MARIO: Motion-Augmented Real-Time Multi-Sensor Inertial Odometry

Yiquan Li[1,*] Taeyoung Yeon[1,*] Chenfeng Gao[1] Vasco Xu[2] Xuanyou Liu[1] Karan Ahuja[1]
[1]Northwestern University [2]University of Chicago
[*]Equal contribution

Abstract

Inertial odometry (IO) using only Inertial Measurement Units (IMUs) provides a lightweight solution for human motion tracking in augmented reality (AR) and wearable devices. Recent learning-based IO methods have improved the generalizability of inertial localization via large-scale pretraining on human motion datasets. Unfortunately, these approaches remain prone to drift and noise because they fail to capture human motion dynamics, especially on daily activity datasets such as Nymeria. In contrast, we propose to ground inertial odometry in human kinematics through a learned IMU-inferred pose prior that promotes the propagation of physically consistent motion constraints. We integrate our pose prior into existing IO architectures and reduce positional drift by up to 36% in the challenging Nymeria dataset (5x larger than prior works). We further showcase improved long-term performance by developing a sensor-fusion framework that incorporates auxiliary signals from other lightweight sensors such as the magnetometer, barometer, and secondary IMU already available on commercial AR glasses. With our fusion strategy, drift is reduced to 42%, improving robustness and generalization across diverse motion conditions. Together, our results establish a new paradigm for inertial and lightweight odometry, unifying human motion kinematics with multimodal sensing, setting a new benchmark for accurate and robust camera-less human tracking. Our website is available at https://spice-lab.org/projects/MARIO/.

1. Introduction

Inertial Measurement Units (IMUs) are compact and low-cost sensors that measure linear acceleration and angular velocity. They are widely used in computer vision, robotics, and extended reality (XR) for motion tracking, operating reliably under challenging conditions with a low power profile and lightweight form-factor. Their small form factor and low power consumption make them a core sensing modality for wearable and mobile devices.

However, estimating position from IMU signals involves integrating noisy acceleration and angular velocity, which leads to cumulative drift. Visual–inertial odometry (VIO) methods [3, 23] mitigate this drift by fusing IMU signals with image features from cameras. While effective, visual methods degrade in high-speed motion or visually challenging conditions such as low light, motion blur, or lack of visual features. Moreover, cameras introduce power, privacy, and environmental limitations, making IMU-only odometry increasingly attractive for lightweight AR and wearable devices.

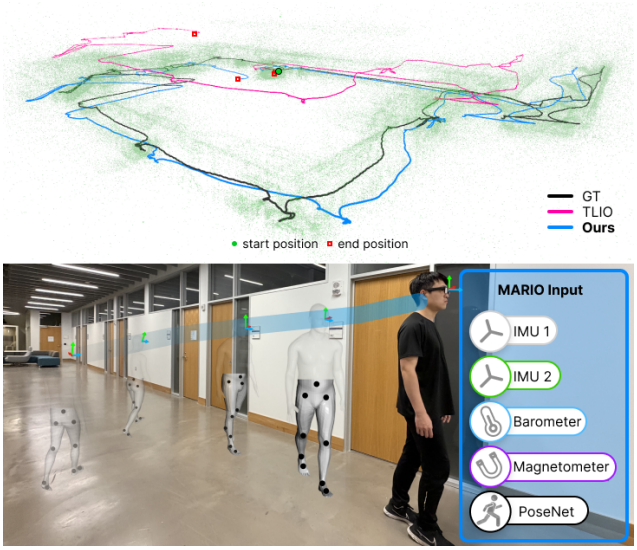

Figure 1. We propose MARIO, an inertial odometry framework that builds on a single-IMU model, grounds motion in lower-body pose predicted by a pretrained PoseNet, and enhances robustness through sensor fusion with a secondary IMU, barometer, and magnetometer.

Recently, learning-based inertial odometry (IO) methods have emerged that mitigate drift through data-driven motion priors learned directly from IMU signals [7, 19, 24, 32]. Despite these advances, most models rely on a single IMU and remain prone to drift and noise, limiting their robustness in tracking scenarios that involve complex human motions and diverse environments.

Our key insight is that on-body IMU tracking captures human kinematics, rather than implicitly integrating accelerometer and gyroscope measurements over time. In this work, we explicitly ground inertial odometry in human kinematics through a learned IMU-inferred pose prior, thereby strengthening motion constraints. We propose PoseNet, which predicts full-body pose in the SMPL body model [17] from a single head-mounted IMU. The learned pose prior serves as a spatio-temporal kinematic anchor, injecting physically meaningful structure into motion estimation and substantially reducing translation error and drift when integrated into existing inertial odometry architectures.

Furthermore, on-body devices such as AR glasses provide readily available, low-power sensors, including a magnetometer, barometer, and dual IMUs. However, the benefits of incorporating these sensors into neural inertial odometry models remain largely unexplored. To further enhance performance, we introduce a Multi-Sensor Fusion Module that incorporates these auxiliary signals, which are already available on commercial AR glasses such as Meta's Aria [4]. These complementary cues provide more accurate heading and elevation information, improving robustness and generalization across diverse motion patterns.

We demonstrate the generality of our framework by integrating it with four state-of-the-art inertial odometry methods and conducting comprehensive evaluations on the Nymeria [19] (300 hrs, 297.87m avg. trajectory length), Aria Everyday Activities (7.3 hrs, 29.35m avg. trajectory length) [18], and TLIO [16] (60 hrs, 119.78m avg. trajectory length) datasets. Quantitative and qualitative analyses show that our approach consistently reduces drift and lowers translation error across all benchmarks, confirming its robustness and generalization for human-centric motion tracking.

Our key contributions are:

- We introduce a human-pose-grounded inertial odometry framework that injects a learnt IMU-inferred kinematics prior into our odometry model, improving stability and reducing drift.
- We demonstrate a sensor-fusion framework that integrates barometer, magnetometer, and dual-IMU signals to provide robust odometry, substantially improving vertical and heading cues.
- We integrate our modules into existing IO architectures (TLIO [16], AirIO [24], RoNIN [32] and EQNIO [7]) and benchmark on Nymeria, Aria Everyday, and TLIO datasets demonstrating consistent improvements across datasets and metrics.

2. Related Work

2.1. Inertial Odometry

Inertial odometry (IO) estimates a device's 6-DoF pose (position and orientation) from IMU signals. Orientation is obtained by integrating gyroscope readings, while double integration of linear acceleration yields position. However, noise and bias in low-cost IMUs accumulate during integration, causing drift over time. Heuristic methods such as pedestrian dead reckoning (PDR) exploit motion regularities (e.g., step detection and stride estimation) but rely on hand-tuned assumptions and degrade in unconstrained motion [9, 10].

Learning-based IO replaces such heuristics with data-driven motion priors that infer displacement or velocity directly from IMU sequences. RIDI [31] regresses short-term velocity to correct low-frequency drift, while IONet [2] and RoNIN [32] integrate learned velocity estimates for accurate 2D trajectory reconstruction. RoNIN further introduces a heading-agnostic coordinate frame (HACF) aligning gravity with the vertical axis. TLIO [16] extends this by combining a learned displacement estimation with an Extended Kalman Filter (EKF) to jointly refine position, orientation, and bias. IDOL [26] explicitly estimates orientation from magnetometer readings and regresses translation with a bi-directional LSTM, improving robustness and reducing long-term drift.

To improve generalization, recent works incorporate structural and equivariant priors. AirIO [24] shows that retaining IMU data in the body frame, rather than transforming it to the global frame, improves drone odometry. RIO [1] leverages rotational equivariance as a self-supervised signal, and EqNIO [7] learns canonical displacement priors in a gravity-aligned frame for rotation consistency. TartanIMU [39] extends IO to a foundation model trained on 100+ hours of IMU data across cars, drones, robots, and humans, achieving cross-domain generalization through low-rank fine-tuning and online adaptation. M2EIT [15] further introduces a mixture-of-experts framework that fuses spatial, temporal, and frequency features for state-of-the-art robustness across motion domains.

Orthogonal to prior work, we introduce a human-pose-guided prior within a multi-stage framework, grounding motion estimation in body dynamics to improve both accuracy and robustness. We further leverage multi-sensor fusion on AR glasses—combining multiple IMUs and auxiliary sensors (barometer, magnetometer) already available on existing devices [4]—to further mitigate long-term instability and bias drift.

2.2. Body Motion Capture for IMU Devices

Inertial measurement units (IMUs) have become an attractive sensing modality for motion capture due to their com-

pact size, low power consumption, and portability. Commercial systems such as Xsens [29] employ many high-grade IMUs (typically 17) for high-fidelity full-body tracking but remain intrusive and impractical for everyday use.

To improve practicality, recent research has explored reconstructing full-body pose from fewer sensors. SIP [28] demonstrated that six Xsens IMUs can reconstruct full-body motion using offline optimization. Deep Inertial Poser (DIP) [6] introduced a bi-directional LSTM model for pose estimation from six IMUs, while TransPose [34] extended this to estimate global translation via a multi-stage architecture. TIP [8] applied Transformers for non-planar motion, and PIP [35], PNP [36], and GlobalPose [37] incorporated physics-based optimization for more plausible motion reconstruction.

Building on these advances, recent work leverages IMUs embedded in everyday devices. IMUPoser [22] demonstrated that arbitrary combinations of IMUs in earbuds, smartwatches, and smartphones can reconstruct full-body pose using masking strategies for missing sensors, while MobilePoser [30] extended this approach with a multi-stage framework that jointly estimates pose and global translation. DiffusionPoser [27] further employed a diffusion-based model to inpaint missing signals during denoising, supporting more flexible sensor placements.

These works pave the way for reconstructing full-body motion from sparse IMU configurations, but they often rely on multiple on-body IMUs. We propose learning a general human-pose prior from a single IMU which, when integrated into inertial odometry models, significantly reduces error and drift.

2.3. Sensor Fusion for IMU Motion Tracking

IMUs are compact and low-cost but are sensitive to noise, leading to drift over time [41]. To mitigate these limitations, extensive research has explored complementing IMUs with other sensors to improve stability and accuracy. The most common is visual-inertial odometry (VIO), which jointly optimizes motion estimates from IMU and camera observations. LiDAR–inertial odometry (LIO) follows a similar principle, using geometric depth constraints from LiDAR scans to improve robustness in low-light or textureless environments. We refer readers to comprehensive surveys for detailed reviews of VIO and LIO methods [5, 12]. Also, MINS [13] proposes a tightly coupled multisensor-aided inertial navigation system that fuses IMU, wheel encoders, cameras, LiDAR, and GNSS to address asynchronous measurements.

Cameras and LiDARs are power-intensive and often impractical for lightweight platforms such as AR glasses. Recent work instead incorporates non-visual modalities—magnetometers, barometers, ultrasound, and ToF sensors—for efficient motion tracking. MagShield [25] uses magnetic cues to detect disturbances and correct orientation errors. BaroPoser [38] exploits barometric pressure on wearables to recover global translation on non-flat terrain. UltraPoser [14] leverages ultrasound via built-in speakers and microphones to expand body-area sensing, while ToF-IP [33] integrates Time-of-Flight depth with sparse IMUs to constrain geometry and reduce drift. There is another line of work that fuses additional sensors into an EKF [11, 20, 21]. However, little work has explored the fusion of easily accessible sensors, such as barometers and magnetometers, for neural inertial odometry.

3. Method

We formulate the inertial odometry (IO) problem as estimating a device trajectory, T, from a time series of IMU measurements comprising linear acceleration and angular velocity. Existing approaches differ in how they reconstruct motion from these signals.

We aim to show that by explicitly introducing a human motion prior—learned from IMU signals—and integrating complementary sensing modalities, inertial tracking can be made more stable, accurate, and generalizable. As illustrated in Figure 2, our framework consists of two main components: (1) PoseNet, which learns a human motion prior from a single IMU to provide kinematic structure and temporal consistency; and (2) a Multi-Sensor Fusion Module, which combines magnetometer, barometer, and secondary-IMU signals through a learned mid-level fusion strategy to enhance robustness and long-term stability.

To validate our approach, we integrate our modules into four representative IO architectures. AirIO [24] predicts body-frame velocity and transforms it to global coordinates using gyroscope-derived rotations before integration. TLIO [16] instead directly regresses 3D positional displacements over fixed 1-second windows using a ResNet architecture with gravity-aligned input. EqNIO[7] extends this displacement-based approach by learning in a canonical O(2) equivariant frame for improved rotation consistency. Finally, RoNIN-LSTM [32] uses bi-directional LSTM for 2D velocity, we modified it to output 3D velocity vectors for fair comparison.

3.1. PoseNet

We introduce PoseNet, a lightweight network that learns a pose prior from a single IMU stream. The network takes linear accelerations, angular velocities, and orientations as input and predicts body pose in SMPL format using a 6D rotation representation [40].

Following AirIO[24] archiecture to extract IMU feature information, PoseNet uses a hybrid convolutional–recurrent architecture designed to capture both short-term inertial dynamics and long-range temporal dependencies. IMU signals are first encoded by a temporal CNN that extracts lo-

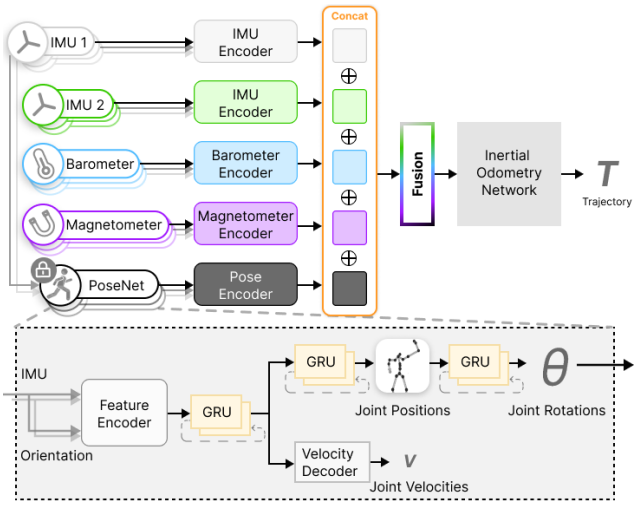

Figure 2. Overview of the MARIO inertial odometry framework. (a) We first learn a PoseNet that takes IMU measurements and orientations as input and predicts human pose (represented by joint rotations). (b) We encode multiple sensor signals—secondary IMU, barometer, magnetometer, and the pose estimated by the frozen PoseNet—concatenate their features through a fusion layer, and then feed the fused representation to an existing inertial odometry model to predict the trajectory.

cal motion patterns, followed by stacked GRUs that sequentially predict joint positions and then rotations in the SMPL format. We estimate only nine joint angles corresponding to the pelvis and lower limbs, focusing on locomotion-related joints that dominate global motion while reducing model complexity. Estimating arm or hand motion from head-mounted IMUs is inherently underconstrained and thus excluded. Therefore, PoseNet will give a 9×6D dimension pose vector each timestamp, which will be fed into our fusion module described in the next subsection. We train the pose prior using an ℓ_2 loss on joint positions and 6D joint rotations, and a Huber loss on joint-angle velocities to encourage temporal smoothness. Let $\mathbf{J}$ and $\hat{\mathbf{J}}$ denote the ground-truth and predicted joint positions, and $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ the corresponding 6D rotations:

$$\mathcal{L}_{\text{pos}} = \left\|\hat{\mathbf{J}} - \mathbf{J}\right\|_2^2, \qquad \mathcal{L}_{\text{ang}} = \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2.$$

Given ground-truth velocity v and predicted velocity $\hat{v}$, we compute the velocity loss defined by a Huber loss function:

$$\mathcal{L}_{\text{vel}} = \begin{cases} \frac{1}{2}(\hat{v} - v)^2, & \text{if } |\hat{v} - v| < \delta, \\ \delta\left(|\hat{v} - v| - \frac{1}{2}\delta\right), & \text{otherwise,} \end{cases}$$

with $\delta = 0.005$. The total training objective is

$$\mathcal{L} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{ang}}\mathcal{L}_{\text{ang}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}},$$

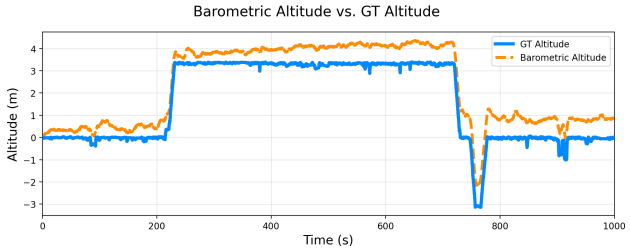

Figure 3. Visualization of altitude from the barometer compared with the ground-truth altitude. This demonstrates that the barometer provides valuable altitude information.

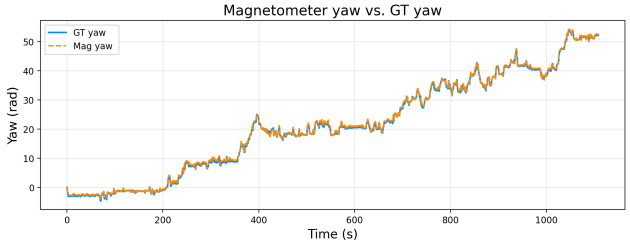

Figure 4. Visualization of magnetometer data compared with ground-truth yaw orientation. The magnetometer signal tracks the ground-truth, demonstrating its utility for constraining heading drift in inertial odometry.

where $\lambda_{\text{pos}} = 0.05$, $\lambda_{\text{ang}} = 0.05$, and $\lambda_{\text{vel}} = 1$.

PoseNet is trained on the Nymeria dataset [19], which contains over 300 hours of synchronized IMU and ground-truth pose data collected from head-mounted devices across diverse activities and environments.

3.2. Multi-Sensor Fusion Module

Predicting motion from IMU signals is inherently prone to drift due to sensor noise and bias. To address this limitation, we introduce a Multi-Sensor Fusion Module that combines complementary signals from a barometer, a magnetometer, and a secondary IMU that are already available on AR glasses [4]. These additional modalities supply grounding for elevation and heading and provide data redundancy that improves robustness and long-term stability.

Barometer A barometer measures ambient air pressure as a scalar time series in pascals. We convert pressure p (Pa) to altitude h (m) using the hypsometric relation under the International Standard Atmosphere (ISA) model, assuming sea-level pressure $P_0 = 101{,}325\,\text{Pa}$, temperature T_0, and lapse rate L:

$$h \approx K\left(1 - \left(\frac{p}{P_0}\right)^n\right), \qquad K = \frac{T_0}{L}, \quad n = \frac{R\,L}{g_0},$$

where R is the specific gas constant for dry air and g_0 is standard gravity. The altitude signal is generally smooth

but may drift slowly due to weather or indoor pressure changes. To reduce noise, we apply a moving-average filter to denoise both the altitude and its derivative vertical velocity, yielding a 1D velocity estimate at each timestamp that is later used in the fusion module. Figure 3 shows that barometric altitude closely follows ground-truth, indicating that pressure provides a strong cue for estimating altitude changes.

Magnetometer A magnetometer measures the local magnetic field as a 3D vector, which can be used to infer heading relative to magnetic north. Despite susceptibility to magnetic disturbances indoors, it provides a valuable absolute orientation cue when calibrated.

We compute the yaw angle by combining the magnetometer reading $\mathbf{m}$ with gravity $\mathbf{g}$ estimated from the IMU. After normalization, $\hat{\mathbf{m}} = \mathbf{m}/\|\mathbf{m}\|$ and $\hat{\mathbf{g}} = \mathbf{g}/\|\mathbf{g}\|$, tilt compensation removes the vertical component:

$$\mathbf{E} = \frac{\hat{\mathbf{m}} \times \hat{\mathbf{g}}}{\|\hat{\mathbf{m}} \times \hat{\mathbf{g}}\|}, \qquad \mathbf{N} = \hat{\mathbf{g}} \times \mathbf{E},$$

and the yaw is computed as

$$\psi = \operatorname{atan2}\big((\mathbf{E})_z,\ (\mathbf{N})_z\big).$$

This 1D yaw signal serves as the magnetometer input to the fusion module. Figure 4 compares the derived yaw with ground-truth, showing alignment and demonstrating that the magnetometer provides a useful heading cue.

Secondary IMU An accelerometer on a moving rigid body measures not only linear acceleration but also rotation-induced terms. For a body with linear acceleration $\mathbf{a}_0$, angular velocity $\boldsymbol{\omega}$, and angular acceleration $\boldsymbol{\alpha}$, an accelerometer at position $\mathbf{r}$ records

$$\mathbf{a} = \mathbf{a}_0 + \boldsymbol{\alpha} \times \mathbf{r} + \boldsymbol{\omega} \times \big(\boldsymbol{\omega} \times \mathbf{r}\big) + \mathbf{b} + \mathbf{n},$$

where $\mathbf{b}$ and $\mathbf{n}$ denote bias and noise. Adding a second, spatially separated IMU improves the observability of rotational motion and helps disentangle translational acceleration from rotation-induced effects. The dual-IMU configuration also provides redundancy that reduces the impact of sensor bias and noise. In the Aria glasses used in the Nymeria [19] and Aria Everyday Activities datasets [18], the primary IMU is located on the right temple and the secondary IMU on the left. Similar to the primary IMU, the secondary IMU provides a 6D (accelerometer + gyroscope) measurement vector at each timestamp for sensor fusion.

3.3. Feature Fusion Strategy

We use mid-level feature fusion tailored to the latent state of inertial odometry (IO). Each auxiliary stream (magnetometer, barometer, secondary IMU) is first time-aligned to the primary IMU. We then encode each stream with a small causal temporal CNN to obtain per-step features $f_{\text{mag}}(t)$, $f_{\text{baro}}(t)$, $f_{\text{imu2}}(t)$ and also for $f_{\text{pose}}(t)$. The primary IMU (accelerometer + gyroscope) is passed through the same architecture to yield $f_{\text{imu}}(t)$. At time t, we concatenate:

$$f_{\text{fusion}}(t) = \big[\, f_{\text{imu}}(t)\,;\, f_{\text{mag}}(t)\,;\, f_{\text{baro}}(t)\,;\, f_{\text{imu2}}(t)\,;\, f_{\text{pose}}(t)\,\big]. \tag{1}$$

and feed $f_{\text{fusion}}(t)$ to a MLP fusion layer followed by existing inertial odometry models to estimate the motion state.

4. Experiments

4.1. Dataset

We evaluate our approach across three different datasets, Nymeria [19] , Aria Everyday Activities [18] and TLIO [16]. The Nymeria dataset serves as our primary training source, partitioned with an 80/10/10 split for train/validation/test. For additional evaluation on household activities, we use the Aria Everyday Activities (AEA) as a test dataset with a model pretrained on Nymeria. For baseline comparisons, we incorporate the TLIO dataset following the original repository's train/validation/test split [16].

Nymeria Dataset The Nymeria dataset [19] comprises 1200 sequences of 15-20 minute duration from 264 subjects in 50 varied indoor and outdoor environments, totaling 300 hours and 400km of movement. The data captures 20 different scenarios including athletic activities, meals, work tasks, outdoor excursions, and cycling. Project Aria glasses record synchronized multimodal signals: dual IMUs at 800Hz and 1kHz, magnetometer at 10Hz, and barometer at 50Hz. Ground-truth full-body pose is captured using an Xsens MVN Link suit with 17 inertial units at 240 Hz. We convert these Xsens poses into the SMPL format to supervise PoseNet. The Project Aria MPS system provides the 3D trajectories at 1 kHz used as ground-truth for inertial odometry.

Aria Everyday Activities Dataset The AEA dataset [18] contains 143 recording sessions across five distinct indoor environments, totaling 7.3 hours. Activities include typical household routines: meal preparation, tidying, eating, and social interaction. Recording setup and ground-truth systems match specifications from Nymeria dataset.

TLIO Dataset The TLIO dataset [16] features approximately 60 hours of walking motion recorded via head-worn hardware with a Bosch BMI055 inertial sensor. The collection focuses on pedestrian movement patterns within interior spaces. Unlike Nymeria and AEA, TLIO provides only IMU measurements without additional sensing modalities,

with ground-truth trajectories derived from Visual Inertial Odometry.

4.2. Implementation Details

We conduct all experiments on NVIDIA GeForce RTX 4090 and A40 GPUs. AirIO uses a sliding window of 1,000 samples (stride 100), while EqNIO, TLIO, and RoNIN-LSTM use 200 samples with stride 10 for TLIO/Aria Everyday datasets and stride 100 for Nymeria. All signals are resampled to 200 Hz; barometer and magnetometer readings are interpolated to match the IMU rate. Unless otherwise specified, models are trained on Nymeria's training split, evaluated on its test split, then assessed on Aria Everyday for cross-dataset generalization. The pose prior is trained exclusively on Nymeria and frozen during training and evaluation.

We use a learning rate of 5×10^{-4} for AirIO and 1×10^{-4} for TLIO. During training, we condition the models on ground-truth orientations, whereas at test time we use orientations inferred from the gyroscope. For TLIO, EqNIO, and RoNIN-LSTM, we train the covariance branch from the first iteration, which we find leads to more stable convergence on Nymeria. For gravity compensation in the raw IMU accelerometers, we subtract gravity using the known device orientation.

4.3. Evaluation Metrics

Performance is reported using Absolute Trajectory Error (ATE), Relative Translational Error at 1 s and 5 s (RTE-1s / RTE-5s), and a segment-wise drift rate ("Drifting"). ATE is computed as the root-mean-square error between estimated and ground-truth positions after temporal synchronization and rigid alignment:

$$\mathrm{ATE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{p}_i - \hat{\mathbf{p}}_i\right\|^2}. \tag{2}$$

To better analyze ATE, we further decompose it into horizontal and vertical components, where the horizontal error is measured in 2D, capturing motion on the ground plane, and the vertical error is measured in 1D, capturing only upward and downward motion. RTE-Δt measures the mean error of relative displacements over sliding windows of duration $\Delta t \in \{1, 5\}$ s:

$$\mathrm{RTE}_{\Delta t} = \frac{1}{n}\sum_{i=1}^{n}\left\|\left(\mathbf{p}_{i+\Delta t} - \mathbf{p}_i\right) - \left(\hat{\mathbf{p}}_{i+\Delta t} - \hat{\mathbf{p}}_i\right)\right\|. \tag{3}$$

Drifting is reported as the mean relative error (percentage) with respect to the ground-truth displacement over the same windows.

4.4. Results

We present results on the Nymeria and Aria datasets, integrating our PoseNet and Multi-Sensor Fusion Module into four IO architectures: AirIO [24], TLIO [16], EqNIO [7], and RoNIN-LSTM [32]. We use (+Pose) to denote models with our pose prior and (+All) for models using both pose and all sensor modalities.

Table 1 and 2 present our main results on Nymeria and Aria Everyday datasets. On Nymeria, our full system (+All) achieves consistent improvements across all architectures: AirIO shows 32% ATE reduction (6.85m $\rightarrow$ 4.64m) and 39% drift reduction (3.56% $\rightarrow$ 2.16%), while TLIO demonstrates even larger gains with 44% ATE improvement (10.19m $\rightarrow$ 5.73m) and 44% drift reduction (6.46% $\rightarrow$ 3.64%). EqNIO and RoNIN-LSTM follow similar trends with 41% and 35% ATE reductions respectively. Adding PoseNet alone delivers substantial benefits, reducing RTE-5s by 11% for AirIO and 13% for TLIO. Individual sensor modalities contribute complementary cues, with the barometer particularly effective for vertical stability. On Aria (Table 2), models pretrained on Nymeria generalize effectively without fine-tuning, maintaining strong performance with 29-32% ATE improvements. The pose prior shows robust cross-domain transfer, reducing drift by up to 53% on Aria. Figure 5 shows this progressive improvement with three sample sequences across TLIO variants. Furthermore, our AirIO model with full sensor fusion and pose prior (+All) achieves 133.6M FLOPs per inference and 315 FPS on NVIDIA A40 GPU, enabling real-time deployment.

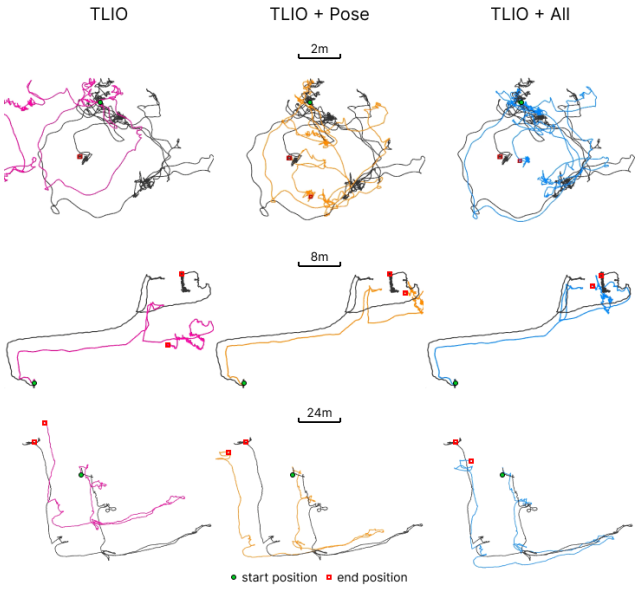

Figure 5. Trajectory visualizations of TLIO, TLIO+Pose, and TLIO+All on Nymeria dataset. TLIO+Pose improves over TLIO, and TLIO+All further reduces drift and tracking errors.

Table 1. Nymeria dataset results. Metrics: ATE (m; horizontal H, vertical V), RTE-5s (m), RTE-1s (m), Drift (%). Lower is better. The best result within each group (**Single IMU** and **Multi-Sensor**) is shown in bold.

		Single IMU		Multi-Sensor			
Model	Metric	Base	+Pose	+sec. IMU	+Baro	+Mag	+All
AirIO	ATE (H,V)	6.85 (6.43, 1.95)	**5.22 (4.89, 1.43)**	5.20 (4.81, 1.63)	5.25 (5.08, 1.00)	5.71 (5.30, 1.71)	**4.64 (4.30, 1.36)**
	RTE-5s	0.363	**0.323**	0.332	0.362	0.360	**0.297**
	RTE-1s	0.099	**0.085**	0.088	0.109	0.098	**0.078**
	Drift	3.56	**2.35**	2.52	2.56	2.77	**2.16**
TLIO	ATE (H,V)	10.19 (9.87, 1.09)	**7.97 (7.52, 1.46)**	6.45 (6.09, 0.93)	5.85 (5.54, 0.81)	8.57 (8.17, 1.45)	**5.73 (5.46, 0.77)**
	RTE-5s	0.322	**0.281**	0.300	0.260	0.304	**0.242**
	RTE-1s	0.103	**0.096**	0.097	0.088	0.099	**0.083**
	Drift	6.46	**4.94**	3.70	3.34	5.17	**3.64**
EqNIO	ATE (H,V)	**7.63 (7.21, 1.09)**	7.65 (7.26, 1.08)	5.49 (5.21, 0.70)	5.27 (5.02, 0.62)	6.85 (6.48, 0.86)	**4.52 (4.18, 0.96)**
	RTE-5s	0.379	**0.316**	0.273	0.270	0.331	**0.231**
	RTE-1s	0.122	**0.101**	0.091	0.093	0.107	**0.081**
	Drift	**3.93**	4.30	2.84	2.85	3.39	**2.40**
RoNIN-LSTM	ATE (H,V)	9.10 (8.77, 0.97)	**6.83 (6.51, 0.92)**	8.76 (8.34, 1.30)	**5.88 (5.59, 0.76)**	5.93 (5.60, 0.95)	5.89 (5.66, 0.64)
	RTE-5s	0.330	**0.281**	0.289	0.261	0.254	**0.246**
	RTE-1s	0.103	**0.092**	0.093	0.088	0.085	**0.084**
	Drift	6.22	**3.99**	6.13	**3.44**	3.64	3.61

Table 2. Aria dataset results using pretrained model from Nymeria dataset. Metrics: ATE (m; horizontal H, vertical V), RTE-5s (m), RTE-1s (m), Drift (%). Lower is better. The best result within each group (**Single IMU** and **Multi-Sensor**) is shown in bold.

		Single IMU		Multi-Sensor			
Model	Metric	Base	+Pose	+sec. IMU	+Baro	+Mag	+All
AirIO	ATE (H,V)	1.25 (1.18, 0.29)	**0.86 (0.79, 0.25)**	1.18 (0.92, 0.66)	1.03 (0.99, 0.22)	1.29 (1.21, 0.33)	**0.89 (0.82, 0.28)**
	RTE-5s	0.334	**0.218**	0.227	0.273	0.300	**0.201**
	RTE-1s	0.106	**0.062**	0.068	0.089	0.099	**0.062**
	Drift	11.38	**5.32**	7.35	8.55	9.48	**6.04**
TLIO	ATE (H,V)	1.51 (1.45, 0.21)	**1.23 (1.13, 0.26)**	1.05 (0.95, 0.25)	**1.00 (0.93, 0.18)**	1.12 (1.00, 0.29)	1.02 (0.96, 0.17)
	RTE-5s	0.246	**0.222**	0.231	0.202	0.231	**0.195**
	RTE-1s	0.081	**0.075**	0.077	0.070	0.078	**0.069**
	Drift	9.69	**8.09**	7.76	**6.68**	8.06	7.67
EqNIO	ATE (H,V)	**1.20 (1.11, 0.24)**	1.24 (1.17, 0.20)	0.92 (0.85, 0.18)	1.03 (0.98, 0.16)	1.08 (1.01, 0.20)	**0.89 (0.84, 0.13)**
	RTE-5s	0.270	**0.233**	0.203	0.212	0.246	**0.186**
	RTE-1s	0.093	**0.079**	0.072	0.075	0.084	**0.067**
	Drift	**8.18**	8.45	**6.43**	8.14	7.28	6.67
RoNIN-LSTM	ATE (H,V)	1.39 (1.31, 0.20)	**1.04 (0.97, 0.19)**	1.45 (1.31, 0.31)	0.99 (0.92, 0.17)	**0.95 (0.88, 0.18)**	1.05 (0.99, 0.14)
	RTE-5s	0.270	**0.212**	0.234	0.206	**0.196**	0.201
	RTE-1s	0.085	**0.073**	0.077	0.071	**0.069**	0.070
	Drift	13.22	**7.60**	12.51	7.74	**7.38**	7.49

In Figure 6, we demonstrate the cumulative distribution function (CDF) of RTE-5s for AirIO, TLIO, EqNIO, and RoNIN-LSTM on the Nymeria dataset. Adding the pose prior consistently lowers RTE-5s for all models, and incorporating all sensors together with the pose prior further reduces RTE-5s across the entire Nymeria test set.

5. Ablation Studies

Pose Prior Ablation In Table 3, we evaluate the pose prior's effectiveness on the TLIO dataset. We evaluate four baseline models—AirIO, TLIO, EqNIO, and RoNIN-LSTM—using PoseNet pretrained on the Nymeria dataset. The results show that incorporating PoseNet consistently improves ATE by 2–16% across all models, indicating that our pose prior generalizes well across different datasets.

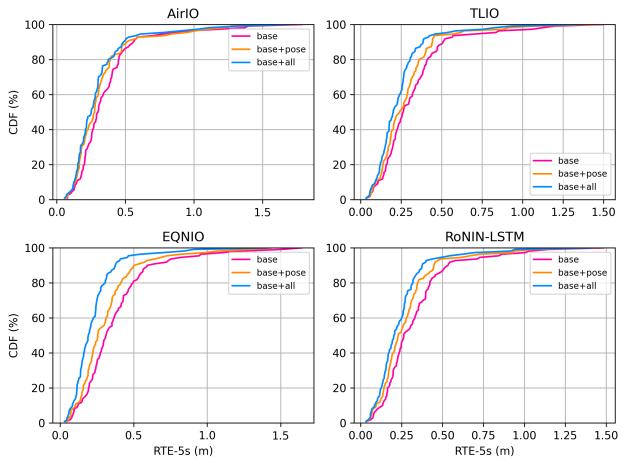

Figure 6. RTE-5s CDF on the Nymeria dataset for AirIO, TLIO, EqNIO, and RoNIN-LSTM. We show the cumulative distribution of 5-second relative trajectory error for all four base models.

Table 3. TLIO dataset result. Metrics reported: ATE (m; horizontal H, vertical V), RTE-5s (m), RTE-1s (m), and % Drifting (lower is better). RoNIN refers to RoNIN-LSTM variant.

Model	ATE (H, V)	RTE-5s	RTE-1s	Drifting
AirIO	2.191 (2.034, 0.632)	0.358	0.092	2.36
AirIO + Pose	**1.950 (1.848, 0.462)**	**0.344**	**0.085**	**1.83**
TLIO	2.415 (2.259, 0.357)	0.314	0.100	3.28
TLIO + Pose	**2.023 (1.890, 0.249)**	**0.307**	**0.097**	**2.70**
EQNIO	2.116 (1.965, 0.319)	0.339	0.104	**2.55**
EQNIO + Pose	**2.073 (1.916, 0.315)**	**0.326**	**0.100**	2.56
RoNIN	2.953 (2.806, **0.335**)	0.345	0.107	4.06
RoNIN + Pose	**2.630** (**2.466**, 0.384)	**0.330**	**0.101**	**3.76**

Gravity Removal Ablation To investigate whether the model benefits from gravity-removed linear acceleration versus raw acceleration, we conduct an ablation study in Table 4. We compare models trained with (w/ g) and without (w/o g) gravitational components in the acceleration data. Results show that removing gravity consistently improves performance across both AirIO and TLIO with all sensors fused. For AirIO, gravity removal reduces ATE from 6.121m to 4.641m and drifting from 3.58% to 2.16%. Similarly, TLIO shows improvements with ATE decreasing from 6.613m to 5.734m when gravity is removed. These results suggest that linear acceleration is more informative than raw acceleration for head tracking tasks.

Fusion Strategy Ablation Table 5 compares our sensor-fusion strategy with alternative fusion modalities. We evaluate (i) direct early-stage concatenation, where pose joint parameters are concatenated with accelerometer and gyroscope measurements, and (ii) cross-attention, where pose parameters are first encoded by a CNN-based pose encoder and the resulting pose features attend to all IMU features via a cross-attention module. The experiments indicate that our approach achieves better ATE, RTE, and drift performance in most cases.

Table 4. Nymeria results with "(+ all)" ablations (with/without gravity). Metrics: ATE (m), RTE-5s (m), RTE-1s (m), and % Drifting (lower is better).

Model	Gravity	ATE	RTE-5s	RTE-1s	Drifting
AirIO + all	w/ g	6.121	0.347	0.110	3.58
AirIO + all	w/o g	**4.641**	**0.297**	**0.078**	**2.16**
TLIO + all	w/ g	6.613	0.287	0.095	4.01
TLIO + all	w/o g	**5.734**	**0.242**	**0.083**	**3.64**

Table 5. Nymeria results with fusion strategy ablations. Metrics: ATE (m), RTE-5s (m), RTE-1s (m), and % Drifting.

Model	Fusion	ATE	RTE-5s	RTE-1s	Drifting
AirIO + Pose	raw concat	9.739	0.509	0.171	6.08
AirIO + Pose	cross attn	5.767	0.398	0.116	2.64
AirIO + Pose	ours	**5.218**	**0.323**	**0.085**	**2.35**
TLIO + Pose	raw concat	8.323	0.282	**0.092**	5.18
TLIO + Pose	cross attn	8.471	0.284	0.093	5.52
TLIO + Pose	ours	**7.972**	**0.281**	0.096	**4.94**

6. Limitations

While the MARIO framework shows clear benefits from integrating human-pose priors and multi-sensor fusion for inertial odometry, several limitations remain. PoseNet introduces additional latency; future work could explore injecting kinematic priors directly into existing inertial odometry models for better efficiency. Magnetometers remain vulnerable to indoor disturbances, and barometric altitude is sensitive to pressure drift and weather changes. Future work includes learned reliability gating for corrupted sensors, self-calibration for misalignment, and further optimization for real-time deployment on resource-constrained wearables.

7. Conclusion

We present MARIO, an inertial odometry framework that augments off-the-shelf IO models with an IMU-inferred pose prior to ground motion estimation in human kinematics. Beyond standard IMU input, we also incorporate lightweight sensing modalities, including a barometer, a magnetometer, and a second on-glasses IMU, through an efficient multimodal fusion strategy. MARIO plugs into existing IMU-based backbones and improves trajectory accuracy by up to 42%.

References

[1] Xiya Cao, Caifa Zhou, Dandan Zeng, and Yongliang Wang. Rio: Rotation-equivariance supervised learning of robust inertial odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6614–6623, 2022. 2

[2] Changhao Chen, Xiaoxuan Lu, Andrew Markham, and Niki Trigoni. Ionet: Learning to cure the curse of drift in inertial odometry. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[3] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem, 2017. 1

[4] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 4

[5] Guoquan Huang. Visual-inertial navigation: A concise review. In *2019 international conference on robotics and automation (ICRA)*, pages 9572–9582. IEEE, 2019. 3

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

[7] Royina Karegoudra Jayanth, Yinshuang Xu, Ziyun Wang, Evangelos Chatzipantazis, Daniel Gehrig, and Kostas Daniilidis. Eqnio: Subequivariant neural inertial odometry. *arXiv preprint arXiv:2408.06321*, 2024. 1, 2, 3, 6

[8] 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. 3

[9] A. R. Jimenez, F. Seco, C. Prieto, and J. Guevara. A comparison of pedestrian dead-reckoning algorithms using a low-cost mems imu. In *2009 6th IEEE International Symposium on Intelligent Signal Processing*, pages 37–42. IEEE, 2009. 2

[10] A.R. Jiménez, F. Seco, J.C. Prieto, and J. Guevara. Indoor pedestrian navigation using an ins/ekf framework for yaw drift reduction and a foot-mounted imu. In *2010 7th Workshop on Positioning, Navigation and Communication*, pages 135–143, 2010. 2

[11] Daniel Laidig and Thomas Seel. Vqf: Highly accurate imu orientation estimation with bias estimation and magnetic disturbance rejection. *Information Fusion*, 91:187–204, 2023. 3

[12] Dongjae Lee, Minwoo Jung, Wooseong Yang, and Ayoung Kim. Lidar odometry survey: recent advancements and remaining challenges. *Intelligent Service Robotics*, 17(2):95–118, 2024. 3

[13] Woosik Lee, Patrick Geneva, Chuchu Chen, and Guoquan Huang. Mins: Efficient and robust multisensor-aided inertial navigation system, 2023. 3

[14] Yadong Li, Shuning Wang, Yongjian Fu, Justin Chen, Xingyu Chen, Ju Ren, Xinyu Zhang, Akshay Gadre, and Ke Sun. Ultraposer: Pushing the limits of imu-based full-body pose estimation with ultrasound sensing on consumer wearables. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, 2025. 3

[15] Yan Li, Yang Xu, Changhao Chen, Zhongchen Shi, Wei Chen, Liang Xie, Hongbo Chen, and Erwei Yin. M2eit: Multi-domain mixture of experts for robust neural inertial tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 28207–28216, 2025. 2

[16] Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I. Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. Tlio: Tight learned inertial odometry. *IEEE Robotics and Automation Letters*, 5(4):5653–5660, 2020. 2, 3, 5, 6

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[18] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024. 2, 5

[19] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild, 2024. 1, 2, 4, 5

[20] Sebastian O. H. Madgwick, Andrew J. L. Harrison, and Ravi Vaidyanathan. Estimation of imu and marg orientation using a gradient descent algorithm. In *2011 IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 1–7, Zurich, Switzerland, 2011. IEEE. 3

[21] Robert Mahony, Tarek Hamel, and Jean-Michel Pflimlin. Nonlinear complementary filters on the special orthogonal group. *IEEE Transactions on Automatic Control*, 53(5): 1203–1218, 2008. 3

[22] 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. 3

[23] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1

[24] Yuheng Qiu, Can Xu, Yutian Chen, Shibo Zhao, Junyi Geng, and Sebastian Scherer. Airio: Learning inertial odometry with enhanced imu feature observability, 2025. 1, 2, 3, 6

[25] Yunzhe Shao, Xinyu Yi, Lu Yin, Shihui Guo, Junhai Yong, and Feng Xu. Magshield: Towards better robustness in

sparse inertial motion capture under magnetic disturbances, 2025. 3

[26] Scott Sun, Dennis Melamed, and Kris Kitani. Idol: Inertial deep orientation-estimation and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6128–6137, 2021. 2

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

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

[29] Xsens Technologies B.V. Xsens IMU Systems. https://www.xsens.com. Accessed: 2024-03-07. 3

[30] Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2024. Association for Computing Machinery. 3

[31] Hang Yan, Qi Shan, and Yasutaka Furukawa. Ridi: Robust imu double integration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 621–636, 2018. 2

[32] Hang Yan, Sachini Herath, and Yasutaka Furukawa. Ronin: Robust neural inertial navigation in the wild: Benchmark, evaluations, and new methods, 2019. 1, 2, 3, 6

[33] Yuan Yao, Shifan Jiang, Yangqing Hou, Chengxu Zuo, Xinrui Chen, Shihui Guo, and Yipeng Qin. Tof-ip: time-of-flight enhanced sparse inertial poser for real-time human motion capture. 2025. 3

[34] 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. 3

[35] 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. 3

[36] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): modeling non-inertial effects in sparse-inertial human motion capture. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

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

[38] Libo Zhang, Xinyu Yi, and Feng Xu. Baroposer: Real-time human motion tracking from imus and barometers in everyday devices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, page 1–9. ACM, 2025. 3

[39] Shibo Zhao, Sifan Zhou, Raphael Blanchard, Yuheng Qiu, Wenshan Wang, and Sebastian Scherer. Tartan imu: A light foundation model for inertial positioning in robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22520–22529, 2025. 2

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. 3

[41] 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 (TOG)*, 44(4):1–14, 2025. 3