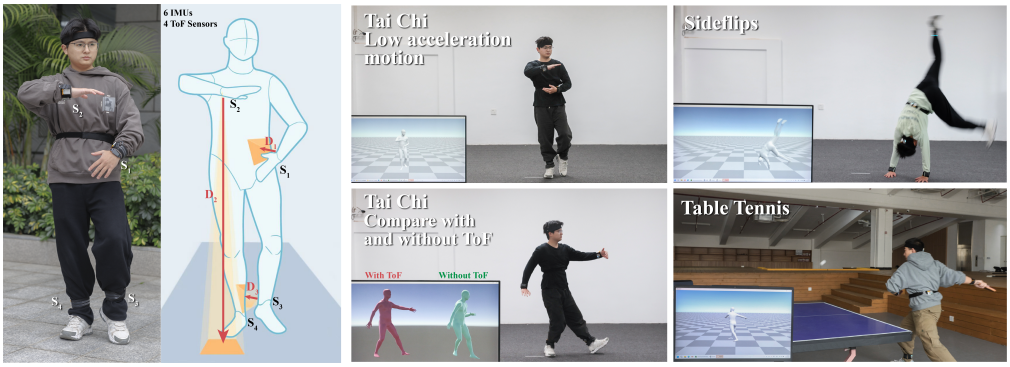

Figure 1: ToF-IP integrates distance maps from four Time-of-Flight (ToF) sensors to overcome the drift and error accumulation inherent in inertial-only motion capture, enabling more accurate and stable motion tracking even for challenging motions such as slow, controlled sequences (e.g., Tai Chi) and rapid, complex actions (e.g., sideflips).

To date, most existing methods focus on providing software-based solutions to the aforementioned challenge. For example, deep learning approaches, such as bidirectional RNNs, have been developed to regress IMU data into pose sequences [8]. Subsequent RNN-based methods further improved pose prediction accuracy and integrated global position estimation [36], with some incorporating more precise dynamic models [37]. Beyond RNNs, alternative architectures have emerged, including attention-based models for capturing physical motion during stationary phases [11]. Nevertheless, software solutions are inherently limited, as they still rely on *indirect* pose estimation and can only mitigate, rather than fully resolve, the problem of error accumulation.

In this paper, we present ToF-IP, a novel 3D full-body pose estimation system that addresses the abovementioned challenge by integrating Time-of-Flight (ToF) distance sensors with sparse IMUs. The distinct advantage of our approach is that ToF sensors provide *direct* distance measurements without relying on *indirect* time-based integration, offering learnable solution space constraints for the position of limb-end joints. A key innovation of our ToF-IP lies in its hardware design, which maintains the lightweight and portable nature of existing sparse IMU systems. Specifically, without altering the standard 6-IMU layout, we incorporate 4 highly integrated ToF sensors based on single-photon avalanche diodes (SPAD) [26] directly onto the IMU circuit boards equipped at left forearm, right forearm, left lower leg, and right lower leg of the human body, ensuring minimal impact on wearability. On the software side, we fully harness the potential of ToF-inertial sensing by proposing a unified Transformer-based framework with two key innovations. First, we introduce a Node-Centric Data Integration strategy that explicitly captures the hierarchical structure of multi-sensor data. Unlike prior methods that flatten all sensor inputs into a single vector-wise organization (discarding the spatial and structural semantics of node), our approach represents each sensing node as an independent token. This token is constructed through intra-node data integration of ToF depth, IMU acceleration and orientation data. These node tokens are then contextually integrated using the self-attention mechanism of a Transformer encoder, which naturally facilitates inter-node communication and dynamic weighting based on task-relevant dependencies. Second, we propose a Dynamic Spatial Positional Encoding (Dyn-PE) method tailored to the unique challenges of wearable sensing. Unlike traditional positional encodings in NLP or vision tasks that assume static or grid-based positions, the spatial configuration of wearable nodes evolves continuously with human motion. To capture this, Dyn-PE models each node's position as a learnable function of global motion signals, generating time-varying encodings that reflect the node's physical displacement in space. This dynamic encoding enhances the model's spatial awareness, allowing it to better resolve ambiguous interactions and motion patterns across nodes. Extensive experimental results show that, compared to state-of-the-art (SOTA) methods, our approach significantly improves joint position estimation, achieving superior accuracy in tracking complex and slow movements like Tai Chi.

In summary, our contributions include:

- We design an in-situ enhanced bimodal wearable sensing platform for 3D full-body tracking, retaining the conventional layout of 6 sensing nodes. The platform allows for the flexible use of either single IMU sensing or IMU+ToF bimodal sensing, with only a 3% increase in volume.

- We propose *ToF-IP*, a novel Transformer-based inertial-ToF motion capture framework that introduces two key innovations on the software side: (i) a Node-Centric Data Integration strategy that preserves the structural semantics of multi-sensor data by treating each sensing node as a token and hierarchically integrating intra- and inter-node information via self-attention; and (ii) a Dynamic Spatial Positional Encoding scheme that models the continuously evolving spatial positions of wearable nodes as motion-conditioned functions, enhancing spatial awareness and robustness to body movement variations. Our approach demonstrates substantial improvements over state-of-the-art methods, delivering higher positional precision and more accurate joint angle estimation, particularly in the upper limbs and legs.
- We propose *ToF-IP-DB*, a large dataset containing over 20 types of motion activities, 208 minutes (749,000 frames) collected from 10 participants (3 male, 7 female), including dynamic motions such as dances and aerobics, as well as slow-paced movements like Tai Chi and Baduanjin. This dataset uniquely combines synchronized ToF distance maps, 6-DoF IMU signals, and SMPL reference poses, with GT motion data.

2 Related Work

2.1 Pose Estimation Using Inertial Sensors

With the rapid advancement of MEMS technology [9], IMUs (Inertial Measurement Units) have become smaller, more power-efficient, and affordable. This has led to numerous works leveraging IMUs for human pose estimation. Despite their independence from external environments, the working principle of IMUs—using accelerometers, gyroscopes, and magnetometers to compute orientation—limits their accuracy. In the commercial market, motion capture systems employing 17–19 IMUs for human pose estimation exist [35, 23], but they require operation in uniform magnetic field environments and have limited usage durations, as accumulated integration errors can lead to model collapse.

The pursuit of lightweight solutions has spurred research into using sparse IMUs—typically six sensors placed at limb extremities. The advent of the SMPL [18] and AMASS [19] datasets has enabled the creation of large-scale mocap/IMU-aligned datasets. Synthetic continuous acceleration and rotation data were generated by placing virtual IMUs on specific body parts in the AMASS dataset, facilitating pose estimation in both offline [31] and real-time settings. Deep learning methods, such as bidirectional RNNs, were designed to regress IMU data to pose sequences [8]. Building on this, RNN-based approaches have improved pose prediction accuracy and incorporated global position estimation [36], with some integrating more precise dynamic models, such as those in [37]. Other methods explore alternative network architectures, including attention-based models for learning physical motion during stationary points [11] and spatiotemporal modules for more accurate pose estimation [34]. Efforts have also been made to enhance comfort and convenience. For example, some methods use VR headsets, smartphones, and other wearable devices to estimate upper-body poses and predict lower-body movements. Zuo et al. integrated IMUs into loose clothing [42], effectively regressing upper-body poses while mitigating artifacts caused by fabric-induced jitter.

Despite such success, challenges remain in pose estimation using IMUs alone [22, 10, 14, 28]. These include IMUs' inability to directly measure velocity or position, accumulation of drift errors [15, 13], and reliance on forward kinematics models [25] composed of bones and joints. Sparse IMU layouts—especially those at the extremities of limbs, such as hands and legs—face difficulties due to the higher degrees of freedom and heavier prediction tasks borne by individual sensors.

2.2 Time-of-Flight Distance Sensors

To mitigate these issues, recent research has explored hybrid solutions that integrate additional sensor modalities [1, 3, 17]. One common strategy is to incorporate global positioning or localization techniques to provide absolute positional constraints [7]. For instance, Zihajehzadeh et al. [41] combined IMUs with ultrawideband (UWB) localization to eliminate yaw angle drift in lower-body tracking, leveraging UWB's absolute position measurements to correct inertial estimates. Similarly, Liu et al. [16] used a micro-flow sensor to estimate motion velocity, enabling accurate extraction of gravitational acceleration from accelerometer data and improving posture tracking stability.

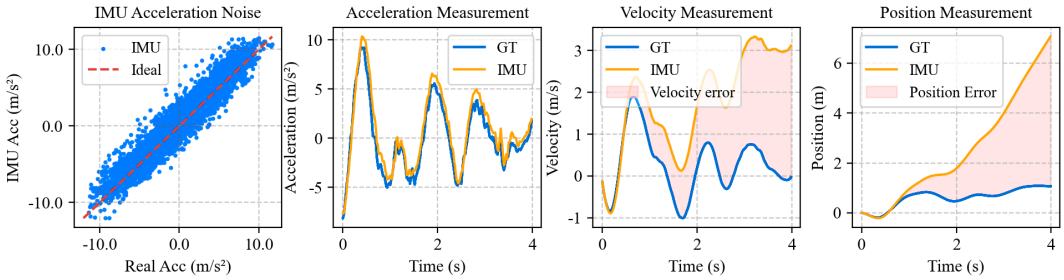

Figure 2: Illustration of error accumulation in velocity and position. Even slight noise in IMU acceleration leads to significant integration errors in velocity and position within just 4 seconds. Acceleration signals are taken from the TIC dataset [43], with ground-truth provided for comparison.

Another class of methods exploits vision-based and depth sensors to compensate for inertial drift [30, 32, 5, 21, 38]. Depth sensors such as LiDAR and structured-light cameras have been explored for markerless tracking. ToF sensors broadly refer to methods for precise distance measurement based on the time taken by light pulses or continuous waves to travel [6, 12]. Unlike vision-based methods, ToF sensors are resistant to variations in lighting and occlusions, providing robust depth measurements even in challenging conditions [33, 24]. Some studies [27, 4] have integrated ToF sensors into external environments for global pose estimation, but their application in wearable systems remains underexplored.

By carefully balancing factors like power consumption and heat dissipation, we augment the conventional six-IMU sparse layout with four low-resolution depth Time-of-Flight (ToF) sensors. This hardware integration enhances inertial pose estimation by capturing inter-limb distances and contact points with the environment, providing real-time, in-situ constraints that mitigate positional drift while preserving the wearability and portability characteristic of traditional inertial tracking systems.

3 Inherent Limitation of Sparse Inertial-only Motion Capture

Under a sparse IMU configuration, only a subset of joint orientations can be directly measured. To compensate for missing measurements, acceleration data is commonly used as an additional input [8], as it carries implicit cues about joint positions that can aid in inferring unobserved joint orientations. In principle, joint positions can be obtained by double-integrating the acceleration signals over time. However, in practice, real-world acceleration measurements are prone to various sources of error, such as sensor noise and signal drift, inevitably resulting in significant error accumulation over time:

Proposition 3.1 (Error Accumulation Analysis)**.** *Following standard statistical practice, we assume that the acceleration measurement error at any timestamp $\tau \in (0, t)$ follows a normal distribution $\epsilon_a(\tau) \sim \mathcal{N}(\mu, \sigma^2)$. Then, we have:*

- *Distribution of joint velocity error:* $\epsilon_v(t) \sim \mathcal{N}(\mu t, \sigma^2 t)$
- *Distribution of joint position error:* $\epsilon_s(t) \sim \mathcal{N}(\frac{1}{2}\mu t^2, \frac{1}{2}\sigma^2 t^2)$

Proof. The proof is provided in Section 2 of the supplementary material. □

Proposition 3.1 and Fig. 2 shows that the joint position error grows quadratically over time, highlighting an inherent limitation of sparse inertial-only solutions.

4 Method

To address the inherent limitation of sparse inertial-only solutions discussed in Sec. 3, we propose integrating Time-of-Flight (ToF) sensors, which provide *direct* distance measurements to mitigate error accumulation from time-based integration.

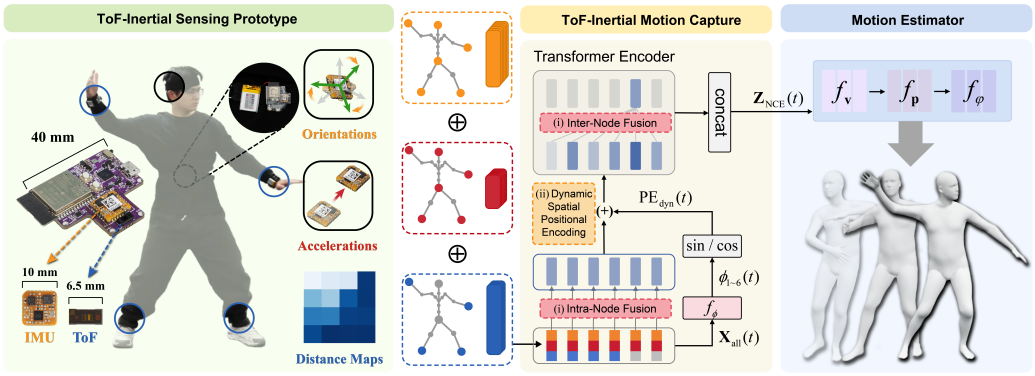

Figure 3: Illustration of Our ToF-Inertial Sensing Prototype and Motion Capture Method. **Left:** Our ToF-Inertial sensing prototype comprises 6 IMU-attached sensing nodes, with 4 of them (located on the left / right forearm and left / right lower-leg) additionally integrated with ToF sensors. **Middle:** Our ToF-Inertial motion capture method comprises two novel techniques: (i) a node-centric data integration strategy based on a Transformer encoder, encompassing intra-node and inter-node data integration between ToF-Inertial sensor; (ii) a dynamic spatial positional encoding to adapt to the dynamic changes in the spatial positions of sensing nodes during motion capture; **Right:** Three sequentially connected LSTM networks ($f_{\text{v}}, f_{\text{p}}, f_\phi$) serve as motion estimators to transform the integrated sensing node data into human body movements.

4.1 ToF-Inertial Sensing Prototype

As shown in Fig. 3 (left), to maximize compatibility with existing inertial-only solutions and minimize impact on wearability, i) we adopt the standard 6-node layout used in prior works [36, 37, 11, 39] and place 6 IMUs on the left/right forearms, left/right lower legs, pelvis and head, respectively; ii) we integrate 4 ToF sensors into the IMU nodes on the left/right forearms and left/right lower legs. Specifically, we mount the ToF sensors on the inner wrists and rear ankles to capture distance measurements from distal joint endpoints to nearby body parts and the ground. Note that we omit ToF sensors from the head and pelvis nodes, as these positions seldom observe relevant surfaces.

4.2 ToF-Inertial Motion Capture

Overview. As shown in Fig. 3 (middle, right), our ToF-Inertial Motion Capture framework adopts a Transformer-based architecture that incorporates two key innovations for integrating IMU and ToF data: (i) Node-Centric Data Integration, which explicitly models intra- and inter-node interactions through tokenized node representations; and (ii) Dynamic Spatial Positional Encoding, which encodes the time-varying spatial positions of sensing nodes using motion-conditioned functions. Following [36, 37, 39], we further employ a cascade of three LSTM networks as motion estimators.

4.2.1 Node-Centric Data Integration

Conditions for Effective ToF Integration. Although the direct distance measurements provided by ToF establish a data foundation for improving joint position estimation, their effectiveness hinges on two key data integration conditions:

- **[Intra-node Integration]** Each ToF sensor must integrate orientation and acceleration data from its co-located inertial sensor within the sensing node to determine the viewing direction and motion state information;
- **[Inter-node Integration]** Since both the motion of the sensing node itself and the captured object can cause changes in the ToF distance map, integrating measurements from other sensing nodes is required to distinguish absolute motion from relative motion.

Limitation of Existing Data Integration Method. Existing methods typically use fully-connected networks (FCNs) for data integration. This approach directly flattens data from multiple nodes into a single vector, losing the inherent data structure organized by sensing nodes and overlooking

intra-node integration. Formally, consider an embedded feature f_i from an FCN layer, we have:

$$f_i = \text{Flatten}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}) \cdot W_i + b_i \tag{1}$$

where $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ are the data of the N nodes. In this formulation, the linear combination of $\mathbf{x}^{(j)}, j = 1, ..., N$ inherently achieves inter-node integration, but overlooks the intra-node data integration within each $\mathbf{x}^{(j)}$.

Node-Centric Data Integration. To address this challenge, we propose replacing fully-connected networks used in previous methods with a Transformer Encoder. In this scheme, each sensing node is converted into an independent token, with all encoding performed at the token (node) level. Consider the encoding process for an arbitrary output token $\mathbf{z}_T^{(n)}$ of node n, which unfolds in two structured steps to explicitly perform intra-node and inter-node data integration:

- *1) Intra-node Data Integration via Tokenization.* The multi-modal data of each node $\mathbf{x}^{(n)} = [d^{(n)}, a^{(n)}, R^{(n)}]$ is encoded into an intra-node token $\mathbf{z}_{\text{intra}}^{(n)} \in \mathbb{R}^{d_{\text{model}}}$ as follow:

$$\mathbf{z}_{\text{intra}}^{(n)} = f_T(\mathbf{x}^{(n)}) + \text{PE}^{(n)} \tag{2}$$

  where f_T is a tokenize function, d_{model} is size of token. $d \in \mathbb{R}^{16}$, $a \in \mathbb{R}^3$, and $R \in \mathbb{R}^{3\times3}$ are ToF depth maps, IMU acceleration and orientation, respectively, $\text{PE}^{(n)}$ is positional encoding of the n-th sensing node. For IMU-only sensing node (head and hip), d is set to zeros.

- *2) Inter-node Data Integration via Self-Attention.* The intra-node tokens $\{\mathbf{z}_{\text{intra}}^{(1)}, \ldots, \mathbf{z}_{\text{intra}}^{(N)}\}$ are then integrated via Transformer Encoder's self-attention mechanism, which computes inter-node interaction weights $A \in \mathbb{R}^{N\times N}$ applied to intra-node tokens:

$$\mathbf{z}_{\text{inter}}^{(i)} = \mathbf{z}_{\text{intra}}^{(i)} + \sum_{j=1}^{N} A_{ij}(\mathbf{z}_{\text{intra}}^{(j)} \cdot \mathbf{W}_{\text{v}}) \tag{3}$$

  where $\mathbf{W}_{\text{v}}$ projects tokens into value space, and the $\mathbf{z}_{\text{intra}}^{(i)}$ is residual connection term. Then each $\mathbf{z}_{\text{inter}}^{(i)}$ will go through layer norm (LN) and feed-forward network (FFN) in the Transformer Encoder and concatenated to produce the final embedding vector $\mathbf{Z}_{\text{NCI}} \in \mathbb{R}^{(N\times d_{\text{model}})}$.

4.2.2 Dynamic Spatial Positional Encoding

Static Positional Encoding. As Eq. 2 shows, positional encoding PE is a fundamental process to incorporating position information into tokens. Existing static positional encoding methods assign **static** positional values ϕ to each token via an addition operation. For example, the static positional encoding in the original Transformer [29] is as follows:

$$\text{PE}_{\text{sta}}^{(n,2i,2i+1)} = [\sin(\omega_i \cdot \phi_{pos}), \cos(\omega_i \cdot \phi_{pos})], \quad \phi_{pos} = n, \quad \omega_i = 10000^{-2i/d_{\text{model}}} \tag{4}$$

where n is the index of the input token, d_{model} is the dimension of the token, ϕ_{pos} is the static positional value determined by the token index (sequential position). The n-th sensing node token processed by static positional encoding can then be represented as $\mathbf{z}_{\text{intra}}^{(n)} = f_T(\mathbf{x}^{(n)}) + \text{PE}_{\text{sta}}^{(n)}$, where $\text{PE}_{\text{sta}}^{(n)}$ is static positional encoding calculated by Eq. 4.

Our Dynamic Spatial Positional Encoding. However, unlike the static and discrete positions in natural language processing and computer vision applications (such as word order or pixel grid coordinates), the positions of sensing nodes are **dynamic** and continuous **spatial** positions that change with human body movements. Considering these characteristic, we model the positions of sensing nodes as continuous functions of motion signals and propose dynamic spatial positional encoding (Dyn-PE) as follows:

$$\text{PE}_{\text{dyn}}^{(n,2i,2i+1)}(t) = [\sin(\omega_i \cdot \phi_n(t)), \cos(\omega_i \cdot \phi_n(t))], \quad \phi_n(t) = f_\phi^n(\mathbf{X}_{\text{all}}(t)), \quad \omega_i = 10000^{-2i/d_{\text{model}}} \tag{5}$$

where $\phi_n(t)$ are the dynamic positional value of sensing node n at time t ($n = 1, 2, ..., 6$), $\mathbf{X}_{\text{all}}(t) = d^{1\to6}(t), a^{1\to6}(t), R^{1\to6}(t)$ are data of all six sensing nodes, serving as human motion signal. The $f_\phi^1, ..., f_\phi^n$ are position estimation functions implemented with a 2-layer MLP. Then, our Dynamic Spatial Positional Encoding can be integrated into Eq. 2 as: $\mathbf{z}_{\text{intra}}^{(n)} = f_T(\mathbf{x}^{(n)}) + \text{PE}_{\text{dyn}}^{(n)}(t)$.

4.2.3 Motion Estimators

Following [36, 37, 39], we feed the sensor data embedding $\mathbf{z}_{\text{NCI}}$ as the shared input to three sequential motion estimators for joint velocity $\mathbf{v} \in \mathbb{R}^{J\times 3}$, joint position $\mathbf{p} \in \mathbb{R}^{J\times 3}$ and joint rotation $\boldsymbol{\varphi} \in \mathbb{R}^{J\times 6}$:

$$\mathbf{v}(t) = f_{\mathbf{v}}(\mathbf{Z}_{\text{NCI}}(t)) \quad \mathbf{p}(t) = f_{\mathbf{p}}(\mathbf{Z}_{\text{NCI}}(t), \mathbf{v}(t)) \quad \boldsymbol{\varphi}(t) = f_{\boldsymbol{\varphi}}(\mathbf{Z}_{\text{NCI}}(t), \mathbf{p}(t)) \tag{6}$$

where $J = 18$ is the total number of tracked joints, $f_{\mathbf{v}}$, $f_{\mathbf{p}}$, $f_{\boldsymbol{\varphi}}$ are RNN-based motion estimators. These motion estimators are trained in a supervised manner using the following motion loss:

$$\mathcal{L}_{\text{motion}} = ||\mathbf{v}(t) - \mathbf{v}^{\text{GT}}(t)||_2^2 + ||\mathbf{p}(t) - \mathbf{p}^{\text{GT}}(t)||_2^2 + ||\boldsymbol{\varphi}(t) - \boldsymbol{\varphi}^{\text{GT}}(t)||_2^2 \tag{7}$$

where GT denotes the ground truth value.

4.2.4 Global Translation Tracking

The global translation tracking in this work is powered by velocity output of Motion Estimator and SMPL kinematic model. Specifically, we first compute the estimated velocity of the four joint endpoints (left and right forearms and lower legs) equipped with ToF-Inertial nodes and convert into pseudo stationary label q_s:

$$q_s^{(i)}(t) = \begin{cases} 1 & \text{if } ||v^{(i)}(t)||_2 < \epsilon \\ 1 - \frac{||v^{(i)}(t)||_2 - \epsilon}{0.2} & \text{if } \epsilon \leq ||v^{(i)}(t)||_2 < \epsilon + 0.2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Where i denote endpoint index, the ϵ is a cut-off threshold to filter small jitter in $||v^{(i)}(t)||_2$ (we use $\epsilon = 0.05m/s$ in this work). Subsequently, based on $\varphi(t)$ provided by the Motion Estimator, we calculate the root translation relative to the endpoints is stationary, denoted as $s_{FK}^{(i)}(t)$, using the forward kinematics:

$$s_{FK}^{(1,2,3,4)}(t) = FK(\varphi(t - \Delta t)) - FK(\varphi(t)) \tag{9}$$

Where Δt is time gap of 2 continuous captures. Then we define FK-based translation as follow:

$$s_{FK}(t) = \frac{\sum_{i=1}^{4} q_s^{(i)}(t) \cdot s_{FK}^{(i)}(t)}{\sum_{i=1}^{4} q_s^{(i)}(t)} \tag{10}$$

Similar to previous works[36], we fusion the $s_{FK}(t)$ with root velocity provide by Motion Estimator to obtain the final translation estimation:

$$\begin{aligned} s_{NN}(t) &= v_{root}(t - \Delta t) \cdot \Delta t \\ s(t) &= (1 - q_m) \cdot s_{FK}(t) + q_m \cdot s_{NN}(t) \end{aligned} \tag{11}$$

Where $q_m = \textbf{min}(q_s^{(1)}, ..., q_s^{(4)})$ denotes the pseudo label of full-body moving (e.g., jumping on the air, sliding), and s_{NN} denotes translation estimation base on neural network (Motion Estimators).

5 Experiment

5.1 Experimental Setup

Synthetic Dataset. We leverage the AMASS dataset [19] to synthesis a large-scale paired ToF-IMU-Motion data for motion estimators pre-training, which includes both IMU and ToF data simulation.

- **IMU Data Simulation**: Similar to previous works [8, 36, 11, 40], the IMU orientation and acceleration are calculated based on the global joint orientation of the SMPL [18] model and the trajectory of the selected mesh vertices.
- **ToF Data Simulation**: ToF data simulation is implemented using Unity. We simulate the ToF sensor with virtual depth cameras positioned on the rendered SMPL body, aligned with the hardware wearing setup. The original depth maps are then down-sampled into 4×4 to fit the configuration of ToF. More detailed settings are provided in the supplementary materials.

Table 1: Comparison of methods on DIP and our ToF-IP-DB datasets across multiple error metrics. We evaluated our ToF-IP on DIP with additional ToF synthesis.

Method	ToF-IP-DB					DIP (with synthesis ToF)				
	SIP Err	Ang Err	Pos Err	EndPos Err	Jitter	SIP Err	Ang Err	Pos Err	EndPos Err	Jitter
Transpose	20.73	14.51	7.58	13.58	0.18	17.06	8.86	6.03	8.73	1.11
TIP	20.37	14.58	7.63	13.97	0.17	16.90	9.07	5.63	8.27	1.56
PIP	20.22	13.85	7.32	12.77	0.12	15.33	8.78	5.12	7.78	0.17
DynaIP	19.04	13.33	7.26	13.05	0.16	13.78	7.07	4.98	7.44	0.18
PNP	18.52	13.23	6.86	12.39	0.12	13.71	8.75	4.97	7.49	0.17
ToF-IP(Ours)	**17.26**	**12.09**	**6.31**	**11.41**	0.12	**13.62**	**6.75**	**4.59**	**6.65**	0.17

ToF-IP-DB Dataset. We collected a full-body motion capture dataset, containing over 20 types of movements, 208 minutes (749,000 frames) from 10 participants (3 male and 7 female) with heights ranging from 170 cm to 185 cm. All participants were informed about the purpose of the experiment and signed consent agreements. Participants were required to perform the following steps for data collection:

- Simultaneously wears an optical motion capture suit (with multiple optical markers attached) along with our 6-node ToF-Inertial motion capture prototypes.
- Performs a T-pose for IMU calibration.

During data collection, the participant is asked to perform diverse types of motion, e.g., dances, aerobics, and daily social activities. Sensor data and motion data are collected synchronously at 60Hz. The motion data is captured using the NOKOV marker-based optical motion capture system, including full-body pose and global translation. Each collection session lasts 6–10 minutes.

Training Settings. All our experiments run on a PC with an Intel(R) Core(TM) i7-13700KF CPU and an NVIDIA RTX 4080 GPU. The model is implemented using PyTorch 1.12.1 with CUDA 11.3. We use the Adam optimizer with a learning rate of $lr = 1 \times 10^{-3}$ and weight decay of $lr = 1 \times 10^{-6}$ during n epochs training. The batch size was set to 512.

Metrics. We use the following five error metrics to evaluate the accuracy and quality of captured motion: ***1) Angular Error*** (°), which represents the global rotation error of all joints; ***2) Positional Error*** (*cm*), which is the joint position error of all joints; ***3) SIP Error*** (°), defined as the global rotation error of hips and shoulders; ***4) Endpoints Positional Error*** (*cm*), positional errors of the four ToF sensor attached joints (the left and right wrist and ankle); ***5) Jitter*** (*km/s*^{3}), denoting the jerk (time derivative of acceleration) of all body joints in the global space.

5.2 Comparison with SOTAs

Quantitative Results. Table 1 shows the evaluation results on the ToF-IP-DB and DIP [8] dataset. The results demonstrate that our method consistently outperforms existing approaches across all metrics, particularly in SIP angular error and positional error. The reduction in SIP error signifies more accurate tracking of upper arm movements and knee lifts, which is attributed to the ToF-enabled improvement in joint position estimation as we expected.

Qualitative Results. As illustrated in Fig. 4, we selected Tai Chi and Baduanjin movements from the ToF-IP-DB dataset, which are characterized by long durations and gentle velocities—conditions where inertial-only position tracking inherently fails to measure accurate joint position. In contrast, our method leverages ToF-derived direct distance measurements and introduce inter-joint distance constraints, leading to marked improvements in overall pose estimation accuracy.

5.3 Ablation Study

Effectiveness of ToF Integration. As shown in Table. 2, removing ToF leads to a noticeable decrease in all metrics, specifically in terms of EndPos Error, this validates our core motivation that the direct distance measurements from ToF sensors can improve the estimation accuracy of joint endpoints, thereby enhancing the estimation of non-sensor-attached measured joints (lower SIP). Notably, introducing ToF without using the proposed Node-centric Data Integration (NCI)

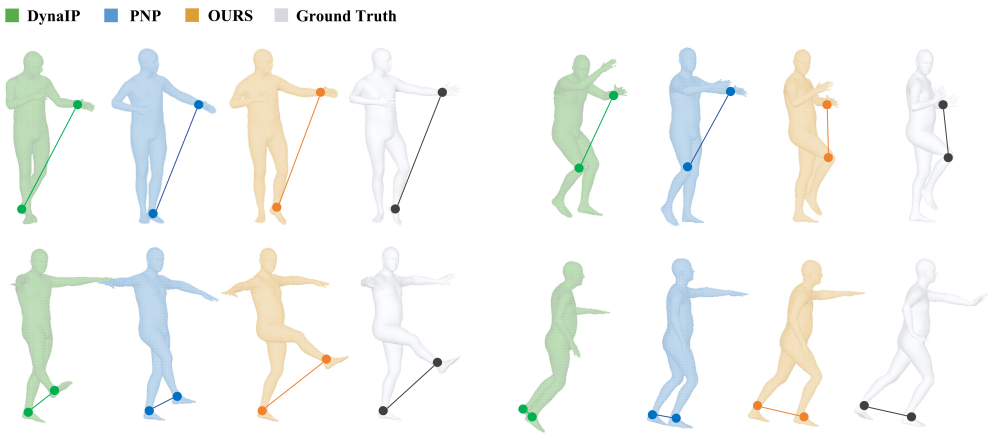

Figure 4: Qualitative comparisons with the state-of-the-art methods on our ToF-IP-DB dataset. We highlight the joint-to-joint distances in the ToF's line-of-sight direction.

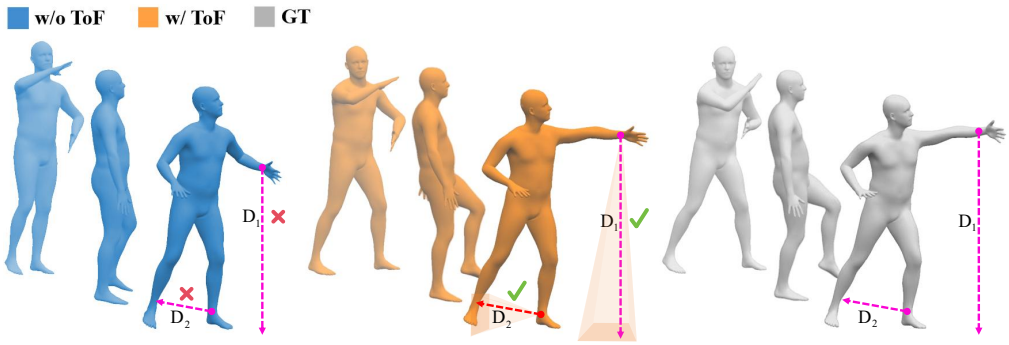

Figure 5: Qualitative comparison of results with and without ToF direct distance measurements. The pose estimation samples are from Case 1 and Case 4 in Table 2.

fails to achieve the desired improvement (Case 1 vs. Case 2), demonstrating the necessity of the proposed NCI for our ToF-Inertial motion capture framework. Qualitative results in Fig. 5 demonstrate how ToF integration significantly enhances inertial-only motion capture. The direct distance measurements provided by ToF effectively reduce the estimation errors of joint positions, thereby ensuring continuous and accurate pose estimation.

Table 2: Ablation study results on DIP and our ToF-IP-DB datasets (ToF integration).

Case	ToF	NCI	ToF-IP-DB					DIP (with synthesis ToF)				
			SIP Err	Ang Err	Pos Err	EndPos Err	Jitter	SIP Err	Ang Err	Pos Err	EndPos Err	Jitter
1	×	×	18.87	13.14	6.88	12.27	0.07	16.34	7.64	5.80	8.49	0.13
2	✓	×	18.95	12.89	6.92	12.39	0.13	16.10	7.50	5.47	8.08	0.17
3	×	✓	17.90	13.18	6.81	12.06	0.13	15.52	7.30	5.30	7.79	0.17
4	✓	✓	**17.26**	**12.09**	**6.31**	**11.41**	0.12	**13.62**	**6.75**	**4.59**	**6.65**	0.17

Effectiveness of Dynamic Spatial Positional Encoding. As shown in Table 3, our proposed Dynamic Spatial Positional Encoding consistently outperforms both the traditional static encoding and its learnable variant (where parameters are optimized during training but remain fixed during inference) [2], demonstrating the effectiveness of modeling position in positional encoding for ToF-Inertial motion capture, Fig.6 further supporting the dynamic nature of our encoding scheme.

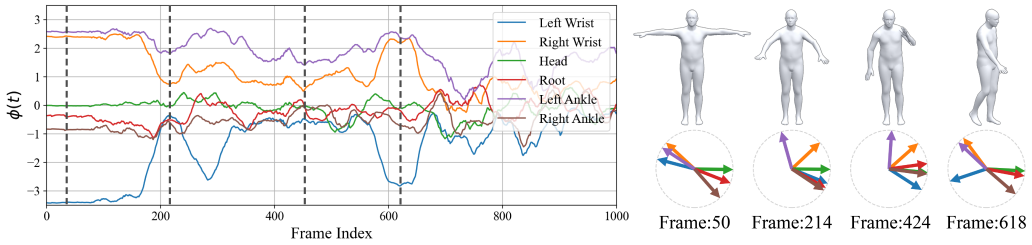

Figure 6: Quantitative visualization of the temporal dynamics of $\phi_n(t)$ across different motion types. Each curve represents the phase-based positional encoding of a specific sensing node over time.

Table 3: Ablation study results on DIP and our ToF-IP-DB datasets (positional encoding).

Positional Encoding	ToF-IP-DB					DIP				
	SIP Err	Ang Err	Pos Err	EndPos Err	Jitter	SIP Err	Ang Err	Pos Err	EndPos Err	Jitter
Static	17.76	12.52	7.60	11.98	0.13	14.20	6.87	4.77	6.95	0.17
Static (Learnable)	17.56	12.20	7.36	11.56	0.14	13.95	6.85	4.62	6.73	0.18
Dynamic Spatial (Ours)	**17.26**	**12.09**	**6.31**	**11.41**	**0.12**	**13.62**	**6.75**	**4.59**	**6.65**	**0.17**

6 Limitations

Despite the superiority of our approach, several limitations highlight avenues for future research. The performance of our method is contingent on the availability and reliability of ToF-based distance measurements, which may be compromised in scenarios with occlusions, or limited field of view. Additionally, the inherent noise in ToF sensors introduces jitter in pose estimation, as reflected in our results. Furthermore, the current evaluation is conducted in controlled environments, and the generalization of our approach to more diverse and dynamic real-world scenarios remains to be validated. Future work could focus on improving robustness to ToF sensor limitations through hybrid models, reducing noise with advanced filtering techniques.

7 Conclusion

We introduced ToF-IP, a novel ToF-inertial motion capture system that overcomes the inherent limitations of sparse IMUs by integrating direct distance measurements from lightweight, body-mounted ToF sensors. Through a hardware-efficient design and a unified Transformer-based framework, ToF-IP achieves accurate joint position estimation while preserving the portability and wearability of existing IMU systems. Our software contributions, including Node-Centric Data Integration and Dynamic Spatial Positional Encoding, enable structured multi-sensor integration and dynamic spatial awareness, which are critical for handling complex, low-velocity, and non-linear human motions. Extensive experiments validate ToF-IP's effectiveness across diverse movement scenarios, setting a new standard for hybrid sensor-based human motion tracking.

8 Acknowledgments

The authors would like to thank Qiannan Cao, Yingqi Yang, Hui Zou, Lishuang Zhan, Jiabao Gan, Yile Pan, Ziyi Shan, Ran Li, Bingling Liu, Wenjing Wu, Jiaqi Li, Yuenan Ji, Ziqian Huang and Shuyang Xing for their help on live demos and dataset collection. This work was supported by the National Natural Science Foundation of China (Nos. 62472364, 62072383), the Public Technology Service Platform Project of Xiamen City (No. 3502Z20231043), the Xiaomi Young Talents Program / Xiaomi Foundation, and the Fundamental Research Funds for the Central Universities (No. 20720240058, "Young Eagle Plan" Top Talents of Fujian Province). This work was also supported by the National Key R&D Program of China (No. 2023YFC3305600).

References

[1] Yiming Bao, Xu Zhao, and Dahong Qian. Fusepose: Imu-vision sensor fusion in kinematic space for parametric human pose estimation. *IEEE Transactions on Multimedia*, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[3] Jiawei Fang, Ruonan Zheng, Xiaoxia Gao, Chengxu Zuo, Shihui Guo, Yiyue Luo, et al. Fip: Endowing robust motion capture on daily garment by fusing flex and inertial sensors. *arXiv preprint arXiv:2502.15058*, 2025.

[4] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 755–762. IEEE, 2010.

[5] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *International Journal of Computer Vision*, 127:381–397, 2019.

[6] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.

[7] Hashim A Hashim, Abdelrahman EE Eltoukhy, and Kyriakos G Vamvoudakis. Uwb ranging and imu data fusion: Overview and nonlinear stochastic filter for inertial navigation. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):359–369, 2023.

[8] 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.

[9] Yunhan Huang, Arvind Sai Sarathi Vasan, Ravi Doraiswami, Michael Osterman, and Michael Pecht. Mems reliability review. *IEEE Transactions on Device and Materials Reliability*, 12(2):482–493, 2012.

[10] Yonatan Hutabarat, Dai Owaki, and Mitsuhiro Hayashibe. Quantitative gait assessment with feature-rich diversity using two imu sensors. *IEEE Transactions on Medical Robotics and Bionics*, 2(4):639–648, 2020. doi: 10.1109/TMRB.2020.3021132.

[11] 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*, pages 1–9, 2022.

[12] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29, pages 141–159. Wiley Online Library, 2010.

[13] Jung Keun Lee, Edward J Park, and Stephen N Robinovitch. Estimation of attitude and external acceleration using inertial sensor measurement during various dynamic conditions. *IEEE transactions on instrumentation and measurement*, 61(8):2262–2273, 2012.

[14] Jie Li, Xiaofeng Liu, Zhelong Wang, Xu Zhou, and Ziyang Wang. Sensor combination selection for human gait phase segmentation based on lower limb motion capture with body sensor network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. doi: 10.1109/TIM.2022.3201947.

[15] Gabriele Ligorio and Angelo M Sabatini. A novel kalman filter for human motion tracking with an inertial-based dynamic inclinometer. *IEEE Transactions on Biomedical Engineering*, 62(8):2033–2043, 2015.

[16] Shi Qiang Liu, Jun Chang Zhang, and Rong Zhu. A wearable human motion tracking device using micro flow sensor incorporating a micro accelerometer. *IEEE Transactions on Biomedical Engineering*, 67(4): 940–948, 2020. doi: 10.1109/TBME.2019.2924689.

[17] Shiqiang Liu, Junchang Zhang, Yuzhong Zhang, and Rong Zhu. A wearable motion capture device able to detect dynamic motion of human limbs. *Nature communications*, 11(1):5615, 2020.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), oct 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013. URL `https://doi.org/10.1145/2816795.2818013`.

[19] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[20] 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*, pages 1–12, 2023.

[21] Md Moniruzzaman, Zhaozheng Yin, Md Sanzid Bin Hossain, Hwan Choi, and Zhishan Guo. Wearable motion capture: Reconstructing and predicting 3d human poses from wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[22] Cecilia Monoli, Juan Francisco Fuentez-Pérez, Nicola Cau, Paolo Capodaglio, Manuela Galli, and Jeffrey A. Tuhtan. Land and underwater gait analysis using wearable imu. *IEEE Sensors Journal*, 21(9):11192–11202, 2021. doi: 10.1109/JSEN.2021.3061623.

[23] Noitom. Perception neuron series. `https://www.noitom.com/.`, n.d. Accessed [2025-02-07].

[24] Simone Pasinetti, M Muneeb Hassan, Jörg Eberhardt, Matteo Lancini, Franco Docchio, and Giovanna Sansoni. Performance analysis of the pmd camboard picoflexx time-of-flight camera for markerless motion capture applications. *IEEE transactions on instrumentation and measurement*, 68(11):4456–4471, 2019.

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[26] François Piron, Daniel Morrison, Mehmet Rasit Yuce, and Jean-Michel Redouté. A review of single-photon avalanche diode time-of-flight imaging sensor arrays. *IEEE Sensors Journal*, 21(11):12654–12666, 2020.

[27] Alice Ruget, Max Tyler, Germán Mora Martín, Stirling Scholes, Feng Zhu, Istvan Gyongy, Brent Hearn, Steve McLaughlin, Abderrahim Halimi, and Jonathan Leach. Real-time, low-cost multi-person 3d pose estimation. *arXiv preprint arXiv:2110.11414*, 2021.

[28] Laura Susana Vargas-Valencia, Felipe B. A. Schneider, Arnaldo G. Leal-Junior, Pablo Caicedo-Rodríguez, Wilson A. Sierra-Arévalo, Luis E. Rodríguez-Cheu, Teodiano Bastos-Filho, and Anselmo Frizera-Neto. Sleeve for knee angle monitoring: An imu-pof sensor fusion system. *IEEE Journal of Biomedical and Health Informatics*, 25(2):465–474, 2021. doi: 10.1109/JBHI.2020.2988360.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] 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.

[31] 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*, volume 36, pages 349–360. Wiley Online Library, 2017.

[32] 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)*, September 2018.

[33] Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, and Felix Heide. Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13252–13262, 2023.

[34] Yinghao Wu, chaoran wang, Lu Yin, Shihui Guo, and Yipeng Qin. Accurate and steady inertial pose estimation through sequence structure learning and modulation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 42468–42493. Curran Associates, Inc., 2024.

[35] Xsens. Xsens 3d motion tracking. `https://www.xsens.com/`, n.d. Accessed [2025-03-01].

[36] 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.

[37] 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.

[38] 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 Trans. Graph.*, 42(4), 2023.

[39] 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.

[40] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024.

[41] Shaghayegh Zihajehzadeh and Edward J. Park. A novel biomechanical model-aided imu/uwb fusion for magnetometer-free lower body motion capture. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(6):927–938, 2017. doi: 10.1109/TSMC.2016.2521823.

[42] Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, and Yipeng Qin. Loose inertial poser: Motion capture with imu-attached loose-wear jacket. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2209–2219, 2024.

[43] Chengxu Zuo, Jiawei Huang, Xiao Jiang, Yuan Yao, Xiangren Shi, Rui Cao, Xinyu Yi, Feng Xu, Shihui Guo, and Yipeng Qin. Transformer imu calibrator: Dynamic on-body imu calibration for inertial motion capture. *ACM Transactions on Graphics*, 44(4), 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discussed them in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: While we provide the core assumptions and sketch key derivation steps for our theoretical results, we do not include full formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we will release the code/data later.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we specified all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, we did not do that.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, For each experiment, we provided sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work does not have direct potential positive societal impacts and negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not have such issues.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all the creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and are the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not have such issues in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes, we included relevant details.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This research involving human participants was reviewed and approved by the Institutional Review Board (IRB). All participants provided informed consent after being fully informed of the potential risks and procedures.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This work does not involve the use of large language models (LLMs) as an important, original, or non-standard component of the core methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.